Risk Group Turnover in Epidemic Models

Abstract

Keywords: mathematical modelling, HIV, STI, risk heterogeneity

Abbreviations: HIV: human immunodeficiency virus, TPAF: transmission population attributable fraction

Contents

1. Introduction

Core group theory has long underpinned the study of epidemics of sexually transmitted infections (STI). The theory posits that heterogeneity in acquisition and transmission risk are sometimes necessary and sometimes sufficient for an STI epidemic to emerge and persist. This heterogeneity is often demarcated by identifying potential cores, comprised of sub-populations or geographies, where risks of acquisition and onward transmission are the highest, such that the core's unmet STI prevention and treatment needs sustain local epidemics (Yorke et al., 1978; Gesink et al., 2011). Such cores often comprise so-called key populations, such as female sex workers, men who have sex with men, and transgender people.

Mathematical models of STI transmission include heterogeneity in risk by stratifying the modelled population by features such as the partner change rate, levels of sexual mixing between subgroups, and partnership types (Mishra et al., 2012). The implications of including heterogeneity, as compared to assumptions of homogeneity, include higher basic reproductive ratios R_0, and lower overall STI prevalence (provided the latter still results in $R_0 > 1$) (Boily and Mâsse, 1997). R_0 and overall STI prevalence are further influenced by mixing between subgroups (Stigum et al., 1994; Boily and Mâsse, 1997).

Mathematical models with multiple risk groups can then be used to quantify the contribution of high-risk groups to overall transmission. For example, the *transmission population attributable fraction* (TPAF) of the group can be computed (Mishra et al., 2012). TPAF is defined as: the fraction of all new infections that stem, directly and indirectly, from a failure to prevent infection in a particular risk group. The TPAF can be used to help guide "prioritized" or "targeted" interventions for groups at highest risk.

In most transmission models, individuals in a particular risk group are assumed to remain in that group for their entire life. Rarely discussed or included in transmission models is the movement of individuals between risk groups, which we refer to as "turnover". For example, consider a high risk group representing sex work, for which there are a larger number of sexual partners as paid clients, and other STI-associated vulnerabilities (Watts et al., 2010). Individuals may retire from sex work but continue to be sexually active at a lower level of risk, or enter into sex work following a period of lower risk (Boily et al., 2015). Individuals may also enter into and exit from a group engaging in multiple partnerships, and so on.

Several authors have implicated risk group turnover as an important factor in model outputs, such as: the predicted equilibrium prevalence (Stigum et al., 1994; Eaton and Hallett, 2014); the fraction of transmissions occurring during acute infection (Zhang et al., 2012); the basic reproductive number R_0 (Henry and Koopman, 2015); and the level of universal treatment required to achieve epidemic control (Henry and Koopman, 2015). Yet, implementations of risk groups and turnover in recent models vary widely. In works by Koopman et al., rates of movement between two risk groups are balanced analytically based on the size of the groups; Boily et al. (2015) use a 100-year burn-in period to equilibrate a complex system of turnover transitions; some authors only model turnover in the direction of high to low risk, (Stigum et al.,

1994; Eaton and Hallett, 2014). These various approaches to turnover are contrasted with implementations of model features like assortative sexual mixing, which typically follow "standard" methods, such as that proposed by Nold (1980).

Perhaps widespread and consistent implementation of turnover dynamics in STI models is made difficult by the following two challenges. First, it is not clear how epidemiological data can be used to inform rates of turnover among risk groups. Second, as shown by Boily et al. (2015), naive selection of turnover rates can result in imbalanced flows, causing risk groups to initially change size over time.

As a solution to these problems, we introduce a unified framework for parameterizing risk group turnover, based on available data. We develop this framework in Section 2. In Sections 3 and 4, we sought to examine the influence of turnover on the estimated TPAF of a high risk group, using an illustrative STI model. However, to understand this influence of turnover on TPAF, we first investigated the influence of turnover on group-specific incidence and prevalence (Experiment 1). Similarly, since the uncertain model parameters are often fitted so that the model reproduces observed prevalence data, we also examined the influence of turnover on fitted parameters (Experiment 2). The influence of turnover on TPAF was then explored, before and after model fitting (Experiment 3). We discuss the implications of these experiments in Section 5.

2. System

TODO: finish integrating SM revisions

In this section we introduce a unified system of equations to describe risk group turnover in deterministic epidemic models with heterogeneity in risk. We then describe how the unified approach can be used in practical terms, based on different assumptions and data available for parameterizing turnover in risk. We then conclude by framing previous approaches to this task using the proposed system.

2.1. Notation

Consider a population divided into G risk groups. We denote the number of individuals in risk group $i \in [1, \ldots, G]$ as x_i and the set of all risk groups as $\boldsymbol{x} = \{x_1, \ldots, x_G\}$. The total population size is $N = \sum_i x_i$, and the relative population size of each group is denoted as $\hat{x}_i = x_i/N$. Individuals enter the population at a rate ν per year, and exit at a rate μ per year. We assume that individuals entering into the population originate from another exogenous population $\boldsymbol{e} = \{e_1, \ldots, e_G\}$. We make this assumption so that the exogenous population $\boldsymbol{e}$ may have different relative risk group sizes from the system population $\boldsymbol{x}$.[1] That is, $\hat{e}_i$ may not necessarily equal $\hat{x}_i$. However, since the entry rate ν is relative to the size N of population $\boldsymbol{x}$, it will be convenient and otherwise inconsequential to assume that population $\boldsymbol{e}$ also has size N. That

[1] We could equivalently stratify the rate of entry ν by risk group while keeping the exogenous population $\boldsymbol{e}$ equal to the system population $\boldsymbol{x}$. However, we find this formulation more difficult to work with in the subsequent sections.

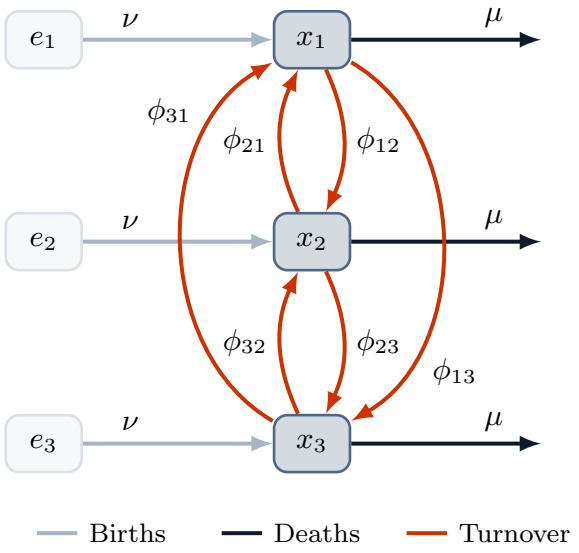


Figure 1: System of risk groups and flows between them for $G = 3$

way, the total number of individuals entering into population $\boldsymbol{x}$ per year is given by νN, and the number of individuals entering into group x_i specifically is given by $\hat{e}_i \nu N$.

Turnover transitions may then occur between any two groups, in either direction. Therefore we denote the turnover rates as a $G \times G$ matrix ϕ. The element ϕ_{ij} corresponds to the proportion of individuals in group x_i who move from group x_i to group x_j each year. An example matrix is given in Eq. (1), where we write the diagonal elements as $*$ since they represent transitions from a group to itself.

$$\phi = \begin{bmatrix} * & x_1 \to x_2 & \cdots & x_1 \to x_G \\ x_2 \to x_1 & * & \cdots & x_2 \to x_G \\ \vdots & \vdots & \ddots & \vdots \\ x_G \to x_1 & x_G \to x_2 & \cdots & * \end{bmatrix} \tag{1}$$

Risk groups, transitions, and the associated rates are also shown for $G = 3$ in Figure 1.

2.2. Parameterization

Next, we consider the goal of constructing a system like the one introduced above which reflects the risk group dynamics observed in a specific context. We assume that the relative sizes of the risk groups in the model ($\hat{\boldsymbol{x}}$) are already known, and should remain constant over time. Thus, what remains is to estimate the values of the parameters: ν, μ, $\hat{\boldsymbol{e}}$, and ϕ, using commonly available sources of data.

2.2.1. Total Population Size

The total population size $N(t)$ is a function of the rates of population entry $\nu(t)$ and exit $\mu(t)$, given an initial size N_0. We allow the proportion entering the system to vary by risk group via $\hat{\boldsymbol{e}}$, while the exit rate has the same value for each group. We assume that there is no disease-attributable death. Because the values of ν and μ are the same for each risk group, they can be estimated independent of $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{e}}$, and ϕ.

The difference between entry and exit rates defines the rate of population growth:

$$\mathcal{G}(t) = \nu(t) - \mu(t) \tag{2}$$

The total population may then be defined using an initial population size N_0 as:

$$N(t) = N_0 \exp\left(\int_0^t \log\big(1 + \mathcal{G}(\tau)\big) d\tau\right) \tag{3}$$

which, for constant growth, simplifies to the familiar expression (Malthus, 1798):

$$N(t) = N_0(1 + \mathcal{G})^t \tag{4}$$

Census data, such as (DataBank, 2019), can be used to source the total population size in a given geographic setting over time $N(t)$, thus allowing Eqs. (3) and (4) to be used to estimate $\mathcal{G}(t)$.

If the population size is assumed to be constant, then $\mathcal{G}(t) = 0$ and $\nu(t) = \mu(t)$. If population growth occurs at a stable rate, then $\mathcal{G}$ is fixed at a constant value which can be estimated via Eq. (4) using any two values of $N(t)$, separated by a time interval τ:

$$\mathcal{G}_\tau = \frac{N(t+\tau)^{\frac{1}{\tau}}}{N(t)} - 1 \tag{5}$$

If the rate of population growth $\mathcal{G}$ varies over time, then Eq. (5) can be reused for consecutive time intervals, and the complete function $\mathcal{G}(t)$ approximated piecewise by constant values. The piecewise approximation can be more feasible than exact solutions using Eq. (3), and can reproduce $N(t)$ accurately for small enough intervals τ, such as one year.

Now, given a value of $\mathcal{G}(t)$, either $\nu(t)$ must be chosen and $\mu(t)$ calculated using Eq. (2), or $\mu(t)$ must be chosen, and $\nu(t)$ calculated. Most modelled systems assume a constant duration of time that individuals spend in the model $\delta(t)$ (Anderson and May, 1991) which is related to the rate of exit μ by:

$$\delta(t) = \mu^{-1}(t) \tag{6}$$

In the context of sexually transmitted infections, the duration of time usually reflects the average sexual life-course of individuals from age 15 to 50 years, such that $\delta = 35$ years. The duration δ may also vary with time to reflect changes in life expectancy. The exit rate $\mu(t)$ can then be defined as $\delta^{-t}(t)$ following Eq. (6), and the entry rate $\nu(t)$ defined as $\mathcal{G}(t) - \mu(t)$ following Eq. (2).

2.2.2. Turnover

Next, we present methods for resolving the distribution of individuals entering the risk model $\hat{\boldsymbol{e}}(t)$ and the rates of turnover $\phi(t)$, assuming that entry and exit rates $\nu(t)$ and $\mu(t)$ are known. Similar to above, we first formulate the problem as a system of equations. Then, we explore the data and assumptions required to solve for the values of parameters in the system. The (t) notation is omitted throughout this section for clarity, though time-varying parameters can be estimated by repeating the necessary calculations for each t.

The number of risk groups G specifies the number of G unknown elements in $\hat{\boldsymbol{e}}$ and $G(G-1)$ unknown elements in ϕ. We collect these unknowns in the vector $\boldsymbol{\theta} = [\hat{\boldsymbol{e}}, \boldsymbol{y}]$, where $\boldsymbol{y} = \text{vec}_{i \neq j}(\phi)$. For example, for $G = 3$, the vector $\boldsymbol{\theta}$ is defined as:

$$\boldsymbol{\theta} = \begin{bmatrix} \hat{e}_1 & \hat{e}_2 & \hat{e}_3 & \phi_{12} & \phi_{13} & \phi_{21} & \phi_{23} & \phi_{31} & \phi_{32} \end{bmatrix} \tag{7}$$

We then define a linear system of equations which uniquely determine the elements of $\boldsymbol{\theta}$:

$$\boldsymbol{b} = A\,\boldsymbol{\theta} \tag{8}$$

where A is a $M \times G^2$ matrix and $\boldsymbol{b}$ is a M-length vector. Specifically, each row in A and $\boldsymbol{b}$ defines a constraint: an assumed mathematical relationship involving one or more elements of $\hat{\boldsymbol{e}}$ and ϕ. For example, a simple constraint could be to assume the value $\hat{e}_2 = 0.20$. Each of the following sections introduces a type of constraint, including: assuming a constant group size, specifying elements of $\boldsymbol{\theta}$ directly, assuming an average duration in a group, and assuming balanced rates of turnover. Constraints may be selected and combined together based on availability of data and plausibility of assumptions, provided a total of $M = G^2$ constraints are defined. The values of $\hat{\boldsymbol{e}}$ and ϕ can then be calculated algebraically using $\boldsymbol{\theta} = A^{-1}\boldsymbol{b}$, for which many algorithms exist (LAPACK, 1992)

Constraint 1: Constant group size. First, we define the "conservation of mass" equation for group x_i, wherein the rate of change of the group is defined as the sum of flows in / out of the group:

$$\frac{d}{dt}x_i = \nu\, e_i + \sum_j \phi_{ji}\, x_j - \mu\, x_i - \sum_j \phi_{ij}\, x_i \tag{9}$$

While Eq. (9) is written in terms of absolute population sizes $\boldsymbol{x}$ and $\boldsymbol{e}$, it is equivalent to divide through by N, yielding a system in terms of proportions $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{e}}$, which can be more useful, since N need not be known. If we assume that the proportion of each group $\hat{x}_i$ is constant over time, then the desired rate of change for risk group i will be equal to the growth of the risk group, $\mathcal{G}x_i$. Substituting $\frac{d}{dt}x_i = \mathcal{G}x_i$ into Eq. (9), and simplifying yields:

$$\nu\, x_i = \nu\, e_i + \sum_j \phi_{ji}\, x_j - \sum_j \phi_{ij}\, x_i \tag{10}$$

Factoring the left and right hand sides in terms of $\hat{\boldsymbol{e}}$ and ϕ, we obtain G unique constraints. For $G = 3$, this yields the following 3 rows as the basis of $\boldsymbol{b}$ and A:

$$\boldsymbol{b} = \begin{bmatrix} \nu x_1 \\ \nu x_2 \\ \nu x_3 \end{bmatrix}; \qquad A = \begin{bmatrix} \nu & \cdot & \cdot & -x_1 & -x_1 & x_2 & \cdot & x_3 & \cdot \\ \cdot & \nu & \cdot & x_1 & \cdot & -x_2 & -x_2 & \cdot & x_3 \\ \cdot & \cdot & \nu & \cdot & x_1 & \cdot & x_2 & -x_3 & -x_3 \end{bmatrix} \tag{11}$$

These G constraints are necessary to ensure risk groups do not change size over time. However, to obtain a unique solution, we still need an additional $G(G-1)$ constraints. For $G = 3$, this corresponds to 6 additional constraints.

Constraint 2: Specified elements. The simplest type of additional constraint is to directly specify the value of individual elements in $\hat{\boldsymbol{e}}$ or ϕ. Such constraints may be appended to $\boldsymbol{b}$ and A as an additional row k using indicator notation.[2] That is, with b_k as the specified value and $A_k = [0, \ldots, 1, \ldots, 0]$ as the indicator vector, with 1 in the same position as the desired element in $\boldsymbol{\theta}$. For example, for $G = 3$, if it is known that 20% of individuals enter directly into risk group x_2 upon entry into the model ($\hat{e}_2 = 0.20$), then $\boldsymbol{b}$ and A can be augmented with:

$$\boldsymbol{b}_k = \begin{bmatrix} 0.20 \end{bmatrix}; \qquad A_k = \begin{bmatrix} \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \tag{12}$$

since $\hat{e}_2$ is the second element in $\boldsymbol{\theta}$. If we do not want any turnover from group i to group j, then Eq. (12) can also be used to set $\phi_{ij} = 0$.

Note that the elements of $\hat{\boldsymbol{e}}$ must sum to one. Therefore, specifying all elements in $\hat{\boldsymbol{e}}$ will only provide $G - 1$ constraints, as the last element will be either redundant or violate the. This relationship is implicit in Eq. (11), so it is not necessary to supply a constraint $1 = \sum_i \hat{e}_i$. Similar redundancies or inconsistencies can emerge for constraints on ϕ, as noted below.

Constraint 3: Group duration. Another useful constraint can be derived from the average duration of individuals in a risk group. This duration δ_i is defined as the inverse of all efferent flow rates:

$$\delta_i = \left(\mu + \sum_j \phi_{ij} \right)^{-1} \tag{13}$$

Average durations could be derived from survey data, including for key populations, or they could be assumed. These values can be used to define constraints on ϕ by rearranging Eq. (13) to yield: ${\delta_i}^{-1} - \mu = \sum_j \phi_{ij}$. For example, if for $G = 3$, the average duration in group x_1 is known to be $\delta_1 = 5$ years, then $\boldsymbol{b}$ and A can be augmented with:

$$\boldsymbol{b}' = \begin{bmatrix} 5^{-1} - \mu \end{bmatrix}; \qquad A' = \begin{bmatrix} \cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \tag{14}$$

[2] Indicator notation, also known as "one-hot notation" is used to select one element from another vector, based on its position. An indicator vector is 1 in the same location as the element of interest, and 0 everywhere else.

As noted above, redundancies can emerge when constraining turnover rates ϕ via duration δ_i. Namely, specifying all efferent flow rates ϕ_{ij} for one group i will fully determine the duration in the group δ_i. In general, each constraint which is not redundant will increase the rank of A by one, and A must be full rank (G^2) to uniquely determine the parameters in $\boldsymbol{\theta}$.

Constraint 4: Balancing rates of turnover. A final constraint is: if we assume that the absolute number of individuals moving between two risk groups is related by a ratio $\frac{r_i}{r_j}$ so that: $\phi_{ij} x_i r_i = \phi_{ji} x_j r_j$. This helps constrain ϕ_{ij} and ϕ_{ji}. For example, for $G = 3$, if we assume that the number of individuals moving between groups x_1 and x_2 is related by $\frac{3}{2}$, then $\boldsymbol{b}$ and A can be augmented with:

$$\boldsymbol{b}' = \begin{bmatrix} 0 \end{bmatrix}; \qquad A' = \begin{bmatrix} \cdot & \cdot & \cdot & 3x_1 & \cdot & -2x_2 & \cdot & \cdot & \cdot \end{bmatrix} \tag{15}$$

If we assume that the absolute number of individuals moving between the groups is equal, then $\frac{r_i}{r_j} = 1$. Again, care should be taken to avoid redundancies and inconsistencies with other constraints.

Solving the System. Now, given a set of sufficient constraints on $\hat{\boldsymbol{e}}$ and ϕ to ensure exactly one solution, the system can be solved using $\boldsymbol{\theta} = A^{-1}\boldsymbol{b}$. The resulting values of $\hat{\boldsymbol{e}}$ and ϕ can then be used to run the model.

However, we may find that we have insufficient constraints, implying that there are multiple definitions of $\boldsymbol{\theta}$ which satisfy our constraints (A is less than full rank). In this situation, we have two options. First, we can pose the problem as a minimization problem, namely:

$$\boldsymbol{\theta}^* = \arg\min f(\boldsymbol{\theta}), \quad \text{subject to: } \boldsymbol{b} = A\,\boldsymbol{\theta};\ \ \boldsymbol{\theta} \geq 0 \tag{16}$$

where f is a function which globally constrains $\boldsymbol{\theta}$, such as: $||\cdot||_2$. This approach will find the smallest values of $\hat{\boldsymbol{e}}$ and ϕ which satisfy the given constraints.[3] Alternatively, since underdetermined systems have infinite solutions defined by a set of basis vectors, the solutions could be sampled randomly during model fitting, provided that the basis vector coefficients are themselves reasonably bounded.

Similarly, we may find that no solution exists for the given constraints, since two or more constraints are incompatible. In this case, the conflict should be resolved by relaxing one of the offending constraints. For example, if a turnover rate ϕ_{ij} and the duration in the group δ_i are both specified, the turnover rate ϕ_{ij} cannot exceed ${\delta_i}^{-1} - \mu$, due to Eq. (13).

2.3. Previous Approaches

While most models have included population growth, many have not included risk heterogeneity within age-sex groups, and still fewer have simulated turnover among risk groups. These three major features about

[3] Numerical solutions to such problems are widely available, such as the Non-Negative Lease Squares solver (Lawson and Hanson, 1995), available in Python: https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.nnls.html.

Box 1: Common assumptions regarding the dynamics of risk groups

1. **Population Growth:** The total population size N increases over time.
   (a) **No:** $\nu = \mu$; Population size N is constant.
   (b) **Yes:** $\nu > \mu$; Population size N increases.
2. **Risk Groups:** Demographic groups are stratified by risk of infection acquisition.
   (a) **No:** $G = 1$; Demographic groups are homogeneous in risk of infection acquisition.
   (b) **Yes:** $G > 1$; Heterogeneity in risk of infection acquisition within demographic groups is considered.
3. **Turnover:** Individuals may move between risk groups.
   (a) **No:** $\phi = 0$; Individuals do not move between risk groups.
   (b) **Yes:** $\phi > 0$; Individuals move between risk groups.

risk group dynamics and the associated assumptions are summarized in Box 1. In each case, option (b) typically represents the more plausible assumption. Experiment 1 in the next section aims to illustrate the potential implications of assuming option (a) in each case.

Among those works which have considered risk groups and turnover among them, implementations have varied widely. In a $G = 2$ system, Stigum et al. (1994) use a parameter κ to define the rate of movement of individuals from a core group into the remaining population, which is balanced by an equal number of individuals moving in the opposite direction. Considered through the proposed framework, this defines a unique system through assuming: a *specified element* $\kappa = \phi_{12}$, *constant group size*, and *balanced turnover*. Henry and Koopman (2015) employ the same assumptions for another $G = 2$ system, but define a "re-selection rate" ω to control the relative rate of turnover, so that $\phi_{12} = \omega\hat{x}_2$ and $\phi_{21} = \omega\hat{x}_1$. In a $G = 3$ system, Eaton and Hallett (2014) use a parameter Ψ to define the rate of movement from high-to-low, high-to-medium, and medium-to-low risk groups each year, but assume no transitions in the opposite direction. Thus, all six elements of ϕ are *specified*, while *constant group sizes* are ensured by computing the required distribution of model entrants $\hat{\boldsymbol{e}}$ via equation [S12] in the Supplemental Information. The vector $\hat{\boldsymbol{e}}$ could equivalently be resolved using the methods outlined in Section 2.2.

The $G = 7$ system used by Boily et al. (2015) is highly contextual, and includes four high-risk groups which transition to specific low-risk groups (e.g. from "high-/low-volume sex work" to "formerly engaged in sex work"). The turnover matrix ϕ is then completely *specified* (though notably sparse) using several assumptions about *group durations*. The distribution of risk groups among individuals entering the model $\hat{\boldsymbol{e}}$ is also *specified*, leaving the distribution of risk groups among individuals in the model $\hat{\boldsymbol{x}}$ as the only

Table 1: Summary of experiments

Experiment	Outputs	Stratifications	Turnover	Treatment	Fitting
1.1.	prevalence	overall	any vs none	fixed, 2 rates	none
1.2.	prevalence, incidence	overall, by group	varied	fixed, 1 rate	none
1.3.	prevalence, incidence	overall, by group	varied	varied	none
2.	inferred contact rate	by group	any vs none	fixed, 1 rate	prevalence by group
3.1.	TPAF	highest risk	any vs none	fixed, 1 rate	none
3.2.	TPAF	highest risk	any vs none	fixed, 1 rate	prevalence by group

TPAF: transmission population attributable fraction

free parameters. For this reason, a 100-year "burn-in" period is required to equilibrate risk-group sizes before introduction of the STI in the model. Note that this approach is not perfectly compatible with the framework presented here, since we assume *constant group sizes* as the basis for the system in Eq. (11), and exclude $\hat{\boldsymbol{x}}$ from the vector of free parameters $\boldsymbol{\theta}$. However, by relaxing these constraints, it should be possible to formulate even the implementation by Boily et al. (2015) in terms of the proposed framework.

In sum, the framework for modelling turnover presented here aims to generalize all previous implementations. In so doing, we hope to clarify the requisite assumptions, dependencies on epidemiological data, and relationships between previous approaches. In Experiments 2 and 3, we leverage this framework to explore the potential influences of turnover on simulated epidemics.

3. Experiment

We aimed to determine and understand the influence of risk group turnover on the contribution of the highest risk group to the overall epidemic, as measured by the transmission population attributable fraction (TPAF). However, to understand the underlying mechanism, we first examined the influence of turnover on the equilibrium incidence and prevalence projected among risk groups, as well as overall. Since the influence may depend on the duration of infectiousness, we also explored the sensitivity of these results to different treatment rates. Next, we examined how turnover may influence inferred model parameters through model fitting. Finally, we compared the TPAF of the highest risk group with and without turnover, before and after model fitting. Our experiments can be summarized as in Table 1.

3.1. Model & Simulations

To run our experiments, we developed a deterministic single-sex SIT model which simulates transmission in a population with heterogeneity in risk. The model is not representative of a specific infection but includes

Figure 2: Modelled health states. $\mathcal{S}$: susceptible; $\mathcal{I}$: infected; $\mathcal{T}$: treated; λ: force of infection; τ: treatment.

balancing contacts as per sexually transmitted infections (Garnett and Anderson, 1994). The model includes three health states: susceptible $\mathcal{S}$, infectious $\mathcal{I}$, and treated $\mathcal{T}$ (Figure 2), and $G = 3$ levels of risk: high H, medium M, and low L. Risk strata are defined by different number of contacts per year so that individuals in risk group i are assumed to form contacts at a rate C_i per year. The probability of contact formation ρ_{ik} between individuals in group i and individuals in risk group k is assumed to be proportionate to the total number of available contacts within each group:

$$\rho_{ik} = \frac{C_k x_k}{\sum_{\mathrm{k}} C_{\mathrm{k}} x_{\mathrm{k}}} \tag{17}$$

The biological probability of transmission is defined as β per contact. Individuals transition from the infectious $\mathcal{I}$ to susceptible $\mathcal{S}$ health-state via a force of infection λ per year, per susceptible in risk group i:

$$\lambda_i = C_i \sum_k \rho_{ik}\, \beta\, \frac{\mathcal{I}_k}{x_k} \tag{18}$$

Individuals are assumed to transition from the infectious $\mathcal{I}$ to treated $\mathcal{T}$ health-state at a rate τ per year, reflecting diagnosis and treatment. The treatment rate is not stratified by risk group. Individuals in the treated $\mathcal{T}$ health-state are not infectious nor susceptible, and individuals cannot become re-infected.

As described in Section 2, individuals enter the model at a rate ν, exit the model at a rate μ, and transition from risk group i to group j at a rate ϕ_{ij}. The turnover rates ϕ and distribution of individuals entering the model by risk group $\hat{\boldsymbol{e}}$ were computed using the methods outlined in Section 2.2.2 and the following assumptions. First, we assumed that the proportion of individuals entering each risk group $\hat{\boldsymbol{e}}$ was equal to the proportion of individuals across risk groups in the model $\hat{\boldsymbol{x}}$. Second, we assumed that the average duration spent in each risk group $\boldsymbol{\delta}$ is known. Third, we assumed that the absolute number of individuals moving between two risk groups in either direction is balanced. The system of equations which results from these assumptions is given in Appendix A.2. To meet all three conditions, there is only one possible value for each element in ϕ and $\hat{\boldsymbol{e}}$ – i.e. A is full rank. In other words, by specifying these three conditions, we ensure that a unique set of ϕ and $\hat{\boldsymbol{e}}$ is computed.

Using the above three assumptions, we need to specify the values of $\hat{\boldsymbol{x}}$, $\boldsymbol{\delta}$, ν, and μ. Such parameters could be derived from data as described in Section 2.2.2; however, in this experiment, we use the illustrative

Table 2: Model parameters. All rates have units year^{-1}; durations are in years; parameters stratified by risk group are written [high, medium, low] risk.

Symbol	Description	Value
$\boldsymbol{\beta}$	transmission probability per contact	0.03
τ	rate of treatment initiation among infected	0.1
N_0	initial population size	1000
$\hat{\boldsymbol{x}}$	proportion of system individuals by risk group	[0.05 0.20 0.75]
$\hat{\boldsymbol{e}}$	proportion of entering individuals risk by risk group	[0.05 0.20 0.75]
$\boldsymbol{\delta}$	average duration spent in each risk group	[5 15 25]
C	rate of contact formation among individuals in each risk grouop	[25 5 1]
ν	rate of population entry	0.05
μ	rate of population exit	0.03

values summarized in Table 2. After resolving the system of equations, $\hat{\boldsymbol{e}}$ is equal to $\hat{\boldsymbol{x}}$ (assumed), and ϕ is:

$$\phi = \begin{bmatrix} * & 0.0833 & 0.0867 \\ 0.0208 & * & 0.0158 \\ 0.0058 & 0.0042 & * \end{bmatrix} \tag{19}$$

We then simulated epidemics using these parameters. The model was initialized with $N_0 = 1000$ individuals who are distributed across risk groups according to $\hat{\boldsymbol{x}}$. We seeded the epidemic with one infectious individual in each risk group at $t = 0$. There were no treated individuals at the start of the epidemic, and so all individuals except the 3 infectious individuals were susceptible. We numerically solved the system of ordinary differential equations in Python[4] using Euler's method with a time step of $dt = 0.1$ years. The full system of model equations is given in Appendix A.1. All comparative analyses are then conducted at equilibrium, defined as a steady state with <1% difference in incidence per year.

3.2. Experiment 1: Influence of turnover on equilibrium incidence and prevalence

Experiment 1 examined the influence of turnover on equilibrium incidence and prevalence, where incidence is defined as λ_i from Eq. (18), and prevalence is defined as $\hat{\mathcal{I}}_i = \dfrac{\mathcal{I}_i}{\mathcal{X}_i}$.

Experiment 1.1: Overall prevalence with vs without turnover. First, we compare the overall prevalence predicted by the model with and without turnover. The model with turnover is as described above. The

[4] Code for all aspects of the project is available at: https://github.com/c-uhs/turnover

model without turnover has all rates $\phi = 0$. Following Eq. (13), this means that in the model without turnover, the time spent within each risk group is equal to the average duration in the modelled population μ^{-1}. The comparison of prevalence is repeated for two different treatment rates, in order to illustrate variability in the relative difference.

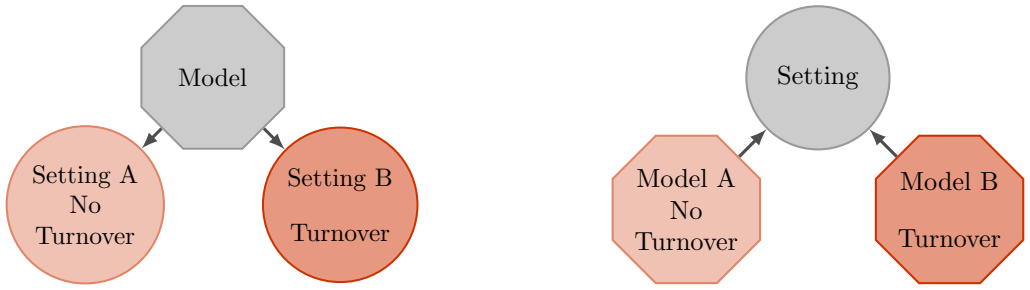
Experiment 1.2: Influence of turnover rates on incidence and prevalence. Second, we determined the influence of different turnover rates on equilibrium incidence and prevalence at a fixed duration of infectiousness (treatment rate). As in similar experiments (Zhang et al., 2012; Henry and Koopman, 2015), the rates of turnover were scaled by a single parameter. However, because the model used here has $G = 3$ risk groups, multiplying by a set of base rates ϕ by a scalar factor would have resulted in changes to the relative population size of risk groups $\hat{\boldsymbol{x}}$. Thus, we controlled the rates of turnover using the duration of individuals in the high risk group δ_H, such that a shorter period of time in the high risk group corresponded to higher rates of turnover among all groups. The duration of individuals in the medium risk group δ_M was then defined as a value between δ_H and the maximum duration μ^{-1} which scales with δ_H following the equation: $\delta_M = \delta_H + \kappa\left(\mu^{-1} - \delta_H\right)$, with $\kappa = 0.3$. The duration of individuals in the low risk group δ_L similarly scaled with δ_H, but the value was not required to calculate ϕ; it can be determined from ϕ afterwards using Eq. (13). In this way, each value of δ_H was used to define a unique set of turnover rates ϕ whose elements all scaled inversely with the duration in the high risk group δ_H. The value of δ_H was then varied from 33 to 3 years to examine the influence of different turnover rates.

Experiment 1.3: Influence of turnover rates at various treatment rates. Next, we conducted a 2-way sensitivity analysis to examine how the influence of turnover might vary at different treatment rates. The treatment rate controls the duration of infectiousness $\delta_\mathcal{I}$ as in $\delta_\mathcal{I} = \tau^{-1}$. Treatment rate τ was varied from 1 to 0.05, implying a duration of infectiousness of 1 to 20 years. The duration of time spent in the high risk group δ_H was varied from 33 to 3 years as in Experiment 1.2. We examined the influence of the rates of turnover on equilibrium prevalence and incidence across the range of treatment rates using multiple 1D plots and 2D surface plots.

3.3. Experiment 2: Inferred risk heterogeneity with vs without turnover

In Experiment 2, we examined the influence of turnover on the parameter values inferred via model fitting. Specifically, we fit the model to 20% infection prevalence among the high risk group, 3% among the low risk group, and 5% overall, with and without turnover . We fit the contact rates C of all risk groups by minimizing the negative log-likelihood of each predicted prevalence versus the target.[5] We then compared

[5] Sample sizes of 50, 750, and 1000 were assumed to generate binomial distributions for the high, low, and overall prevalence targets respectively, and the minimization was performed using the SLSQP method (Kraft, 1988) from the SciPy Python package (`scipy.optimize.minimize`).



(a) Experiment 3.1: same parameters, different prevalence (no model fitting)

(b) Experiment 3.2: same prevalence, different parameters (fitted model)

Figure 3: Two approaches to comparing TPAF with and without turnover

the inferred contact rates C in the model with versus without turnover. The ratio of fitted (or posterior) contact rates C_H / C_L represents the degree of risk heterogeneity in the population, after fixing all other parameters, which produces the given infection prevalence.

3.4. Experiment 3: Influence of turnover on the TPAF of the high risk group

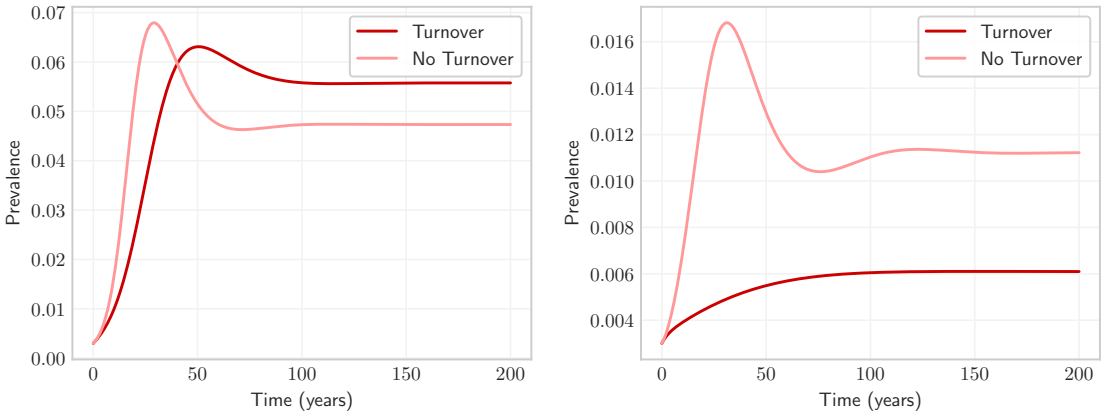
Finally, Experiment 3 sought to examine how the estimated contribution of highest risk group to overall transmission, as measured by the transmission population attributable fraction (TPAF), varies with versus without turnover. The TPAF of a risk group i is defined as:

$$\text{TPAF}_i(t) = \frac{I_0(t) - I_i(t)}{I_0(t)} \tag{20}$$

where $I_0(t)$ is the cumulative number of new infections by time t under usual conditions, and $I_i(t)$ is the cumulative number of new infections assuming no transmission from risk group i. Both $I_0(t)$ and $I_i(t)$ are calculated starting from a system at equilibrium. There are two ways to consider the comparison of TPAF with versus without turnover; these are illustrated in Figure 3, and explained in the following sections.

Experiment 3.1: Two settings. First, we compared two hypothetical settings, which had identical populations (including behaviour), except that setting A had no turnover: $\phi = 0$, while setting B had turnover: ϕ from Eq. (19). As a result, model parameters were identical (except turnover), but the infection prevalence predicted for each setting was different. Following equilibration of the model in both settings, the TPAF of the high risk group was then estimated.

Experiment 3.2: Two models. Second, we compared two models, which were identical in structure except that Model A had no turnover and Model B had turnover. In this case, both models were fitted to the same "setting", as defined by overall and group-specific equilibrium infection prevalence (from Experiment 2). As a result, prevalence was the same in both models, but the group-specific contact rates inferred via fitting were different. As before, the TPAF of the high risk group was then estimated in each model after equilibration.



(a) Treatment rate $\tau = 0.1$

(b) Treatment rate $\tau = 0.2$

Figure 4: Overall projected prevalence with and without risk group turnover under two different treatment rates τ.

4. Results

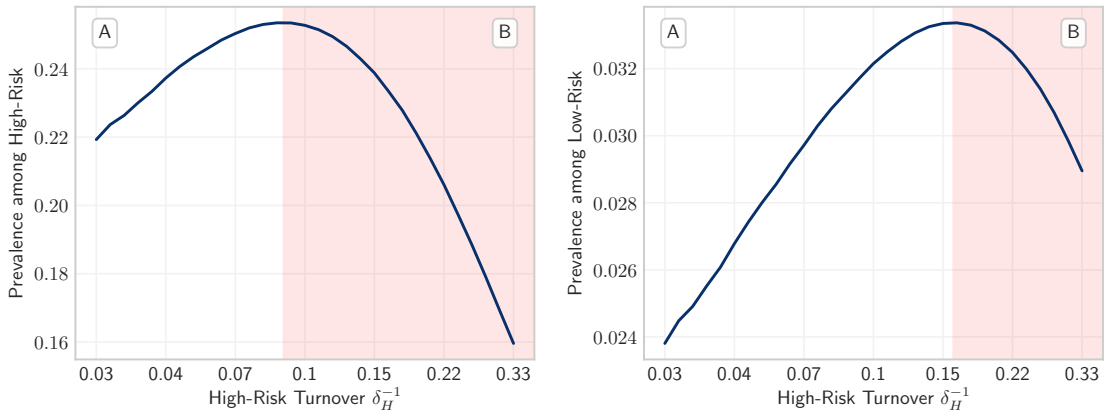
4.1. Experiment 1: Influence of turnover on equilibrium incidence and prevalence

First, we present general trends in equilibrium prevalence and incidence with respect to turnover.

Experiment 1.1: Overall prevalence with vs without turnover. Figure 4 shows the influence of turnover on overall prevalence at two different treatment rates. In both scenarios, turnover appeared to slow the initial epidemic growth as indicated by overall prevalence. However, at equilibrium, the influence of turnover on overall prevalence depended on the treatment rate: prevalence was higher with turnover for $\tau = 0.1$ (Figure 4a), while it was higher without turnover for $\tau = 0.2$ (Figure 4b). Experiments 1.2 and 1.3 aimed to clarify and explain this influence through exploration of group-specific incidence and prevalence at equilibrium under different rates of turnover ϕ and treatment τ.

Experiment 1.2: Influence of turnover rates on incidence and prevalence. Figure 5 illustrates trends in equilibrium prevalence versus turnover among the high and low risk groups for fixed treatment rate ($\tau = 0.1$). For both groups, the same profile was observed: at low turnover, increasing turnover increased prevalence, up to a maximum value (region A), and increasing turnover beyond this point (region B) then decreased prevalence. In the high risk group (Figure 5a), this transition occurred at a lower rate of turnover, while in the low risk group (Figure 5b), the transition occurred at a higher rate of turnover. This peaked prevalence profile can be explained by the interaction between two factors: the movement of individuals between risk groups and incidence.

The movement of individuals between risk groups is depicted in Figure 6 for four rates of turnover and a simplified system (two risk groups). Recall that, by our assumptions, the distribution of health states among individuals leaving a risk group was equal to the distribution of health states within the group. Therefore,



(a) High risk

(b) Low risk

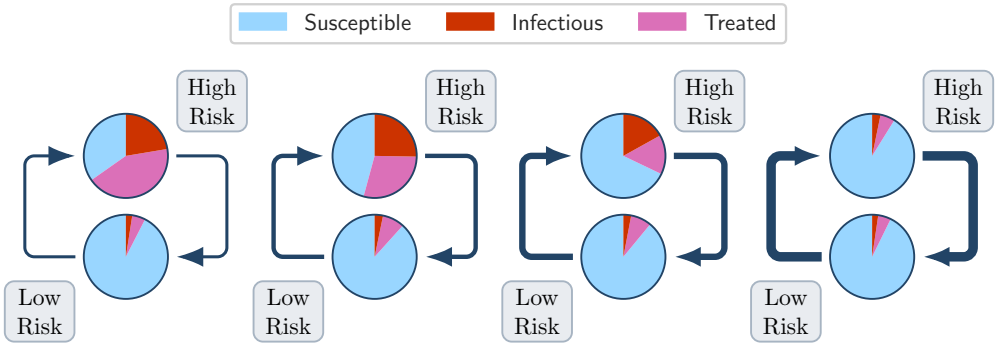
Figure 5: Equilibrium prevalence among high and low risk groups versus turnover, as controlled by the duration in the high risk group δ_H. Turnover shown in log scale.

a higher proportion of individuals leaving the high risk group were infectious as compared to individuals entering the high risk group, who were mostly susceptible. Thus, in the high risk group, turnover yielded a net replacement of infectious individuals with susceptible individuals. However, turnover similarly replaced treated individuals with susceptible individuals in this group. If incidence is sufficiently high, infection of these susceptible individuals can outpace the loss of infectious individuals via turnover. As a result, prevalence among the high risk group can actually increase with increasing turnover. In fact, this is what we observed in this model for moderate rates of turnover (Figure B.2b), yielding the increase in prevalence among high risk individuals shown in region A of Figure 5a.[6] Among the low risk group, turnover yielded a net exchange susceptible individuals for infectious and treated individuals. As a result, moderate rates of turnover also increased prevalence among the low risk group, as shown in region A of Figure 5b.

Next, and in order to explain the reversal of these trends at higher rates of turnover (region B), we examined the second factor in the influence of turnover on infection prevalence: incidence. Consider the force of infection equation, Eq (18). As shown in Appendix A.3, the dynamic component in this expression is the proportion of available partnerships which are offered by infectious individuals, denoted as $\boldsymbol{C}_{\mathcal{I}}$. This component can be further broken down into the following two sub-factors: $\hat{C}_{\mathcal{I}}$ the average contact rate among infectious individuals, and $\hat{\mathcal{I}}$ overall prevalence. Thus the influence of turnover on incidence can be understood through the influence of turnover on these two sub-factors.[7]

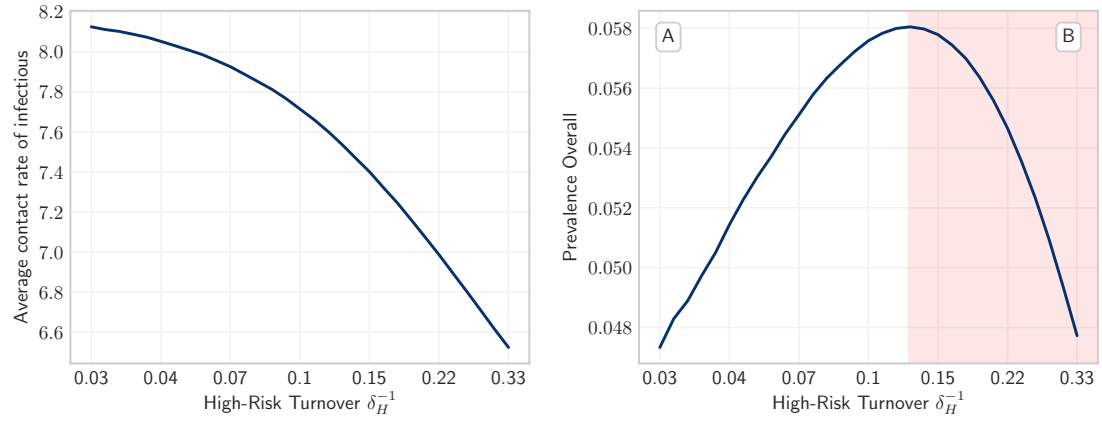
[6] If prevalence is defined as to include both infectious individuals and individuals on treatment, such as in the context of HIV, then infection prevalence among the high risk group monotonically decreased with increasing turnover (Knight et al., 2019).

[7] Since we assumed proportionate mixing among risk groups, and only consider heterogeneity in contact rate C_i, incidence in each risk group is the same as overall except for a scale factor proportional to C_i.



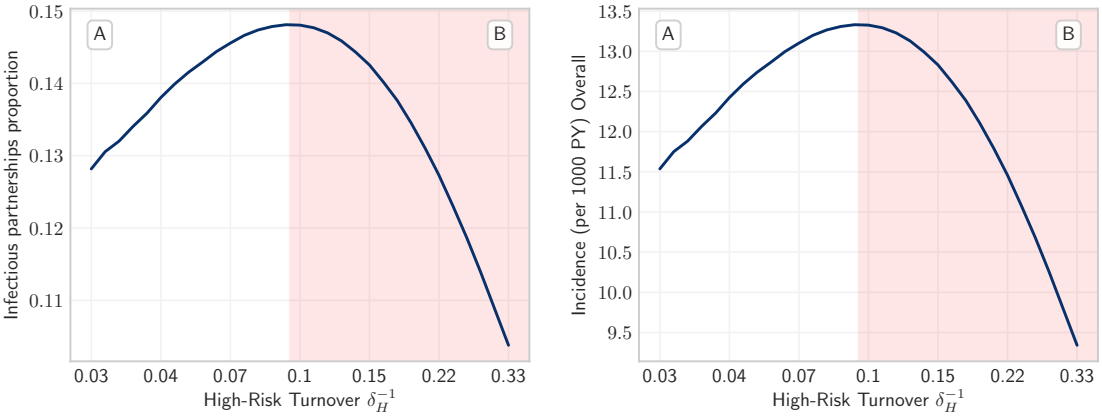
(a) Low turnover (b) Moderate turnover (c) High turnover (d) Very high turnover

Figure 6: Average health states of individuals moving between high and low risk groups due to different rates of turnover.



(a) Average contact rate among infectious individuals $\hat{C}_{\mathcal{I}}$

(b) Overall prevalence $\hat{\mathcal{I}}$



(c) Proportion of available partnerships which are offered by infectious individuals $\boldsymbol{C}_{\mathcal{I}}$

(d) Overall incidence λ

Figure 7: Incidence and the dynamic factors of incidence versus turnover. The product of components (a) and (b) is proportional to (c) the proportion of total available contacts which are with infectious individuals and (d) overall incidence.

As shown in Figure 7a, turnover decreased the first sub-factor: $\hat{C}_{\mathcal{I}}$ the average contact rate among infectious individuals. This is because, as noted above, turnover causes in a net movement of infectious individuals from high to low risk. However, at low to moderate rates of turnover, turnover increased the second sub-factor: $\hat{\mathcal{I}}$ overall prevalence (Figure 7b, region A). Now, under the conditions in region A, overall prevalence $\hat{\mathcal{I}}$ increased faster with turnover than the average contact rate of infectious people $\hat{C}_{\mathcal{I}}$ decreased. Thus, as a product of these two sub-factors, $\boldsymbol{C}_{\mathcal{I}}$ increased with turnover in region A (Figure 7c) and incidence increased proportionally (Figure 7d).

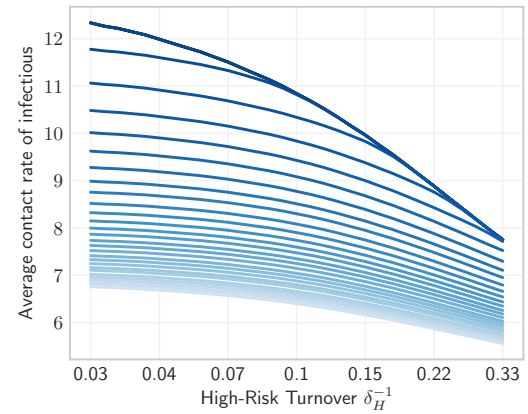
It therefore follows that the peak in incidence, and the transition between regions A and B in Figure 7c occurred when the dominating sub-factor between $\hat{\mathcal{I}}$ and $\hat{C}_{\mathcal{I}}$ reversed. That is, the transition occurred when the average contact rate of infectious people $\hat{C}_{\mathcal{I}}$ decreased with turnover faster than prevalence $\hat{\mathcal{I}}$ increased with turnover. Then, as rates of turnover continued to increase, declining incidence also reversed the upward trend in prevalence, and incidence and prevalence decreased across all groups in a snowball effect. This mechanism then explains the observations shown in region B throughout.

Finally, we note that the rates of turnover which maximized incidence (Figure 7d) were lower than those which maximized overall prevalence (Figure 7b). This is because incidence represents the number of new infections per susceptible, not per individual; so a system at slightly higher prevalence can be maintained by a slightly lower incidence, provided the proportion of susceptibles has decreased more than prevalence has increased. We also note that infection prevalence among the high risk group peaked at lower rates of turnover than prevalence among the low risk group. This is because turnover yields a net movement of infectious individuals from high to low risk; so prevalence among the high risk group can decrease with turnover even as incidence increases, whereas prevalence among the low risk group can increase with turnover even as incidence decreases.

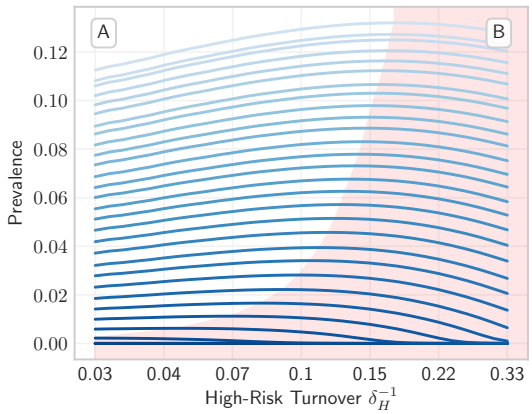
Experiment 1.3: Influence of turnover rates at various treatment rates. So far, the influence of turnover on equilibrium incidence and prevalence was explored for a single treatment rate. Experiment 1.3 explored additional treatment rates τ.

First, the sub-factors of incidence are again shown in Figure 8. Increasing the treatment rate τ increased the average contact rate of infectious individuals $\hat{C}_{\mathcal{I}}$ (Figure 8a). This is because increasing treatment concentrated infections in the highest risk group (Figure B.7), so that $\hat{C}_{\mathcal{I}}$, on average, increased. However, increasing treatment reduced overall prevalence $\hat{\mathcal{I}}$ (Figure 8b), due to the herd immunity effect of treated individuals. The dominant effect was that of $\hat{\mathcal{I}}$, which tended towards zero faster than $\hat{C}_{\mathcal{I}}$ tended towards infinity, and so incidence declined with treatment (Figure 8d).

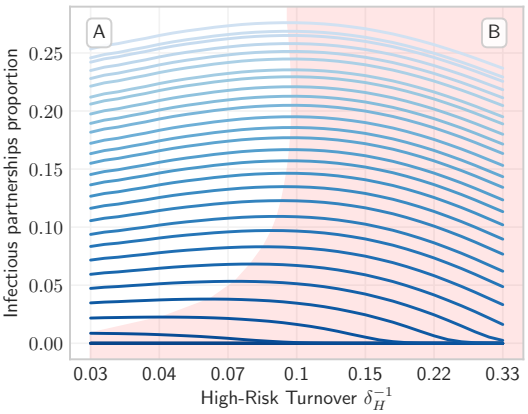
Figure 8d also shows that the rate of turnover which maximized incidence decreased with treatment. That is, as treatment rates increased, turnover was more likely to decrease incidence than it was to increase incidence (region B grew). This effect can be explained as follows. Recall that the mechanism by which



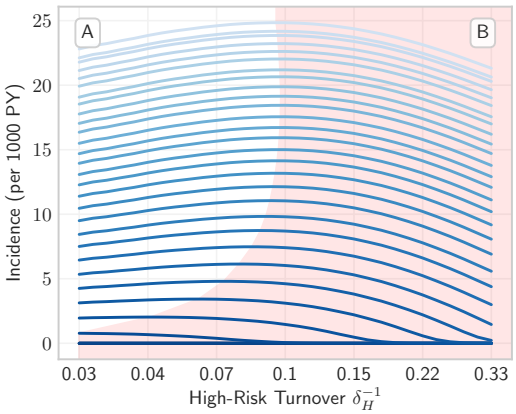
(a) Average contact rate among infectious individuals $\hat{C}_{\mathcal{I}}$



(b) Overall prevalence $\hat{\mathcal{I}}$



(c) Proportion of available partnerships which are offered by infectious individuals $\boldsymbol{C}_{\mathcal{I}}$

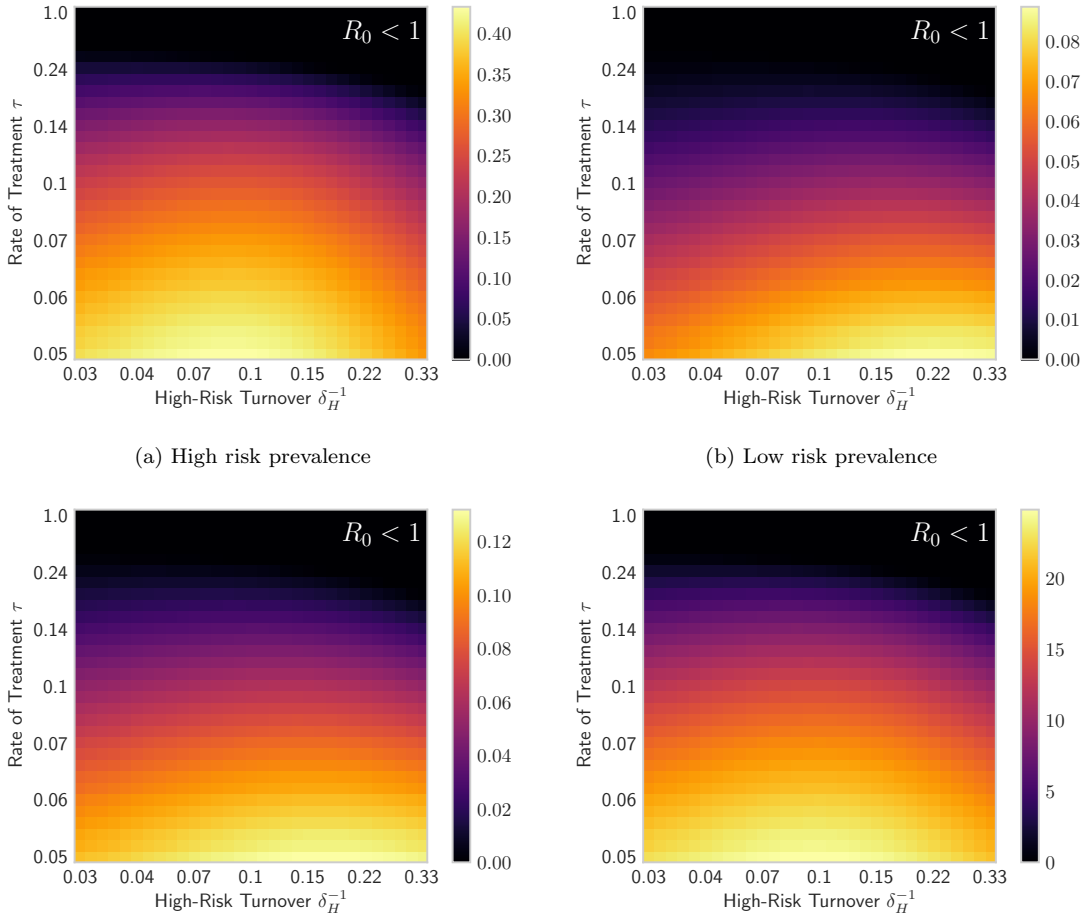


(d) Overall incidence λ

Figure 8: Incidence and the dynamic factors of incidence versus turnover, for a range of treatment rates. Darker blue indicates higher treatment rate. The product of components (a) and (b) is proportional to (c) the proportion of total available contacts which are with infectious individuals and (d) overall incidence.

turnover decreased incidence was through reduction of the average contact rate of infectious individuals $\hat{C}_{\mathcal{I}}$, due to net movement of infectious individuals from high to low risk (Figure 6). If treatment increased the concentration of infections in the high risk group, then the average contact rate of infectious individuals $\hat{C}_{\mathcal{I}}$ would be more sensitive to redistribution of those infectious individuals to lower risk groups via turnover. In Figure 8a, this is shown as the larger downward slope of $\hat{C}_{\mathcal{I}}$ versus turnover at higher treatment rates (darker blue). Therefore, at higher treatment rates, turnover was more likely to decrease incidence because movement of infectious individuals from high to low risk had a larger impact on the average contact rate of infectious individuals.

Finally, Figure 9 summarizes trends in overall equilibrium incidence and group-specific prevalence with



(a) High risk prevalence

(b) Low risk prevalence

(c) Overall prevalence

(d) Overall incidence

Figure 9: Equilibrium prevalence and incidence for different rates of turnover ϕ (log scale) and treatment τ (log scale).

respect to both turnover ϕ and treatment rate τ.[8] Treatment consistently decreased equilibrium incidence (as noted above), as well as prevalence, at all rates of turnover. As suggested in Experiment 1.2, infection prevalence increased with turnover for moderate rates of turnover and treatment, among all groups, and overall. However, for higher rates of turnover, prevalence among each risk group peaked, and then declined. As shown in Figure 8, the rate of turnover at which the peak occurs decreased with treatment rate. Finally, for high rates of treatment and/or turnover, the product of the average contact rate of infectious individuals $\hat{C}_{\mathcal{I}}$ and prevalence $\hat{\mathcal{I}}$ was too low to sustain the epidemic. That is, the basic reproductive number R_0 declined to less than one, and no epidemic was observed.[9]

[8] Figures 9c and 9d are the surface projections of the profiles shown in Figures 8b and 8d, respectively.

[9] In fact, it can be shown that for extreme rates of turnover, a heterogeneous system (e.g. Full model) will converge on a homogeneous system (e.g. model V1). This result is shown in Figure B.9.

Table 3: Equilibrium contact rates C and prevalence P among the high H and low L risk groups predicted by the models with and without turnover, before and after model fitting.

Context	C_H	C_L	C_H / C_L	P_H	P_L	P_H / P_L
No Turnover (Setting A)	25.0	1.0	25.0	22%	2%	**9.2**
Turnover (Setting B)	25.0	1.0	25.0	22%	3%	**6.7**
No Turnover [fit] (Model A)	23.5	1.6	**15.1**	20%	3%	6.6
Turnover [fit] (Model B)	24.3	1.0	**23.8**	20%	3%	6.7

4.2. Experiment 2: Inferred risk heterogeneity with vs without turnover

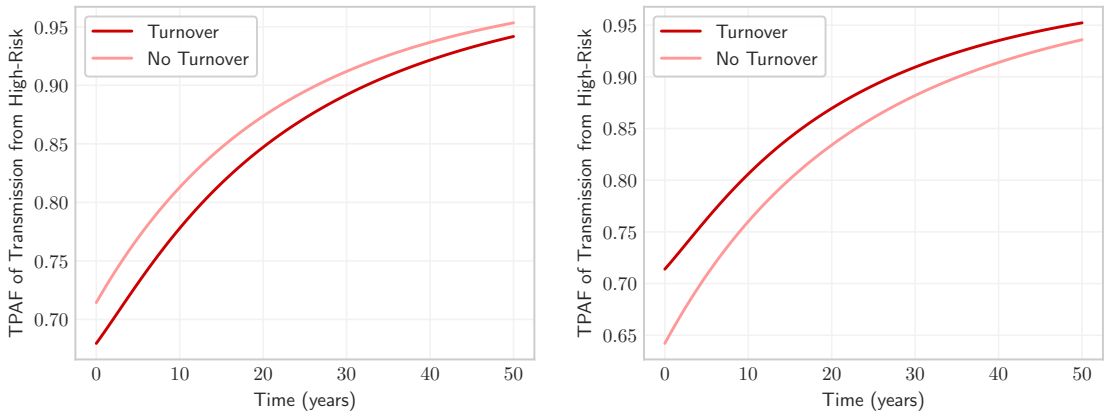
Before model fitting, the predicted prevalence ratio between high and low risk groups was lower with turnover than without: 6.7 vs 9.2. This reflects the "homogenizing" effect of turnover on the average risk experienced by individuals in the model. As shown in Figure B.7b, the high-to-low prevalence ratio consistently declined with turnover for all turnover and treatment rates explored. Thus, when fitting the model to target prevalence values, the fitted contact rates C would have to compensate for this difference.

After fitting the contact rates, both models predicted the desired equilibrium infection prevalence values of 20%, 3%, and 5% among the high risk group, low risk group, and overall. However, in order to do so, the ratio of fitted contact rates between high and low risk groups (C_H / C_L) was higher with turnover than without: 23.8 vs 15.1. That is, the inferred level of risk heterogeneity was higher in the model with turnover than in the model without turnover. This is because, in order to observe the same prevalence ratio in a system with turnover, the "risk homogenizing" effects of turnover must be overcome by greater heterogeneity in risk, as compared to a system without turnover. These results are also summarized in Table 3.

4.3. Experiment 3: Influence of turnover on the TPAF of the high risk group

Finally, we compared the predicted TPAF of the high risk group with and without turnover in: two settings (same parameters, different prevalence); and two models (same prevalence, different parameters). These results are shown in Figure 10. The TPAF approaches 1 for all models over the 50 year period, indicating that unmet treatment needs of the high risk group are central to epidemic persistence in all scenarios. Additionally, no TPAFs intersect during this period, so relative differences between TPAFs can be described irrespective of time horizon.

Experiment 3.1: Two settings. Figure 10a shows the estimated TPAF of the high risk group in two different settings – with and without turnover. In this case, the estimated TPAF is lower in Setting B with turnover versus in Setting A without turnover. This can be attributed to the larger equilibrium prevalence ratio without model fitting (Experiment 2, Table 3), which results in more onward transmission from the high



(a) No fitting

(b) Fitted contact rates

Figure 10: Transmission population attributable fraction (TPAF-from) of the high risk group with and without turnover, and with and without fitted contact rates to group-specific prevalence.

prevalence high risk group. In other words, the importance of reaching the high risk group in a context without turnover is higher than in a context with turnover, all other factors being equal.

Experiment 3.2: Two models. Figure 10b shows the estimated TPAF of the high risk group by two different models – with and without turnover – both fitted to the same prevalence. Now the estimated TPAF is higher in Model B with turnover versus in Model A without turnover. This result, opposite to Experiment 3.1, can be explained by two factors. First, the equilibrium prevalence ratios predicted by the models with and without turnover are equalized through model fitting to the same targets. As a result, differences in prevalence in the high risk group no longer contribute to an increased TPAF estimate by Model A without turnover. Second, as shown in Experiment 2, the ratio of fitted contact rates C_H / C_L in model B with turnover are higher than in Model A without. This affords a higher risk of onward transmission to the high risk group in Model B with turnover, and thus an increased TPAF. This result then implies that models which fail to capture turnover dynamics which are present in reality may underestimate the TPAF of high risk groups. Consequently, the importance of prioritizing high risk groups to achieve epidemic control may be similarly underestimated by such models.

5. Discussion

Epidemic models have rarely considered turnover of individuals among risk groups. In this work we have presented a new framework for modelling such turnover. This framework subsumes and generalizes previous approaches to this task, and clarifies requisite data and assumptions. We then used this framework to simulate turnover in an illustrative model of an STI epidemic, and explored how and why turnover influences

the following model outputs: group-specific equilibrium prevalence and incidence; inferred heterogeneity during model fitting; and projected TPAF of a high-risk group.

Turnover framework. Our framework uses a flexible system of linear equations to resolve the rates of turnover ϕ and the proportions of individuals entering into each risk group upon entry into the model $\hat{\boldsymbol{e}}$. This approach has several advantages. First, as shown in Section 2.3, this formulation subsumes previous implementations of turnover, and thereby hopefully clarifies relationships among them. Second, it facilitates parameterization of turnover among any number of risk groups in a scalable, flexible way. Finally, the framework makes the parameterization of turnover systems more transparent, through explicit incorporation of constraints as rows in the linear system. We have also summarized the major constraints which could be used, so that future modellers may choose among these based on available data and epidemic context. It is our hope that this framework will facilitate inclusion of turnover in future epidemic models.

The major data needs implicated by this approach are then as follows. First, the proportion of total individuals in each risk group $\hat{\boldsymbol{x}}$ is required. This can generally be obtained from representative surveys of the modelled population, and from key populations surveys for under-represented risk groups. Similarly, the risk behaviours of individuals immediately after entering the model can be used to specify elements in $\hat{\boldsymbol{e}}$. Such behaviours could be approximated using data from the youngest age group in the above surveys, or using surveys among young people specifically. Third, it is useful to know the average duration of individuals within each risk group $\boldsymbol{\delta}$. Among key populations, survey questions like "For how many years have you been a sex worker?" can be used to obtain estimates of duration in sex work (Watts et al., 2010). Among the general population, longitudinal data on individual-level changes in risk behaviour can be used to estimate duration in each risk group, although cross-sectional surveys unfortunately rarely capture such information. Finally, similar data, if available, can also be used to estimate the rates of transition between specific risk groups (ϕ).

Influence of turnover on incidence & prevalence. In Experiment 1, we showed that turnover can have a large influence on the equilibrium STI incidence and prevalence projected among different risk groups, as well as overall. In general, turnover yields a net movement of infectious individuals from high to low risk. In an analogy with heat transfer mechanics, incidence represents infection "conduction", whereas turnover represents infection "convection", and both processes represent important mechanisms of infection spread. As such, it would be important to simulate turnover in epidemic models if such dynamics are known or assumed to be present in reality.

Turnover also acts to homogenize the risk experienced by individuals in the model. Recall from core group theory that risk heterogeneity is sometimes necessary for persistence of an STI epidemic. Therefore, homogenization of risk through turnover could make epidemic control easier to achieve, compared to a setting with less or no turnover, as demonstrated by Henry and Koopman (2015). We corroborate this hypothesis in

Figure 9c, where we see that increasing turnover reduces the universal treatment rate τ required to achieve zero prevalence ($R_0 < 1$).

Implications for interventions. Mathematical models are often used to quantify the contribution of high-risk groups to overall transmission, in order to help prioritize interventions. For example, the transmission population attributable fraction (TPAF) represents the maximum proportion of infections which could be averted with a perfect intervention in one risk group. In Experiment 3.1, we showed that the TPAF of a high risk group was lower in a setting with turnover than in a setting without. This was attributable to a higher prevalence ratio in the absence of the homogenizing effects of turnover, resulting in more transmissions from the high risk group. Thus, reaching the high risk group with interventions would be more important in a setting without turnover.

However, in many epidemic models, the values of uncertain parameters are inferred by fitting to group-specific prevalence data before analysis of interventions. In Experiment 2, we showed that the heterogeneity in risk-associated parameters inferred via model fitting was higher in a model with turnover than in a model without. As a result, the TPAF of the high risk group was higher in the model with turnover than in the model without (Experiment 3.2). Therefore, if turnover dynamics present in reality are not simulated in the associated transmission model, the TPAF of high risk groups may be underestimated. That is, if risk group turnover is not simulated in fitted models, the impact of interventions prioritizing high risk groups may be systematically underestimated.

Limitations. There are several limitations to this work, and the results shown here are specific to the model structure and assumptions employed. For example, we assumed proportional contact formation across risk groups, whereas a degree of so-called "assortative mixing" by risk group is often assumed. We briefly explored the potential influence of assortative mixing on the results of Experiment 3 in Appendix B.7.1. We found that the TPAF of the high risk group would likely be further underestimated under assortative mixing if turnover is omitted from fitted models. We also did not consider disease-attributable mortality,[10] or reinfection. However, following similar analyses in Appendices B.7.2 and B.7.3, respectively, we found that inclusion of disease-attributable mortality or reinfection would also result in greater underestimation of high risk group TPAF by models without turnover. That is, when assortative mixing, disease-attributable mortality, or reinfection are considered, fitted models without turnover would likely underestimate the importance of reaching high risk groups with interventions even more than shown in the original Experiment 3 results.

[10] Since disease-attributable mortality disproportionately affects higher risk groups, risk groups may change size after including this dynamic in the model. The proposed turnover framework does not yet include mechanisms to re-balance such changes to group sizes. In fact, such re-balancing would imply the existence of "market" forces to maintain specific risk group sizes, and it's not clear if such forces exist, or if they dominate individual-level factors driving movement between risk groups.

We have also not explored: more complex or infection-specific natural disease history, which may include periods of reduced or increased infectivity; interactions between risk group turnover and age stratifications, which are likely important; or sensitivity of the results shown to other model parameters, such as the number and sizes of risk groups $\hat{\boldsymbol{x}}$, the probability of transmission β, and population entry and exit rates ν and μ. Finally, in examining the influence of turnover on STI prevalence and incidence, we have focused on systems at equilibrium, motivated by our analysis of TPAF. Turnover likely also has an important influence on temporal dynamics of infection transmission. For example, Henry and Koopman (2015) derived the analytical relationship between turnover and the basic reproduction number R_0, and showed that R_0 decreases monotonically with turnover. These limitations and gaps in our analysis thus represent interesting opportunities for future work.

References

Anderson, R. M. and May, R. M. (1991). Infectious diseases of humans: dynamics and control. *Infectious diseases of humans: dynamics and control.*

Boily, M. C. and Mâsse, B. (1997). Mathematical models of disease transmission: A precious tool for the study of sexually transmitted diseases. *Canadian Journal of Public Health*, 88(4):255–265.

Boily, M. C., Pickles, M., Alary, M., Baral, S., Blanchard, J., Moses, S., Vickerman, P., and Mishra, S. (2015). What really is a concentrated HIV epidemic and what does it mean for West and Central Africa? Insights from mathematical modeling. *Journal of Acquired Immune Deficiency Syndromes*, 68:S74–S82.

DataBank (2019). Population estimates and projections.

Eaton, J. W. and Hallett, T. B. (2014). Why the proportion of transmission during early-stage HIV infection does not predict the long-term impact of treatment on HIV incidence. *Proceedings of the National Academy of Sciences*, 111(45):16202–16207.

Garnett, G. P. and Anderson, R. M. (1994). Balancing sexual partnership in an age and activity stratified model of HIV transmission in heterosexual populations. *Mathematical Medicine and Biology*, 11(3):161–192.

Gesink, D. C., Sullivan, A. B., Miller, W. C., and Bernstein, K. T. (2011). Sexually transmitted disease core theory: Roles of person, place, and time. *American Journal of Epidemiology*, 174(1):81–89.

Henry, C. J. and Koopman, J. S. (2015). Strong influence of behavioral dynamics on the ability of testing and treating HIV to stop transmission. *Scientific Reports*, 5(1):9467.

Knight, J., Wang, L., Ma, H., Schwartz, S., Baral, S., and Mishra, S. (2019). The influence of risk group turnover in STI/HIV epidemics: mechanistic insights from transmission modeling. In *STI & HIV 2019 World Congress*, Vancouver, BC, Canada.

Koopman, J. S., Jacquez, J. A., Welch, G. W., Simon, C. P., Foxman, B., Pollock, S. M., Barth-Jones, D., Adams, A. L., and Lange, K. (1997). The role of early HIV infection in the spread of HIV through populations. *Journal of Acquired Immune Deficiency Syndromes*, 14(3):249–58.

Kraft, D. (1988). A software package for sequential quadratic programming. Technical Report DFVLR-FB 88-28, DLR German Aerospace Center — Institute for Flight Mechanics, Koln, Germany.

LAPACK (1992). LAPACK: Linear Algebra PACKage.

Lawson, C. L. and Hanson, R. J. (1995). *Solving least squares problems*, volume 15. SIAM.

Malthus, T. R. (1798). *An Essay on the Principle of Population.*

Mishra, S., Steen, R., Gerbase, A., Lo, Y. R., and Boily, M. C. (2012). Impact of High-Risk Sex and Focused Interventions in Heterosexual HIV Epidemics: A Systematic Review of Mathematical Models. *PLoS ONE*, 7(11):e50691.

Nold, A. (1980). Heterogeneity in disease-transmission modeling. *Mathematical Biosciences*, 52(3-4):227–240.

Stigum, H., Falck, W., and Magnus, P. (1994). The core group revisited: The effect of partner mixing and migration on the spread of gonorrhea, chlamydia, and HIV. *Mathematical Biosciences*, 120(1):1–23.

Watts, C., Zimmerman, C., Foss, A. M., Hossain, M., Cox, A., and Vickerman, P. (2010). Remodelling core group theory: the role of sustaining populations in HIV transmission. *Sexually Transmitted Infections*, 86(Suppl 3):iii85–iii92.

Yorke, J. A., Hethcote, H. W., and Nold, A. (1978). Dynamics and control of the transmission of gonorrhea. *Sexually Transmitted Diseases*, 5(2):51–56.

Zhang, X., Zhong, L., Romero-Severson, E., Alam, S. J., Henry, C. J., Volz, E. M., and Koopman, J. S. (2012). Episodic HIV Risk Behavior Can Greatly Amplify HIV Prevalence and the Fraction of Transmissions from Acute HIV Infection. *Statistical Communications in Infectious Diseases*, 4(1).

A. Supplemental Equations

Table A.1: Notation

Symbol	Definition
i	risk group index
j	risk group index for "other" group in turnover
k	risk group index for "other" group in incidence
$\mathcal{S}_i$	Number of susceptible individuals in risk group i
$\mathcal{I}_i$	Number of infectious individuals in risk group i
$\mathcal{T}_i$	Number of treated individuals in risk group i
N	total population size
ν	rate of population entry
μ	rate of population exit
ϕ_{ij}	rate of turnover from group i to group j
λ_i	force of infection among susceptibles in risk group i
τ	rate of treatment initiation among infected
$\hat{x}_i$	proportion of individuals in risk group i
$\hat{e}_i$	proportion of individuals entering into risk group i
δ_i	average duration spent in risk group i
C_i	contact rate among individuals in risk group i
β	probability of transmission per contact
ρ_{ik}	probability of contact formation between risk groups i and k

A.1. Model Equations

$$\frac{d}{dt}\mathcal{S}_i(t) = \sum_j \phi_{ji}\mathcal{S}_j(t) - \sum_j \phi_{ij}\mathcal{S}_i(t) - \mu\mathcal{S}_i(t) + \nu\hat{e}_i N(t) - \lambda_i(t)\mathcal{S}_i(t) \tag{A.1}$$

$$\frac{d}{dt}\mathcal{I}_i(t) = \sum_j \phi_{ji}\mathcal{I}_j(t) - \sum_j \phi_{ij}\mathcal{I}_i(t) - \mu\mathcal{I}_i(t) + \lambda_i(t)\mathcal{S}_i(t) - \tau\mathcal{I}_i(t) \tag{A.2}$$

$$\frac{d}{dt}\mathcal{T}_i(t) = \sum_j \phi_{ji}\mathcal{T}_j(t) - \sum_j \phi_{ij}\mathcal{T}_i(t) - \mu\mathcal{T}_i(t) + \tau\mathcal{I}_i(t) \tag{A.3}$$

A.2. Complete Example Turnover System

$$
\begin{array}{r}
\text{conservation of mass} \\ \\ \\
\text{entering distribution} \\ \\ \\
\text{group duration} \\ \\ \\
\text{balance turnover} \\ \\ \\
\end{array}
\left\{
\begin{array}{c}
\begin{bmatrix} \nu x_1 \\ \nu x_2 \\ \nu x_3 \\ e_1^* \\ e_2^* \\ e_3^* \\ {\delta_1}^{-1} - \mu \\ {\delta_2}^{-1} - \mu \\ {\delta_3}^{-1} - \mu \\ 0 \\ 0 \\ 0 \end{bmatrix}
\end{array}
\right.
=
\begin{bmatrix}
\nu & \cdot & \cdot & -x_1 & -x_1 & x_2 & \cdot & x_3 & \cdot \\
\cdot & \nu & \cdot & x_1 & \cdot & -x_2 & -x_2 & \cdot & x_3 \\
\cdot & \cdot & \nu & \cdot & x_1 & \cdot & x_2 & -x_3 & -x_3 \\
1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 \\
\cdot & \cdot & \cdot & x_1 & \cdot & -x_2 & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & x_1 & \cdot & \cdot & -x_3 & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & x_2 & \cdot & -x_3
\end{bmatrix}
\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \phi_{12} \\ \phi_{13} \\ \phi_{21} \\ \phi_{23} \\ \phi_{31} \\ \phi_{32} \end{bmatrix}
\tag{A.4}
$$

A.3. Factors of Incidence

Rearranging the force of infection λ_i to isolate the dynamic (not constant) component $(*)$:

$$
\begin{aligned}
\lambda_i &= C_i \sum_k \rho_{ik} \beta \frac{\mathcal{I}_k(t)}{\mathcal{X}_k} \\
&= C_i \beta \sum_k \frac{C_k \mathcal{X}_k}{\sum_{\mathrm{k}} C_{\mathrm{k}} \mathcal{X}_{\mathrm{k}}} \frac{\mathcal{I}_k(t)}{\mathcal{X}_k} \\
&= C_i \beta \underbrace{\frac{\sum_k C_k \mathcal{I}_k(t)}{\sum_{\mathrm{k}} C_{\mathrm{k}} \mathcal{X}_{\mathrm{k}}}}_{*}
\end{aligned}
\tag{A.5}
$$

This component $(*)$ is: ***$C_{\mathcal{I}}$ the proportion of available partnerships which are offered by infectious individuals.*** As the only dynamic component, only this component can be affected by turnover.

Now consider that $\boldsymbol{C}_{\mathcal{I}}$ can be written in terms of the following three factors:

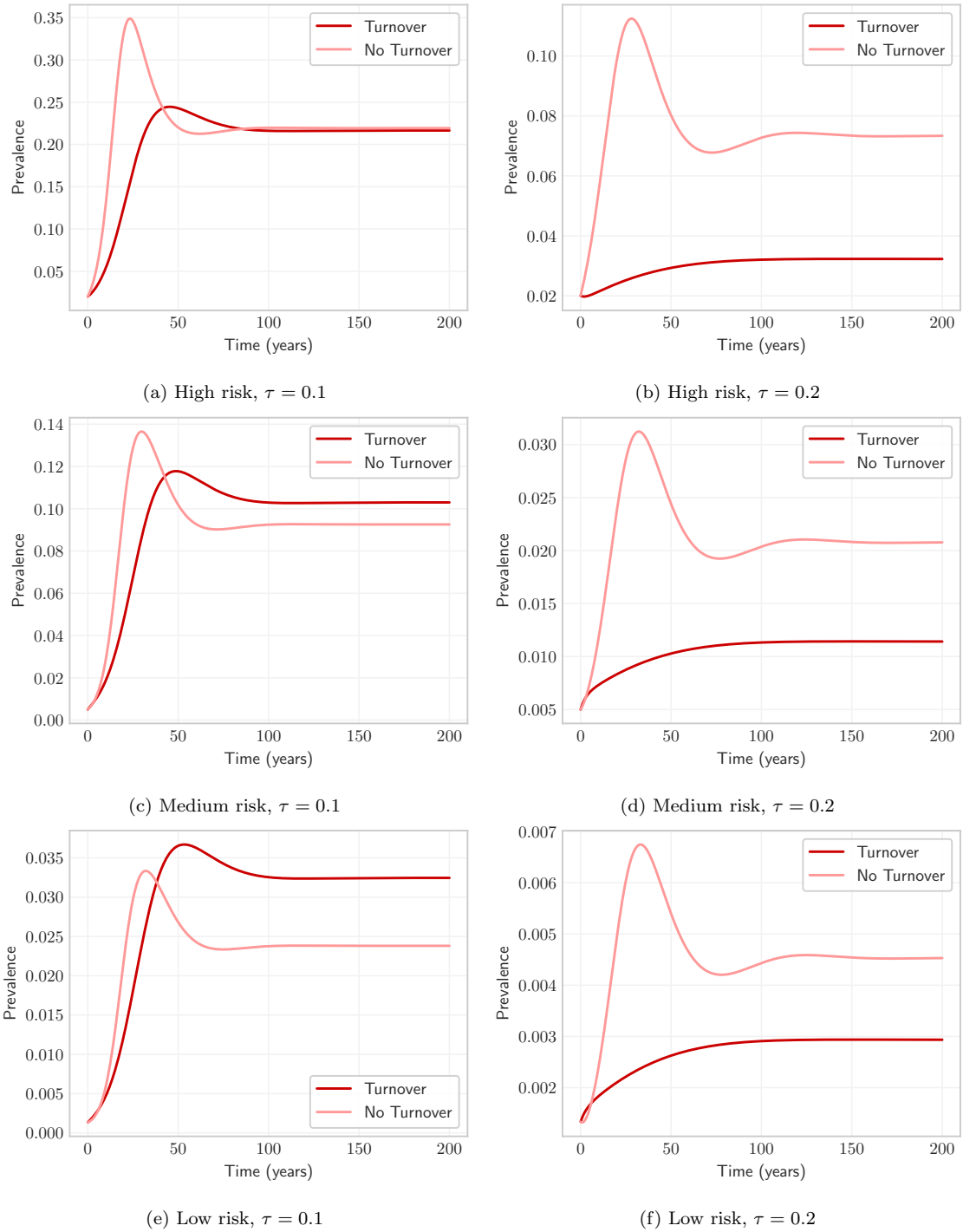
- The average contact rate among infectious individuals $\hat{C}_{\mathcal{I}} = \frac{\sum_k C_k \mathcal{I}_k}{\sum_k \mathcal{I}_k}$
- The proportion of the population who are infectious (prevalence) $\hat{\mathcal{I}} = \frac{\sum_k \mathcal{I}_k}{\sum_k \mathcal{X}_k}$
- The average contact rate among all individuals (constant) $\hat{C} = \frac{\sum_k C_k \mathcal{X}_k}{\sum_k \mathcal{X}_k}$

$$
\begin{aligned}
\boldsymbol{C}_{\mathcal{I}} &= \hat{C}_{\mathcal{I}} \times \hat{\mathcal{I}} \times \hat{C}^{-1} \\
&= \frac{\sum_k C_k \mathcal{I}_k}{\sum_k \mathcal{I}_k} \times \frac{\sum_k \mathcal{I}_k}{\sum_k \mathcal{X}_k} \times \frac{\sum_k \mathcal{X}_k}{\sum_k C_k \mathcal{X}_k} \\
&= \frac{\sum_k C_k \mathcal{I}_k}{\sum_{\mathrm{k}} C_{\mathrm{k}} \mathcal{X}_{\mathrm{k}}}
\end{aligned}
\tag{A.6}
$$

Therefore, actually only two dynamic factors control the force of infection: 1) the average contact rate among infectious individuals $\hat{C}_{\mathcal{I}}$, and 2) the proportion of the population who are infectious $\hat{\mathcal{I}}$; and the product of these factors (scaled by $\hat{C}^{-1}$) gives $\boldsymbol{C}_{\mathcal{I}}$. Overall incidence is then directly proportional to $\boldsymbol{C}_{\mathcal{I}}$, following Eq. (A.5). In fact, the incidence in each group individually is proportional to $\boldsymbol{C}_{\mathcal{I}}$, as C_i is only factor depending on i.

B. Supplemental Results

B.1. Prevalence with and without turnover



(a) High risk, $\tau = 0.1$

(b) High risk, $\tau = 0.2$

(c) Medium risk, $\tau = 0.1$

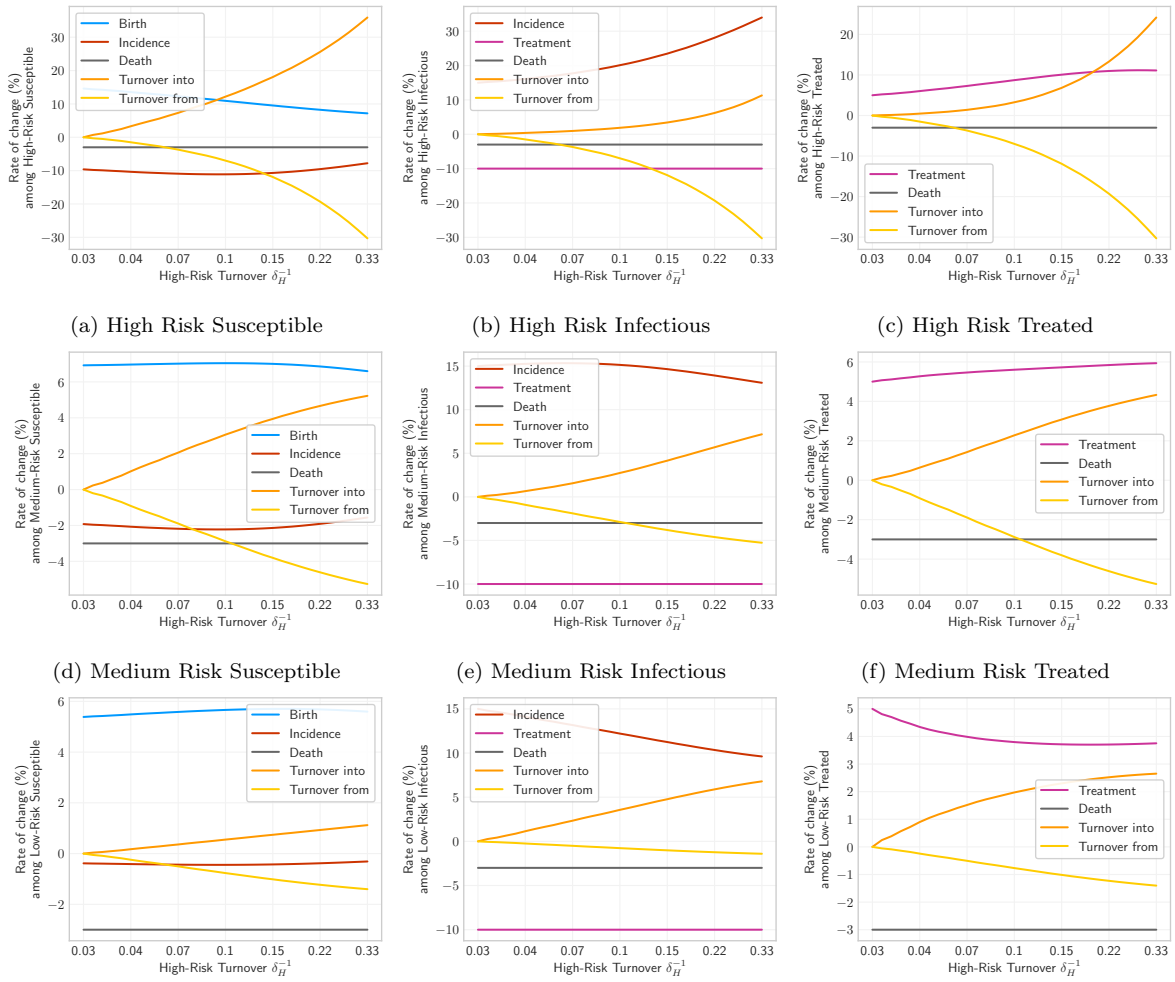
(d) Medium risk, $\tau = 0.2$

(e) Low risk, $\tau = 0.1$

(f) Low risk, $\tau = 0.2$

Figure B.1: Comparison of projected prevalence for each group with and without risk group turnover under two different treatment rates τ. Note that differences in initial trajectories are attributable arbitrary initialization of health states.

B.2. Rates of transition at equilibrium



(a) High Risk Susceptible (b) High Risk Infectious (c) High Risk Treated

(d) Medium Risk Susceptible (e) Medium Risk Infectious (f) Medium Risk Treated

(g) Low Risk Susceptible (h) Low Risk Infectious (i) Low Risk Treated

Figure B.2: Rates of transition among risk groups and health states at equilibrium. All transition rates are shown relative to the named group, for both afferent and efferent transitions. Note that, although each system is at equilibrium, profiles should not sum to zero due population growth.

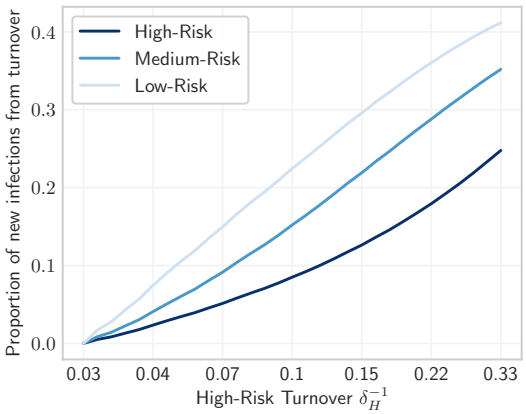
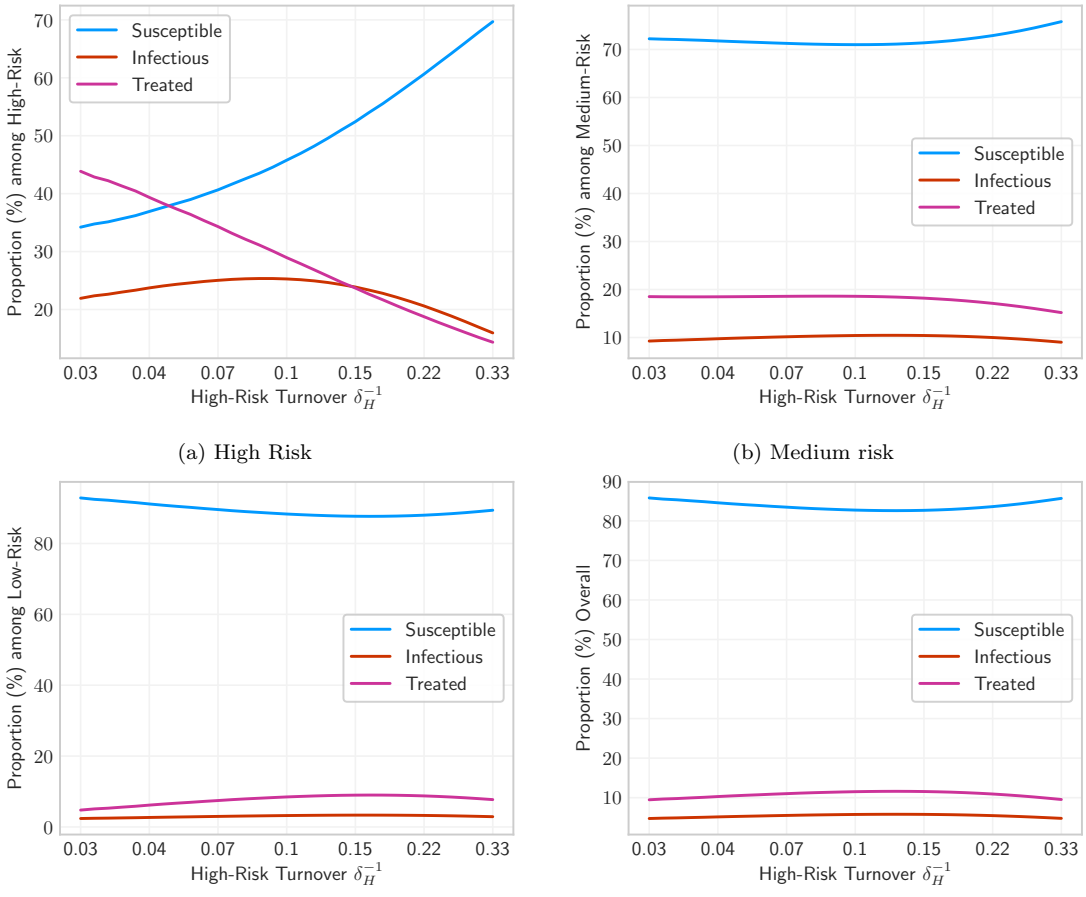


Figure B.3: Proportion of new infectious individuals in each risk group which are from turnover of infectious individuals, as opposed to infection of susceptible individuals in the risk group ($\tau = 0.1$).

B.3. Distribution of health states at equilibrium



(a) High Risk

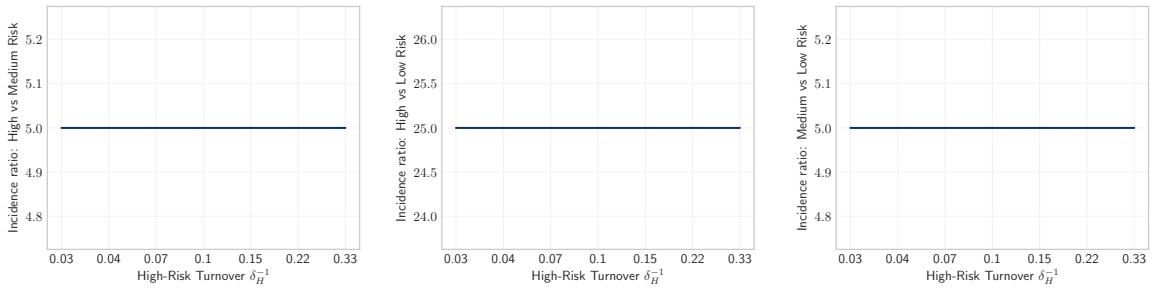
(b) Medium risk

(c) Low Risk

(d) Overall

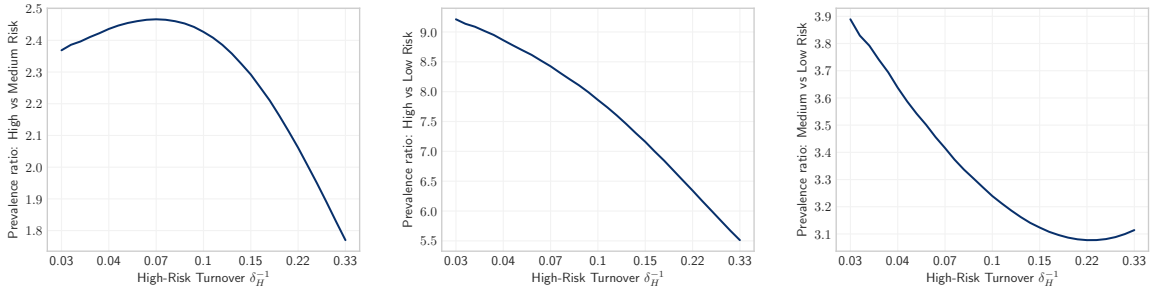
Figure B.4: Equilibrium incidence ratios between risk groups under different rates of turnover.

B.4. Equilibrium Incidence and Prevalence Ratios



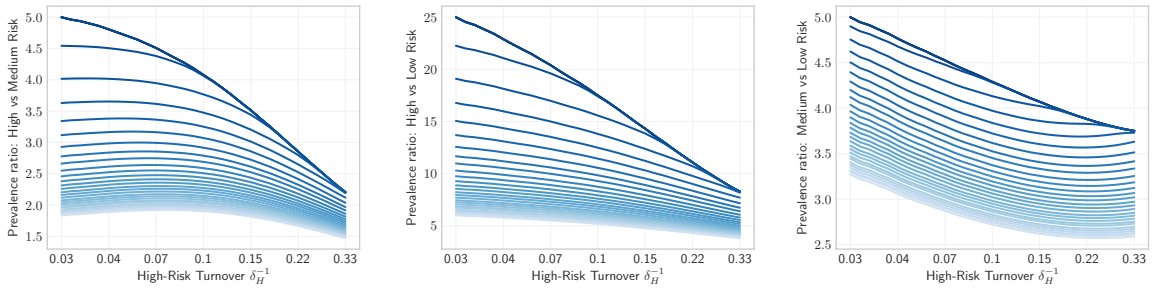
(a) High vs Medium Risk (b) High vs Low risk (c) Medium vs Low Risk

Figure B.5: Equilibrium incidence ratios between risk groups under different rates of turnover ϕ for a treatment rate $\tau = 0.1$. Incidence ratios do not depend on turnover nor treatment – see Eq. (18).



(a) High vs Medium Risk (b) High vs Low risk (c) Medium vs Low Risk

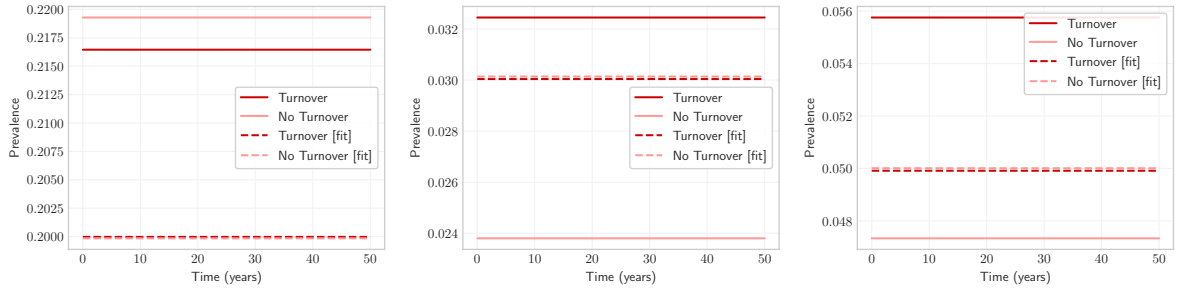
Figure B.6: Equilibrium prevalence ratios between risk groups under different rates of turnover ϕ for a treatment rate $\tau = 0.1$.



(a) High vs Medium Risk (b) High vs Low risk (c) Medium vs Low Risk

Figure B.7: Equilibrium prevalence ratios between risk groups under different rates of turnover ϕ and treatment τ. Turnover increases left to right; treatment increases light to dark.

B.5. *Equilibrium prevalence before and after model fitting*



(a) High risk (b) Low risk (c) Overall

Figure B.8: Equilibrium prevalence among high and low risk groups as well as overall, with and without turnover, and with and without fitted C_i to group-specific prevalence.

B.6. *Extreme turnover converges on a homogeneous system*

TODO: update

Figure B.9: Overall prevalence predicted by a heterogeneous system for a range of high turnover rates. Note how the heterogeneous model ($G = 3$) converges on a homogeneous model ($G = 1$) with very high turnover rates. Compared to the Base model, transmission probability is increased to $\beta = [TBD]$ in order to yield non-zero equilibrium prevalence in the homogeneous model.

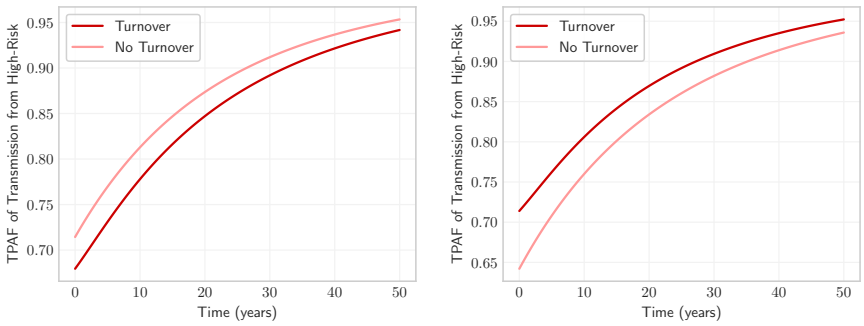
B.7. *"What if" analysis*

This section presents hypotheses and preliminary analysis regarding the potential impacts of alternative model assumptions on the results of Experiment 3. For reference, Figures B.10a and B.10b show the original results of Experiment 3 (copied from the main text).

B.7.1. *Assortative mixing*

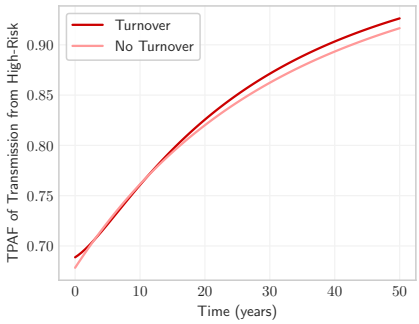
Assortative mixing increases the concentration of transmission within the high risk group, decreasing the TPAF, regardless of turnover. However, turnover counteracts this effect by redistributing infectious individuals from high to low risk, permitting greater onward transmission from infectious formerly high risk individuals. Therefore, while turnover may influence the TPAF of the high risk group through the STI prevalence before model fitting, and the inferred risk heterogeneity after model fitting, we hypothesize that the secondary effect of turnover to circumvent assortative mixing would increase the TPAF in both cases.

To explore this hypothesis, we re-ran Experiment 3 under an assortative mixing scenario, with $\epsilon = 0.5$ (Nold, 1980). Before model fitting and under assortative mixing (Figure B.10c), the TPAF of the high risk group with turnover is similar to the TPAF without turnover, whereas under proportional mixing it was lower with turnover than without turnover (Figure B.10a). After model fitting and under assortative mixing (Figure B.10d), the TPAF of the high risk group is even higher with turnover than without turnover,

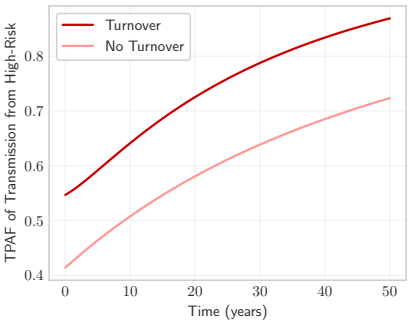


(a) No fitting, original results

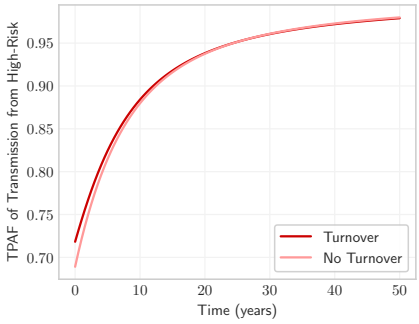
(b) Fitted, original results



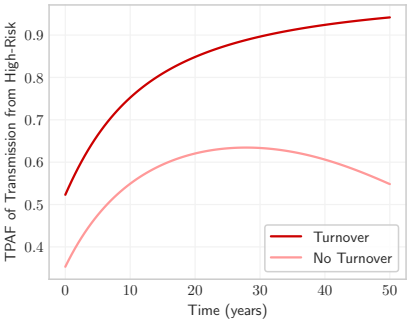
(c) No fitting, assortative mixing



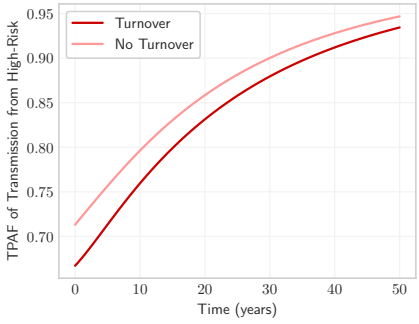
(d) Fitted, assortative mixing



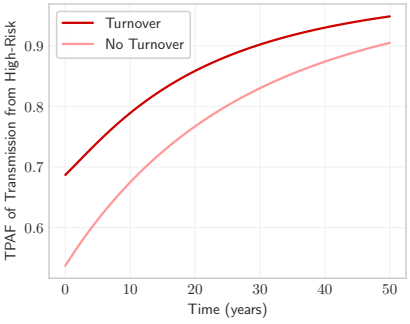
(e) No fitting, disease-attributable mortality



(f) Fitted, disease-attributable mortality



(g) No fitting, reinfection



(h) Fitted, reinfection

Figure B.10: TPAF of the high risk group with and without turnover, and with and without fitted contact rates to group-specific prevalence. Comparison of original results, under proportionate mixing, with no disease-attributable mortality, and without reinfection (a, b), with models considering: assortative mixing (c, d), disease-attributable mortality (e, f), and reinfection (g, h).

as compared to proportional mixing (Figure B.10b). Both of these results support our hypothesis above, though more comprehensive experiments are needed.

B.7.2. Infection-attributable mortality

Disease-attributable mortality disproportionately affects the highest risk group due to higher infection prevalence. As a result, the size of the group decreases,[11] which reduces the TPAF of the group. Similar to assortative mixing, this reduction in TPAF is again counteracted by turnover, through re-supply of individuals from lower-risk groups. Thus, we hypothesize that while disease-attributable mortality would decrease the TPAF of the high risk group, this effect is reduced by turnover, both before and after model fitting.

To explore this assertion, we again re-ran Experiment 3, this time assuming a disease-attributable mortality rate of 0.2 per year among untreated infected individuals $\mathcal{I}$. Since establishment of the epidemic was more difficult under these conditions, we also doubled the probability of transmission per contact β from 0.03 to 0.06. As with assortative mixing, before model fitting and when disease-attributable mortality is included (Figure B.10e), the TPAFs with and without turnover are more similar than when no disease-attributable mortality is considered (Figure B.10a). Similarly, after model fitting and when disease attributable mortality is included (Figure B.10f), the TPAF of the high risk group is even higher with turnover than without turnover, as compared to when no disease-attributable mortality is considered (Figure B.10b). Again, these results support our hypothesis.

B.7.3. Reinfection

TODO: update; Figures OK

[11] In this case, we assume that no "market" forces maintain risk group proportions against changes due to unequal death.