Tutorial A

Introduction to TranStat

A.0.a Learning objective(s):

- To understand the capabilities of TranStat
- To understand the types of analyses that can be conducted with TranStat
- To understand the data requirements for TranStat

A.0.b Take away messages:

- TranStat is a software application designed to estimate the transmissibility of infections within groups of individuals.
- Given the availability of a minimal set of information (group membership, infection/disease status, and disease onset time) for each member of sufficiently large groups of individuals within which an outbreak is observed, TranStat should be able to estimate the probability that group members will transmit infection to each other and the competing risk that group members were infected by exposure to sources of infection located outside of the groups included in the study population.
- TranStat will allow the user to account for the level of existing immunity to the infection of interest.
- As with all standard statistical analysis packages, TranStat allows for estimation of the effects of individual- and group- level characteristics on the level of existing immunity, on the transmission of infection, and on the risk of developing clinically-evident disease once infected.
- TranStat provides analysis results in a very simple format, listing for each model parameter and covariate effect: the estimate, standard error, lower bound of the 95% confidence interval, and the upper bound of the 95% confidence interval. For covariates effects, TranStat additionally provides the p-value from a hypothesis test that the estimated covariate effect is not equal to the null odds ratio of 1.0.
- TranStat will also allow the analyst to implement statistically-appropriate methods and models to account for missing information related to existing immunity, infection and disease status, and the onset times for disease.

A.1. Background

Infectious pathogens continue to contribute significantly to human and animal morbidity and mortality. Understanding how well a pathogen transmits and the determinates of its transmission potential are important for ultimately designing interventions and programs to control and/or prevent morbidity and mortality attributed to this infectious agent. TranStat is a freely-available software application that is designed to estimate the transmission potential of a pathogen from epidemiologic data collected within clusters of infected/diseased individuals (cases) and their contacts. In addition, TranStat estimates the effects of covariates on transmission parameters and is capable of sophisticate imputation for missing information related to susceptibility, infection, disease, and infectiousness.

TranStat is designed to analyze data from a large variety of study designs. The following list of study designs that can be analyzed by TranStat is by no means comprehensive. In general, TranStat can analyze data from any study design targeting clusters of individuals with an outcome or outcomes that is/are known or suspected to be caused by an agent that transmits between the individuals, either directly or through a vector. We will focus on epidemiologic designs for the investigation of the transmission of infectious diseases and its mitigation. These designs include group-randomized intervention trials and individually-randomized studies that involve the enrollment of individuals who interact with other enrollees. Data from non-randomized prospectively and retrospectively followed cohort studies enrolling

groups of individuals who interact with other enrollees may also be analyzed using TranStat. The final and possibly most abundantly available form of data involves a group of studies that we classify as being of the case-ascertained design (Yang et al. 2006). This latter group of studies that have many different titles, including outbreak investigations, case-contact tracing studies, and index-cluster surveys, but all of which have a central feature. A case of the infection or disease of interest is brought to the attention of the public health community through its detection by some form of surveillance system (hence case-ascertained), and then the decision is made to more intensively study that case and any additional cases that may have occurred among associates, friends, family, and other individuals that s/he may have interacted with or exposed to infection.

Before we learn more about the specific capabilities of TranStat, as well as how to use the software, we will spend some time learning about the general conceptual frameworks that underlie the statistical models implemented in TranStat. Specifically, we will discuss frameworks for the natural history and transmission of infection and for the structure of the population of individual hosts where this transmission occurs.

A.1.1. General framework for infection, disease, and transmission

Figure A.1 illustrates the general framework for the natural history of infection for an average member of the general population who is susceptible to infection (blue timeline). Upon infection after exposure to an infectious individual (brown timeline in Figure A.1), two parallel processes occur within the newly infected individual. The infection process is directly linked to a pathogen's need to transmit to other hosts. The pathogenic process can augment the transmission potential of a pathogen, but is often also linked to the host immune response. Here we illustrate the full versions of each of these processes, but, in theory, an infected individual could experience all or only the first part of each process, for example, infection without onset of infectiousness while still experiencing symptoms.

As part of the infection process, the pathogen will invade the target tissue(s) of the body and begin to replicate. After a period of time (the latent period), sufficient replication has occurred such that the infected individual begins to shed pathogen (onset of infectiousness). Often shedding occurs for a defined finite period of time, referred to as the infectious period. The level or degree of shedding may vary over the course of the infectious period. Transmission to other hosts can occur as the result of exposure to an infected individual during its infectious period. In the case of pathogens that depend upon an intermediate vector for host-to-host transmission, such as mosquitoes for dengue and malaria or contaminated water for cholera, the duration of time that the pathogen remains viable while transiting through the vector must be factored into the distribution of the infectious period.

At a certain point during the ongoing replication of the pathogen in the infected individual, the effects of infection itself and/or the host immune response can lead to the development of clinically apparent symptoms (for example, fever, cough, and sore throat with influenza infection). The appearance, or onset, of symptoms is a key event in the pathogenic process, and the period of time from infection to onset of symptoms is referred to as the incubation period. Symptoms tend to last for a finite period of time. The onset of symptoms may occur before or after (situation shown in Figure A.1) the onset of infectiousness, and symptoms may continue beyond the duration of infectiousness or end while the host is still infectious.

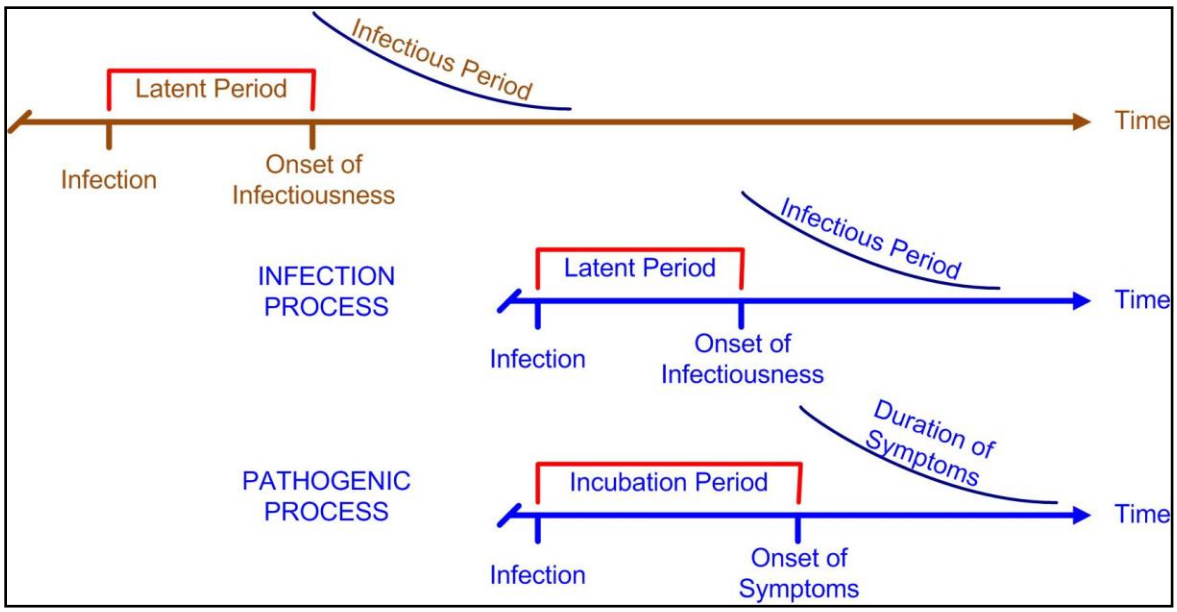

Figure A.1. Conceptual diagram for the natural history of infection.

Figure A.1 also illustrates the basic conceptual model for host-to-host transmission. To be considered at-risk for possible infection (*i.e.*, susceptible), an individual (here illustrated by the blue timeline) must be sufficiently exposed to the pathogen being shedding by another infectious host (illustrated by the brown timeline). Much of the analytic work done with TranStat aims to estimate the probability that a susceptible individual will become infected as a result of this type of exposure.

Therefore, the next topic of interest is how to define an exposure. To reduce the possibility of confusion with regard to the use of the term exposure, we will use the term 'contact' *in lieu* of exposure. For the purposes of our discussion, we will define a contact as the occurrence of a distinct event during which there is a reasonable probability that an infectious dose of pathogen could have been transferred into the body of a susceptible individual from an infectious host. The generalized nature of this definition allows for its application to a diverse set of pathogen-host systems. For influenza, a virus that relies primarily on transmission via respiratory droplets and aerosols, a contact might be defined as being within 3 feet of an infectious individual without any physical barrier separating the two individuals. This definition of a contact for influenza transmission represents a 'distinct event' where respiratory droplets ('infectious dose') exhaled from the infectious individual could be reasonable expected to reach the lungs or nasal mucosa of the susceptible individual with a non-zero probability. Another example of a contact might be the possible introduction of dengue infection to a susceptible child via a mosquito that was infected by an infectious adult. Our general definition of a contact should also permit the application of TranStat to non-human pathogen-host systems.

A.1.2. General framework for the contact structure within the host population

Now that we have established a general framework for the natural history of infection within individuals and for how a pathogen is physically transmitted, we need to establish a model that describes the population and contact structures of the host individuals. The statistical models implemented by TranStat conceive of hosts self-organizing into the groups or clusters of individuals who are more likely to experience a contact with each other than with other individuals in the population. Individual hosts could be members of more than one cluster and there could be more than one type of cluster within a population. For infections that are transmitted through various modes, the type of contact could be further defined by the mode of

transmission. If the complete contact structure is observed, then TranStat can estimate the probability of individual-to-individual transmission between hosts for each type of contact.

Rarely does a study observe the entire contact structure of a population of hosts. Instead, a selection of the clusters in a population are either chosen by investigators for longitudinal follow-up (prospective studies) or are ascertained and evaluated retrospectively, typically with regard to the first onset of infection or disease among the members of a cluster (case-ascertained studies). Examples of these two general types of studies include prospective cohort studies of households for respiratory viral infection and a retrospective investigation of an outbreak of foodborne illness, respectively. For both of these types of studies, the complete contact structure is not observed. Specifically, the unobserved portion of the contact structure is the contacts between the members of the clusters enrolled in the study and other members of the larger population, as well as contacts among members unobserved clusters. If we want to obtained unbiased estimates of transmission probabilities, the possible transmission of infection from infectious members of the unobserved proportion of the population to susceptible members of the enrolled clusters can not be ignored. Therefore, we estimate a general community-to-person (*i.e.*, host) infection probability (CPI), which basically represents the risk of infection due to exposure to sources of infection which are not represented by members of the clusters enrolled in the study.

Figure A.2 illustrates our general model for the contact structure for a population of hosts under the scenarios of complete and incomplete observation of the contact structure. Our model for the contact structure of a population is quite flexible, permitting the specification of multiple types of host clusters, between cluster contacts, and CPI's. Since the case-ascertained studies represent a large proportion of the data collected about infectious diseases occurring in populations (for example, outbreak investigations and contact-tracing studies), Section A.1.4 discusses mitigation of selection/ascertainment bias that certainly is present when trying to obtain unbiased estimates of the transmission potential of an pathogen within the general population, using data from a case-ascertained study.

For the purposes of our ongoing discussion, *person-to-person* (P2P) contact will be an additional term used to refer to contacts between individual hosts enrolled in the study. The terms, *common-source* and *community-to-person* (C2P), will be used to reference exposure to the sources of infection located in the larger community within which the study hosts are embedded.

A.1.3. The basic statistical model implemented by TranStat

Here we provide a brief description of the general statistical model implemented by TranStat (see Yang et al. 2006 for a more complete description of the model). TranStat employs maximum likelihood estimation to fit a transmission model data from the hosts enrolled in a study. These data include the membership of clusters, the contact structure, the infection and/or pathogenicity status of each host, the timing of the onset of symptoms or infectiousness among infected hosts, and the duration of time that each enrolled host was followed (prospectively and/or retrospectively) by study investigators. In addition, this model can estimate the level of association between individual and group-level covariates and P2P and C2P transmission probabilities. It is assumed that the user can specify plausible or known population-level distributions for the incubation, latent, and infectious periods. Users must also be able to fully specify the contact structure between members of hosts enrolled in the study.

A.1.4. Selection and ascertainment bias in case-ascertained studies

In retrospective investigations of an outbreak of an infectious pathogen, the inclusion of a cluster of individual hosts typically depends 1) on the occurrence of infection or resulting pathogenic disease within at-least one member of the cluster and 2) upon this individual being detected by a surveillance system. For case-ascertained studies, we will refer to the individual

member(s) of an enrolled cluster who were detected via surveillance (*i.e.*, the reason for the enrollment of the cluster into the study) as the index case(s). In humans, the complex nature of the process defining whether an infected/symptomatic individual will be detected by surveillance system can lead to selection or ascertainment bias with regard to factors that are associated both with the probability of enrollment and the risk of transmission. For example, members of the clusters enrolled in a case-ascertained study may have a higher pre-disposition than the general population for seeking healthcare, which could mean that the index cases of these clusters are, on average, in better (or worse) health than otherwise similar members of the general population. An index case's general health may affect his/her level of infectiousness, which in turn should affect the probability of P2P transmission. It is generally difficult to fully characterize the nature of this potential selection bias, which makes it difficult to directly model the ascertainment process. Therefore, TranStat can implement a statistical model that only considers the risk of transmission to non-index members of the enrolled clusters, thereby accounting for the effect of this selection bias. For the statistically-minded reader, this model conditions the index-case members of the enrolled clusters out of the likelihood for the individual-to-individual transmission probabilities and the CPI's. Please see Yang et al. 2006 for more details about this aspect of the transmission models implemented in TranStat.

A.1.5. Additional features implemented in version 5 of TranStat

Version 5 of TranStat implements methods for estimating the pathogenicity proportion for an infection (the risk of developing clinically apparent symptoms, given infection). This version also permits the estimation of the proportion of the members of the study population who were not susceptible to infection at the start of the study's follow-up period (for example, immune). TranStat 5 also implements Expectation Maximization (EM) and Monte Carlo EM algorithms that are designed to appropriately account for missing information about the susceptibility, infection, and/or symptomatic status of an enrolled host, as well as the onset time of infectiousness or symptoms for infected and symptomatic individuals, respectively. TranStat is also capable of accounting for right-censoring in the observation of the study outcome, whether that is infection or symptomatic infection.

A.1.6. Additional uses for TranStat

TranStat is designed for the analysis of infectious disease transmission, but this software could be used to for the analysis of any phenomenon or outcome which is 'transmitted', *i.e.*, moves from one distinct entity to another, within clusters of entities. Some examples of non-infectious disease related research topics for which TranStat may be useful include the transmission of addictive behaviors or socially-volatile ideas/ideologies within social mixing groups. The definitions for a 'contact' and for the incubation, latent, and infectious periods would have to be altered to fit these novel contexts.

A.2. Data requirements for TranStat

TranStat implements a discrete time model for transmission. All inputs that are indexed by time must be defined on an integer scale. TranStat will not translate dates to an integer scale. The user must preprocess his/her data to conform TranStat's input requirements. Typically, input values are indexed on a time-scale that is referenced (*i.e.*, time unit = 0) to a time point that occurs prior to the earliest onset date recorded in the study data. For the purposes of reducing the influence of potential biases, it is considered good practice to reference the time scale to a time unit that occurs at least the following amount of time before the earliest onset date in the study data: the sum of the maximum lengths of the latent and infectious periods.

A.2.1. Data required for individual hosts

- cluster membership
- any information or assumptions about whether the individual host was already immune to infection at the beginning of the study follow-up period
- infection status
  If infected,
  - Whether a host was the first infected member of the cluster that was detected by a surveillance system, also known as an index case (applies to studies of a case-ascertained design)
  - pathogenicity status (for example, symptomatic versus asymptomatic)
  - onset time for symptoms or infectiousness. Since the latter is rarely known, the former often serves as a proxy measure.
  - Outcome and date of this outcome for those with clinical disease (symptomatic)
- Other individual-level characteristics that may be associated with susceptibility to infection (including immunity status at the start of the study follow-up period) and/or the infectiousness of an infected individual. Covariate values may be indexed to a specific time unit or set of units or they may be indexed to the entire study follow-up period (*i.e.*, time independent).
- Available information about the individual-to-individual contact structure within each cluster and between clusters (if relevant), as well as data on the level of exposure to the general common sources of infection located outside of the clusters of hosts enrolled in the study. As a quick reminder, these two types of contact are also referred to as P2P and C2P contacts respectively. TranStat is quite flexible in its ability to accommodate multiple types of P2P contact structures, including, for example, structures that are dynamic or static over the course of the study follow-up period. The same applies for the nature of exposure to community sources of infection (C2P contacts). Increased flexibility is available through the ability to specify host-specific P2P and C2P contact histories. If there are multiple types of either P2P or C2P contacts, then the type must be specified for each contact.
- TranStat is capable of implementing data augmentation procedures to account for certain types of missing data related to the outcome of interest. If data is missing regarding whether a host was 1) immune at the beginning of the study follow-up period, 2) infected during the study follow-up period, and/or 3) developed symptomatic disease after infection, the user will be required to construct a separate file that indicates, by individual host, which information is missing. TranStat will then apply the implemented data augmentation procedures to account for additional uncertainty (reflected in the size of standard errors for model estimates) and bias associated with this missing data. TranStat is also capable of conducting the same data augmentation procedures for missing information regarding the onset time for symptoms/infectiousness.

A.2.2. Data required for cluster

We need to know the time units corresponding the beginning and end of the period of time during which a cluster of hosts was followed for study data collection.

A.3. Installing TranStat 5 on your computer

To install TranStat version 5 on your personal computer, you can either use the already compiled binaries for MS Windows machines or those for Mac OSX machines (binaries available for 10.1, 10.2, 10.3 and later, and for 10.10.4).

A.3.1. Installing TranStat 5 using the binaries.
Download the appropriate binary executable file from http://www.cidid.org/transtat/?rq=transtat. Microsoft Windows and Mac OS X versions of the latest version of TranStat are provided.

A.3.2 Compiling the TranStat 5 executable
You will need to download the latest version of the TranStat source code from the Center for Inference and Dynamics of Infectious Diseases (http://www.cidid.org/transtat/?rq=transtat). Uncompress/Unzip the contents of the source file archive into a directory of choice on your machine, which we will hereafter refer to as the *TranStat source folder*.

A.3.2.1 Compiling in Microsoft Window

A.3.2.1.1 Required Software: MINGW32 (or equivalent Linux/Unix emulators for MS Windows) with *gcc* and associated libraries.

A.3.2.1.2 Creating the necessary folder structure and setup files:

a. Open the MINGW32 command prompt.

b. Navigate to the TranStat source folder using the command prompt.

c. Type “runc main.c main.exe” into the MINGW32 command prompt. Ignore the warning messages that appear in the command window. These message should pertain to the redefinition by TranStat of the value for the INFINITY.

Note: If the computer that you are using to compile this TranStat source code contains a Pentium 3 or earlier generation processor, you may be required to perform the following step before the command above will work.

Use Windows Explorer to navigate to the *runc* file in the *bin* folder, and open it with your favorite text editor.

Alter the contents of the *runc* file in the following manner:

- The contents of this file should look similar to the following:

  gcc -O3 -march=***pentium4*** -g -lm $LINK_CNL $CFLAGS $1 -o $2

- Replace the portion shown in bold italics with the type of processer that your machine contains. Consult (https://gcc.gnu.org/onlinedocs/gcc/x86-Options.html#x86-Options) for more information.

c. The TranStat executable should compile as main.exe, which you can copy to any folder where you are storing analysis input files. If the main.exe file is created or is replaced (*i.e.*, if this is not your first time compiling the executable), ignore the warning messages delivered by Cygwin.

A.3.2.2. Compiling in Linux and Unix
Since MINGW32 is a Microsoft Windows emulator for Linux/Unix, the instructions given in Section A.3.1 can be followed to complete compilation of the TranStat main.exe in Linux or Unix.

A.3.2.3. Compiling in Mac OS X

NOTE: The Windows and Linux versions of TranStat has been extensively used and tested. Though the Mac OS X version is simply a re-compilation of the raw C-code using the Mac OS X GCC compiler, we are currently unable to guarantee that all features will work as well or efficiently as they do in the Windows or Linux versions. Therefore, The Mac OS X version is definitely a beta version, and we welcome any negative or positive feedback that the user community may wish to provide (email: jons@fhcrc.org).

A.3.2.3.1. Required software

To generate a native version of TranStat for Mac OS X, you must install a GCC compiler. The XCode application includes a viable GCC compiler. The following instructions will assume that XCode has already been installed.

A.3.2.3.2. Compiling TranStat files for Mac OS X

1. Download and uncompress the latest version of the TranStat source code to you selected TranStat source folder.
2. Open a Terminal Window (Go/Applications/Utilities/Terminal)
3. Use the command line in the terminal window to navigate to the uncompressed archive folder. (HINT: 'cd ..' will allow you to navigate back toward the root directory on your hard drive, and 'cd *foldername*' will allow you to navigate to a subdirectory, *i.e.*, away from the root directory.)
4. Type the following command 'gcc main.c'.
5. The GCC compiler should generate a new file named 'a.out' in the same folder. This is the new TranStat executable for Mac OS X. You should be able to ignore any warning messages generated by the GCC compiler. These message pertain to the redefinition of the meaning of INFINITY for the purposes of running this program.
6. You may wish to rename 'a.out' to 'main.out'.
7. To use this new TranStat executable, you must copy it to the same directory on your hard drive as any data input files that you are planning to analyze.

Note: If you Mac OSX version is not the most recent version available, then you may wish to insert the "-mmacosx-version-min=*version_number*" flag after the gcc command. For example, if your machine is running version 10.5, then you could use the command 'gcc -mmacosx-version-min=10.5 main.c'.

A.4. Input file formats for TranStat version 5

Note: All data input files, with the exception of the configuration file (config.file), must be saved as tab or space delimited text files with the extension *.dat. The order of the columns is crucial, and do not include column headers. Starting with TranStat version 5, the user must insert an additional suffix to each of the files described below, with the exception of the config.file. The suffix takes the form '*_integer*', where *integer* represents positive integers starting with 1 and counting upward in value. This change was made to ensure greater consistency with how TranStat specifies file names for simulation studies. As an example, the next section describes the 'community.dat' file. Under the new naming scheme, this file should be named 'community_1.dat' for the typical analysis (i.e., not involving simulation studies).

A.4.1. Community File: community.dat

A.4.1.1. Description: Defines the clusters of individuals, for example, households or schools, for the enrolled study population.

A.4.1.2. Columns from left to right

- *Unique Cluster Identification Number*: Each cluster must be assigned a unique identification number. Integer values must be used, starting with 1 and proceeding to *N* (number of clusters).
- The time step corresponding to the beginning of the analysis period for the cluster: This time step (often a day) is typically specified by subtracting the sum of the maximum durations of the latent and infectious periods from the earliest onset time reported for the members of the cluster. For example, if the index case for a cluster has onset of symptoms on Day 15 and the maximum lengths of the latent and infectious periods are 3 and 7 days respectively, then typically the cluster would be assigned the value of day 5 (15-[7+3]) for this column.
- The time step corresponding to the end of study follow-up for the cluster: For example, if a cluster is followed to Day 29, then the value for this column would be set to 29.

A.4.2. Population File: pop.dat

A.4.2.1. Description: Defines the individual members of the study population, their immune status at the beginning of study follow-up, and their infection and pathogenicity related information.

A.4.2.2. Columns from left to right

- Unique identification number for each individual host enrolled in the study: Each individual must be assigned a unique identification number. Integer values must be used starting with 1 and proceeding to *S* (number of individuals).
- Unique identification number for the primary cluster of which the individual is a member: This number should match one of the clusters listed in the community.dat file.
- Pre-immune status: This indicator variable (1=Pre-immune, 0=Not pre-immune) records whether or not the individual was known to be immune to infection at the start of the study follow-up period. Used in conjunction with the Pre-existing Immunity Type column (see below), TranStat permits specification of both all-or-none immunity or multiple levels of pre-existing immune protection against infection.
- Infection status: This is an indicator variable for whether the individual was infected during the study follow-up period (1=Infected, 0=Escaped infection).
- Symptomatic indicator: This indicator variable specifies whether an individual developed symptoms consistent with the study's case definition for symptomatic disease (1=symptomatic, 0=asymptomatic).
- Onset time: This variable records the integer time step for the onset of symptoms or infectiousness (as mentioned earlier, typically the former is used). The permitted values are those from the 0 to infinity, with -1 indicating that data is either missing or not relevant (*i.e.*, the individual was not a symptomatic infection).
- Index indicator: This variable only applies to data collected as part of a case-ascertained study, but the user must always specify values for this variable. This indicator variable denotes whether or not an infected individual was the first case in a cluster to be detected by surveillance (1=Index case, 0=Not an index case).
- Disease resolution indicator: This indicator variable (1=outcome occurred; 0=outcome did not occur or Symptomatic indicator=0) denotes whether or not an individual who was symptomatic experienced a subsequent event that ended their

potential to transmit to other members of the study population. The definition for such an event is situationally dependent, but a common example might be death which typically terminate the infectious period of an ongoing shedder. Ebola is a notable example of where deceased individual can still contribute to transmission.

- Disease resolution time: This variable records the integer time step for when disease resolution occurred. The permitted values are those from the 0 to infinity, with -1 indicating that data is either missing or not relevant (*i.e.*, the individual was not a symptomatic infection). For conceptual plausibility, the disease resolution time should be greater than or equal to the time of the symptom onset.
- Pathogenicity Type: This variable provides the user with the ability to specify multiple pathogenicity types. For example, if a user has information about multiple clinical case definitions that represents increasing levels of disease severity, s/he may wish to separately estimate the effects of a covariate on the risk of developing each of the levels of disease severity, given infection. The permitted values for this variable are integers starting from 0 ranging to *U*, which is equal to the number of pathogenicity types minus one. If there is only one type of pathogenicity (*i.e.*, symptomatic or not), then the value for every host should be set to 0.
- Pre-existing Immunity Type: The variable provides the user with the option to specify one or more types of pre-existing immunity, for example, multiple levels of immune protection against infection ranging from none to complete. The permitted values for this variable are integers starting from 0 ranging to *Q*, which is equal to the number of pre-existing immunity types minus one. If there is only one type of pre-existing immunity (*i.e.*, all-or-none immunity), then the value for every host should be set to 0.
- Person-level frequency weight: This variable provides the user with the ability to condense his/her pop.dat by using one line to represent multiple individuals who all have the same values for all other quantities that are associated with them and their cluster (i.e., there is nothing in the input data to distinguish these individuals from each other). Integer values of 1 or greater are permitted.
- Ignore Indicator: This indicator (1=Ignore, 0=Include) provides the user with the ability to have TranStat exclude an individual and all associated data from the analysis.

A.4.3. Community to Person Contact File: c2p_contact.dat

A.4.3.1. Description: Defines the exposure history of the study population to the community source(s) of infection. This file is organized such that each row defines a period of continuous exposure to a particular community source of infection. The identification of the type of C2P contact and the beginning and end (inclusive) of the exposure period are defined. The user may define individual, cluster-level, or study-wide histories of exposure to a community source of infection, but only one format may be used for each c2p_contact.dat file.

A.4.3.2. Study-wide format: Columns from left to right

This format of the c2p_contact.dat can be used when the analysis assumes that all of the members of the enrolled study population have the same history of exposure to community-based source(s) of infection.

- Start time for the C2P exposure: This variable records the integer time step marking the beginning of a period of continuous exposure to a community-based source of infection.

- Stop time for the C2P exposure: This variable records the integer time step marking the end of a period of continuous exposure to a community-based source of infection.
- Type of C2P exposure: Also known as the C2P contact mode, this variable records the type of community-based source of infection to which the study clusters are all exposed. The permitted values for this variable are integers starting from 0 ranging to *D*, which is equal to the number of C2P contact types minus one.
- CP2 Offset: This rarely-used field permits the user to assign importance weights to each C2P contact. Any real number value may be entered, but Transtat will normalize the weights. Assigning the same value to every row in the c2p_contact.dat file tells TranStat that every community-to-person exposure is equally important. The default unweighted value is typically 0.
- C2P Ignore Indicator: This field indicates to TranStat whether a specific C2P contact should be ignored (1=Ignore, 0=Include), *i.e.*, not incorporated into the likelihood.

A.4.3.3. Cluster-level format: Columns from left to right

This format of the c2p_contact.dat can be used when the analysis assumes that all of the members of a cluster have the same history of exposure to community-based source(s) of infection.

- Unique identification number for the cluster: This value must match a value listed for the same variable in the community.dat file.
- Start time for the C2P exposure: This variable records the integer time step marking the beginning of a period of continuous exposure to a community-based source of infection.
- Stop time for the C2P exposure: This variable records the integer time step marking the end of a period of continuous exposure to a community-based source of infection.
- Type of C2P exposure: Also known as the C2P contact mode, this variable records the type of community-based source of infection to which the cluster is exposed. The permitted values for this variable are integers starting from 0 ranging to *D*, which is equal to the number of C2P contact types minus one.
- CP2 Offset: This rarely-used field permits the user to assign importance weights to each C2P contact. Any real number value may be entered, but Transtat will normalize the weights. Assigning the same value to every row in the c2p_contact.dat file tells TranStat that every community-to-person exposure is equally important. The default unweighted value is typically 0.
- C2P Ignore Indicator: This field indicates to TranStat whether a specific C2P contact should be ignored (1=Ignore, 0=Include), *i.e.*, not incorporated into the likelihood.

A.4.3.4. Individual-level format: Columns from left to right

This format of the c2p_contact.dat can be used when the analysis assumes that each member of the enrolled study population has an independent history of exposure to community-based source(s) of infection.

- Unique identification number for the individual: This value must match a value listed for the same variable in the pop.dat file.
- Start time for the C2P exposure: This variable records the integer time step marking the beginning of a period of continuous exposure to a community-based source of infection.

- Stop time for the C2P exposure: This variable records the integer time step marking the end of a period of continuous exposure to a community-based source of infection.
- Type of C2P exposure: Also known as the C2P contact mode, this variable records the type of community-based source of infection to which the cluster is exposed. The permitted values for this variable are integers starting from 0 ranging to *D*, which is equal to the number of C2P contact types minus one.
- CP2 Offset: This rarely-used field permits the user to assign importance weights to each C2P contact. Any real number value may be entered, but Transtat will normalize the weights. Assigning the same value to every row in the c2p_contact.dat file tells TranStat that every community-to-person exposure is equally important. The default unweighted value is typically 0.
- C2P Ignore Indicator: This field indicates to TranStat whether a specific C2P contact should be ignored (1=Ignore, 0=Include), *i.e.*, not incorporated into the likelihood.

A.4.4. Person to Person Contact File: p2p_contact.dat

A.4.4.1. Description: Defines the P2P contact history between enrolled members of the study population. This file is organized such that each row defines a period of continuous contact between enrolled members of the study population. The identification of the individuals involved in the contact and the beginning and end time points (inclusive) for the contact period are defined. The user may define individual or cluster-level histories, but not both in the same p2p_contact.dat file.

A.4.4.2. Cluster-level format: Columns from left to right
This format of the p2p_contact.dat can be used when the analysis assumes that all of the members of a cluster have the same history of contact with each other.

- Unique identification number for the cluster: This value must match a value listed for the same variable in the community.dat file.
- Start time for the P2P contact: This variable records the integer time step marking the beginning of a period of continuous contact between members of the same cluster.
- Stop time for the P2P contact: This variable records the integer time step marking the end of a period of continuous contact between members of the same cluster.
- Type of P2P contact: Also known as the P2P contact mode, this variable records the type of person-to-person exposure that occurred. The permitted values for this variable are integers starting from 0 ranging to *P*, which is equal to the number of P2P contact types minus one.
- PP2 Offset: This rarely-used field permits the user to assign importance weights to each P2P contact. Any real number value may be entered, but Transtat will normalize the weights. Assigning the same value to every row in the p2p_contact.dat file tells TranStat that every person-to-person contact is equally important. The default unweighted value is typically 0.
- P2P Ignore Indicator: This field indicates to TranStat whether a specific P2P contact should be ignored (1=Ignore, 0=Include), *i.e.*, not incorporated into the likelihood.

A.4.4.3. Individual-level format: Columns from left to right

This format of the p2p_contact.dat can be used when the analysis assumes that each member of the enrolled study population has an independent history of contact with every other member of the study population.

- Start time for the P2P contact: This variable records the integer time step marking the beginning of a period of continuous contact between members of the same cluster.
- Stop time for the P2P contact: This variable records the integer time step marking the end of a period of continuous contact between members of the same cluster.
- Unique identification number for the individual who is infected: This value must match a value listed for the same variable in the pop.dat file.
- Unique identification number for the individual who is not infected: This value must match a value listed for the same variable in the pop.dat file.
- Type of P2P contact: Also known as the P2P contact mode, this variable records the type of person-to-person exposure that occurred. The permitted values for this variable are integers starting from 0 ranging to *P*, which is equal to the number of P2P contact types minus one.
- PP2 Offset: This rarely-used field permits the user to assign importance weights to each P2P contact. Any real number value may be entered, but Transtat will normalize the weights. Assigning the same value to every row in the p2p_contact.dat file tells TranStat that every person-to-person contact is equally important. The default unweighted value is typically 0.
- P2P Ignore Indicator: This field indicates to TranStat whether a specific P2P contact should be ignored (1=Ignore, 0=Include), *i.e.*, not incorporated into the likelihood.

A.4.5. Time Independent Covariate File

A.4.5.1. Description: This file stores the values for covariates that are not indexed by a time point or period. Covariates values may be specified as any real number. Covariates are specified at the individual-level. Currently, TranStat does not provide for imputation of missing covariate values, so covariate values must be specified for every member of the study population. Therefore, either the user must impute missing covariate values before importing data into TranStat or only include covariates with complete information.

A.4.5.2. Columns from left to right

- Unique identification number for the individual: This value must match a value listed for the same variable in the pop.dat file.
- One column for each time-independent covariate. Remember the order, because this is important for specifying models in the configuration file.

A.4.6. Time Dependent Covariate File

A.4.5.1. Description: This file stores the values for covariates that are indexed by a time point or period. Covariates values may be specified as any real number. Covariates are specified at the level of each combination of individual and time period. Currently, TranStat does not provide for imputation of missing covariate values, so covariate values must be specified for every combination of individual and time period. Therefore, either the user must impute missing covariate values before importing the data into TranStat or only include covariates with complete information.

A.4.5.2. Columns from left to right

- Unique identification number for the individual: This value must match a value listed for the same variable in the pop.dat file.
- Start time for the time period to which the covariate value pertains: This variable records the integer time step marking the beginning of a period to which the covariate values pertain.
- Stop time for the time period to which the covariate value pertains: This variable records the integer time step marking the end of a period to which the covariate values pertain
- One column for each time-dependent covariate. Remember the order, because this is important for specifying models in the configuration file.

A.4.7. Imputation File

A.4.7.1. Description: This file specifies the individual-level parameters that will control the imputation of outcome-related missing data. One row should be specified per individual member of the enrolled study population for whom at least one outcome-related quantity is missing.

A.4.7.2. Columns from left to right

- Unique identification number for the individual with missing data: This value must match a value listed for the same variable in the pop.dat file.
- Possible Pre-immunity Status: This indicator variable (1=Possible pre-immune, 0=Not possible pre-immune) is used to specify whether there is uncertainty with regard to the immune status of the individual at the beginning of the study follow-up period.
- Possible Escape Status: This indicator variable (1=Possible escape, 0=Not possible escape) is used to specify whether there is uncertainty with regard to whether or not an individual escaped infection by the end of the study follow-up period.
- Possible Symptomatic Infection Status: This indicator variable (1=Possible symptomatic infection, 0=Not possible symptomatic infection) is used to specify whether there is uncertainty with regard to whether or not an individual experienced symptomatic infection by the end of the study follow-up period.
- Possible Symptomatic Start Time: For possible symptomatic infections, this variable records the earliest time step during which the possible symptoms may have begun. A value of '-1' should be entered if an individual was not a possible symptomatic infection.
- Possible Symptomatic Stop Time: For possible symptomatic infections, this variable records the latest time step during which the possible symptoms may have begun. A value of '-1' should be entered if an individual was not a possible symptomatic infection.
- Possible Asymptomatic Infection Status: This indicator variable (1=Possible asymptomatic infection, 0=Not possible asymptomatic infection) is used to specify whether there is uncertainty with regard to whether or not an individual experienced asymptomatic infection by the end of the study follow-up period.
- Possible Asymptomatic Start Time: This variable records the earliest time step during which the infectiousness may have begun for the possible asymptomatic infection. A value of '-1' should be entered if an individual was not a possible asymptomatic infection.
- Possible Asymptomatic Stop Time: This variable records the latest time step during which the infectiousness may have begun for the possible asymptomatic infection. A

value of '-1' should be entered if an individual was not a possible asymptomatic infection.

A.4.8. Configuration File

A.4.8.1. Description: This uniquely-formatted file is used to define the statistical model, its constraints, and other aspects of the estimation process. The general format for this file takes the form of a list of settings. Each item on the list consists of a header line, beginning with the character '**#**', and then followed by a line or set of lines that define the values for that item. The header text for each item is prescribed, so the user must use an exact copy of the header text listed below. The next section lists the item in order of appearance from the beginning to the end of the config.file.

A.4.8.2. Configuration settings in order from beginning to end
The description for each of the following items includes a Format section. In this section quantities listed in *italics* font denote values that must be specified by the user.
The configuration file requires that each section begin with a header-line containing a comment detailing the sections contents. Each comment is proceeded by a '**#**' and a space. The comments MUST be formatted EXACTLY as listed below.

A.4.8.2.a. Defining the input directory path
The user can define a specific path to the directory where the necessary input files are stored for the analysis. Leaving this section blank amounts to assuming that the input files and TranStat executable are stored in the same directory. The directory path should be specified in a manner consistent with your operating system. Windows users should include a terminal '/'.

Format

 # input-path
String had coding the location of the directory where input files are stored.

A.4.8.2.b. Defining the output path for saving results
The user can define a specific path to the directory where output file(s) will be stored for the analysis. Leaving this section blank amounts to assuming that the output files will be stored in the same directory as the TranStat executable. The directory path should be specified in a manner consistent with your operating system. If the user would like to specify a prefix for the output file, then this should be included at the end of this string. Windows users should include a terminal '/' when specifying only an output directory (i.e., no prefix for the output file).

Format

 # output-path
String had coding the location of the directory where input files are stored.

A.4.8.2.c. Defining the distribution for the latent/incubation period
The user defines the time period relative to the onset of infectiousness/symptoms when the infection may have occurred. This time period is defined by the minimum and maximum number of time steps into the past relative to the onset of infectiousness/symptoms, as well as the probability of infection having occurred during each of the time steps falling within this period.

Format

min-max-days-and-probs-of-incubation-period
Number of different incubation period distributions provided
For each incubation period distribution two lines are specified:
Minimum maximum duration (separated by a space)
A space-delimited list of the daily probabilities that infection occurred on each day of the incubation period (should sum to 1), in left-to-right order from minimum to maximum duration

A.4.8.2.d. Defining the distribution for the infectious period
The user defines the time period relative to the onset of infectiousness/symptoms during which a case remains infectious. This time period is defined by the minimum and maximum possible length of the infectious period, as well as the probability of remaining infectious on each day of the period.

Format

primary-lower-upper-bounds-and-probs-of-infectious-days-relative-to-symptom-onset-day
Number of different infectious period distributions provided
For each infectious period distribution two lines are specified:
Minimum maximum duration (separated by a space)
A space-delimited list of the daily probability of remaining infectious for each day of the infectious period, in left-to-right order from minimum to maximum duration

A.4.8.2.e. Defining the number of types of C2P contacts included in the model
The user defines the number of types of C2P contacts to model (this must match the number of C2P types listed in the c2p_contact.dat file). A separate community probability of infection is estimated for each type of C2P contact.

Format

number-of-c2p-transmission-probabilities
Number of C2P contact types

A.4.8.2.d. Defining the number of types of P2P contacts included in the model
The user defines the number of types of P2P contacts to model (this must match the number of P2P types listed in the p2p_contact.dat file). A separate person-to-person probability of infection is estimated for each type of P2P contact.

Format

number-of-p2p-transmission-probabilities
Number of P2P contact types

A.4.8.2.f. Defining the number of types of pathogenicity
The user defines the number of categories of pathogenic disease (*i.e.*, the number of case definitions for clinical disease) that are to be incorporated into the model. This value should match the number of pathogenicity types listed in the pop.dat.

Format

number-of-pathogenicity-groups
Number of pathogenicity types

A.4.8.2.g. Defining the number of types of pre-existing immunity
The user defines the number of categories/levels for immunity at the beginning of the study follow-up period. This value should match the number of pre-immunity types listed in the pop.dat.

Format

number-of-preseason-immunity-groups
Number of pre-immunity types

A.4.8.2.h. Defining the number of time-independent covariates
The user defines the number of time-independent covariates included in the time_ind_covariate.dat file.

Format

number-of-time-independent-covariates
Number of time-independent covariates

A.4.8.2.i. Defining the number of time-dependent covariates
The user defines the number of time-dependent covariates included in the time_dep_covariate.dat file.

Format

number-of-time-dependent-covariates
Number of time-dependent covariates

A.4.8.2.j. Defining covariate effects on susceptibility to infection due to exposure through C2P contact
This item allows the user to specify which covariate will affect susceptibility to infection due to C2P exposure. TranStat estimates the cooperate effect of a covariate on all types of C2P contact, though the user may relax this constraint by including additional covariates specifying the interaction between type of C2P contact and a covariate. Covariate effects are specified in the following manner on a single line. An initial integer between 0 and the total number of covariates (time-independent plus time-dependent) indicates the number of covariates associated with susceptibility to infection due to C2P exposure. This number is followed immediately by a ":" and then a space. After that, a space-delimited list includes the index of each covariate associated with susceptibility to infection due to C2P exposure. Covariates are indexed, as integers starting with 1, by the order that they appear in the time-independent and then the time-dependent covariate data files.

Format

covariates-affecting-susceptibility-for-c2p-transmission
Number of covariates modifying C2P: *a space-delimited list of the covariate index numbers*

A.4.8.2.k. Defining covariate effects on susceptibility to infection due to exposure through P2P contact

This item allows the user to specify which covariate will affect susceptibility to infection due to P2P exposure. TranStat estimates the cooperate effect of a covariate on all types of susceptibility to infection through P2P contact, though the user may relax this constraint by including additional covariates specifying the interaction between type of P2P contact and a covariate. Covariate effects are specified in the following manner on a single line. An initial integer between 0 and the total number of covariates (time-independent plus time-dependent) indicates the number of covariates associated with susceptibility to infection due to P2P exposure. This number is followed immediately by a ":" and then a space. After that, a space-delimited list includes the index of each covariate associated with susceptibility to infection due to P2P exposure. Covariates are indexed, as integers starting with 1, by the order that they appear in the time-independent and then the time-dependent covariate data files.

Format

covariates-affecting-susceptibility-for-p2p-transmission
Number of covariates modifying P2P: *a space-delimited list of the covariate index numbers*

A.4.8.2.l. Defining covariate effects on infectiousness through P2P transmission

This item allows the user to specify which covariate affect infectiousness through P2P transmission. TranStat estimates the cooperate effect of a covariate on all types of infectiousness through P2P contact, though the user may relax this constraint by including additional covariates specifying the interaction between type of P2P contact and a covariate. Covariate effects are specified in the following manner on a single line. An initial integer between 0 and the total number of covariates (time-independent plus time-dependent) indicates the number of covariates associated with infectiousness through P2P transmission. This number is followed immediately by a ":" and then a space. After that, a space-delimited list includes the index of each covariate associated with infectiousness through P2P transmission. Covariates are indexed, as integers starting with 1, by the order that they appear in the time-independent and then the time-dependent covariate data files.

Format

covariates-affecting-infectiousness-for-p2p-transmission
Number of covariates modifying P2P infectiousness: *a space-delimited list of the covariate index numbers*

A.4.8.2.m. Defining covariates for which there is an interaction between the covariate's effect on susceptibility to infection through P2P transmission and its effect on subsequent infectiousness through P2P transmission

If the effect of a covariate on an individual's infectiousness to other members of the enrolled study population (i.e., via P2P transmission) is modified by, or, in other words, interacts with that same covariate's effect on the individual's susceptibility to infection (i.e., before they become infectious), then this is where such an interaction may be specified. Covariate effects are specified in the following manner on a single line. An initial integer between 0 and the total number of covariates (time-independent plus time-dependent) indicates the number of covariates for which interaction between the effects on susceptibility and infectiousness will be modeled. This number is followed immediately by a ":" and then a space. After that, a space-delimited list includes the index of each covariate for which interaction is modeled. Covariates

are indexed, as integers starting with 1, by the order that they appear in the time-independent and then the time-dependent covariate data files.

Format

interactions-for-p2p-transmission
Number of covariates with interaction between susceptibility and infectiousness effects: *a space-delimited list of the covariate index numbers*

A.4.8.2.n. Defining covariate effects on pathogenicity
This item allows the user to specify which covariate affect pathogenicity. TranStat estimates the cooperate effect of a covariate on all types of pathogenicity, though the user may relax this constraint by including additional covariates specifying the interaction between type of pathogenicity and a covariate. Covariate effects are specified in the following manner on a single line. An initial integer between 0 and the total number of covariates (time-independent plus time-dependent) indicates the number of covariates associated with pathogenicity. This number is followed immediately by a ":" and then a space. After that, a space-delimited list includes the index of each covariate associated with pathogenicity. Covariates are indexed, as integers starting with 1, by the order that they appear in the time-independent and then the time-dependent covariate data files.

Format

covariates-affecting-pathogenicity
Number of covariates modifying pathogenicity: *a space-delimited list of the covariate index numbers*

A.4.8.2.o. Defining covariate effects on existing immunity at the start of study follow-up (pre-immunity)
This item allows the user to specify which covariate affect pre-immunity. TranStat estimates the cooperate effect of a covariate on all types of pre-immunity, though the user may relax this constraint by including additional covariates specifying the interaction between type of pre-immunity and a covariate. Covariate effects are specified in the following manner on a single line. An initial integer between 0 and the total number of covariates (time-independent plus time-dependent) indicates the number of covariates associated with pre-immunity. This number is followed immediately by a ":" and then a space. After that, a space-delimited list includes the index of each covariate associated with pre-immunity. Covariates are indexed, as integers starting with 1, by the order that they appear in the time-independent and then the time-dependent covariate data files.

Format

covariates-affecting-preseason-immunity
Number of covariates modifying pre-existing immunity to infection: *a space-delimited list of the covariate index numbers*

A.4.8.2.p. Defining which covariate effects can be combined into a single effect
There are often situations where it is advisable to combine covariate effects into a single effect. For example, say the user defines an effect for an individual's age on susceptibility to infection due to C2P contact and a separate effect for age on susceptibility due to P2P contact. But, there may be no epidemiological or biological reason to believe that age will have different

effects on susceptibility for P2P versus C2P exposure, so to reduce the number of parameters that the model estimates (thereby increasing the power to detect an age effect) the user could tell TranStat to essentially combine these two different effects of age into a single effect. On the first line below the header text, the user will enter a single number representing the updated total number of covariate effects after this combining process has been completed. The number entered on this line will dictate the number of subsequent lines to follow, and an entry of 0 is equivalent to indicating that the model should estimate every covariate effect defined in sections A.4.8.2.i to A.4.8.2.n. If the user enters any number on this line that is greater than 0 and there are covariate effects specified in sections A.4.8.2.i to A.4.8.2.n, then the user must include every covariate effect in a 'combined effect', even if the number of covariates included in a group is 1.

Each of these subsequent lines should take the following format. An initial integer between 0 and the total number of covariates (time-independent plus time-dependent) indicates the number of covariates that will be aggregated into the combined effect being represented by this line. This number is followed immediately by a ":" and then a space. After that, a space-delimited list includes the index of each covariates being combined. This is the last item in the config.file where covariates are indexed, as integers starting with 1, by the order that they appear in the time-independent, then the time-dependent covariate data files, and then in each of the sections from A.4.8.2.i to A.4.8.2.n in descending order.

After the next section (A.4.8.2.p), all covariate indices will be for the combined effects listed in this section. The combined effects will be index, starting with the integer of 1, in the order that they are listed in this section. If 0 combined effects are defined in this section, then TranStat will continue to index all covariate effects by the order that they are listed in sections A.4.8.2.i to A.4.8.2.n.

Format

equal-parameters
Number of new combined effects
Number of lines equal to the "Number of combined effects", for each line include: *Number of covariates included in the combined effect: a space-delimited list of the covariate index numbers*

A.4.8.2.q. Defining which covariate effects or transmission parameters will be fixed.
This section is used to defined which covariates effects or transmission parameters will be fixed. TranStat will not estimate values for these fixed parameters, treating them as known quantities. The primary utility of fixing covariate values is to allow the analyst to include a covariate effect or transmission parameter that is strongly suspected to be operating on the transmission of the pathogen being analyzed, but for which insufficient data is present to allow for estimation of this parameter. This feature can be used to conduct sensitivity analyses for the fixed parameters, i.e., through multiple runs of TranStat where the value at which the covariate effect(s) are fixed is changed with each run.

Format

fixed-parameters
Number of parameters with fixed values, followed by a colon
Number of lines equal to the "Number of parameters with fixed values", for each line include: *Index of the parameter whose value will be fixed, a colon, a space, and then the value at which the parameter will be fixed.*

A.4.8.2.r. Indicator for whether or not to perform a simulation study using the population and model structure defined in prior sections
TranStat can perform simulations studies that simulate outbreaks using the population and model structure defined above, along with user-defined values for the parameters that would be estimated by the model (those listed under section A.4.8.2.o). Performing a simulation study will mean that TranStat will not infer parameters from the observed data. Instead, TranStat will instead simulate outbreaks using the population structure and defined model, conduct inference on the simulate outbreak, perform the first two steps for a user-defined number of simulations, and report the results of all simulations.

Format

perform-simulation
Indicator for whether or not to perform simulation study (1=Perform, 0=Do not perform)

A.4.8.2.s. For simulation studies, defining the proportion of non-cases with ambiguous immune versus escaped-infection status
For members of the study population who did not become symptomatic cases by the end of the study's observation period, there may be ambiguity whether or not these individuals escaped infection or were already fully/partially immune to infection before the study began. This ambiguity can be captured in simulation studies by entering a value (range 0 to 1) for the simulated proportion of non-cases who may or may not have been already immune to the infection at the start of the study.

Format

proportion-with-ambiguity-about-preimmunity-and-escape-status
Proportion of non-cases with ambiguous pre-immune versus escaped status

A.4.8.2.t. Defining the parameter inputs for simulation.
Here the user defines the values for the non-fixed model parameters (i.e., those defined in A.4.8.2.o) to be used for simulation.

Format

parameters-for-simulation
One line per parameter, with parameters listed in the order they are defined in A.4.8.2.o. Each line should contain the following: a set of string characters identifying the parameter (ignored by TranStat, but useful cue for user), a space, and the value for the parameter to be used for simulation.

A.4.8.2.u. Defining the number of outbreaks to simulate.
The user provides and non-zero integer for the number of simulation runs to conduct using the simulation settings provided in Sections A.4.8.2.r-s. For each simulation run, TranStat provides estimates of the model parameters from Section A.4.8.2.o as a single row in a long output file.

Format

number-of-simulations
Non-zero integer indicating the number of outbreaks to simulate.

A.4.8.2.v. Defining the maximum likelihood optimization routine to use for inference
Two routines are provided: the Newton-Raphson (option 0), the Downhill-Simplex (option 1), and the combination of the two routines (option 2).

Format

optimization-choice
Indicator for which maximum likelihood optimization routine to use (0,1, or 2).

A.4.8.2.w. Define the cut-off value for when convergence is achieved by the maximum likelihood optimization routine
This section allows the user to define whether or not to define convergence criteria, and then to define values for each parameter listed in Section A.4.8.2.o.

Format

converge-criteria
The first line is a binary switch for whether (1) or not (0) the user defines convergence criteria. The subsequent lines contain one line per parameter defined in Section A.4.8.2.o (using the same order as defined in that section). Each line should contain the following: a set of string characters identifying the parameter (ignored by TranStat, but useful cue for user), a space, and the value for the convergence cut-off. As a rule of thumb, this value should not be larger than 0.00001 (i.e., 1e-5).

A.4.8.2.x. Define initial estimates for each parameter defined in Section A.4.8.2.o.
TranStat requires initial starting values for the maximum likelihood optimization routine. The user must define the number of starting points, whether or not these will be user-defined, and, if user-defined, a value for each parameter at each starting point (i.e., number of starting points times the number of parameters defined in A.4.8.2.o).

Format

initial-estimates
Number of initial starting points; followed by a colon; and followed by the indicator (1=Yes, 0=No) for whether or not the user will define the values for each parameter at each starting point.
If the user chooses to define the values for each parameter at each starting point, then the next set of lines will define a parameter value on each line. There will be a number of lines equal to the number of parameters defined in Section A.4.8.2.o times the number of starting points. TranStat will read the lines in the following manner: one line for each parameter values for starting point 1 (parameter order should be the same as that of Section A.4.8.2.o), then one line per parameter for starting point 2, etc… *Each line should contain the following: a set of string characters identifying the parameter (ignored by TranStat, but useful cue for user), a space, and the value for the initial value for that parameter at that starting point.*

A.4.8.2.y. Defining upper and lower boundaries for the parameter space over which TranStat will optimize the likelihood
This section allows the analyst to defined whether or not bounds will be defined, and, if user-defined bounds are specified, the upper and lower search bounds for each parameter. Search

bounds must be positive and consistent with the nature of the parameter (for example, probabilities can not have an upper bound equal to or greater than 1.0).

Format

search-bounds
The first line is a binary switch for whether (1) or not (0) the user defines search bounds. The subsequent lines contain one line per parameter defined in Section A.4.8.2.o (using the same order as defined in that section). Each line should contain the following: a set of string characters identifying the parameter (ignored by TranStat, but useful cue for user), a space, the value for the lower search bound, and the value for the upper search bound

A.4.8.2.z. Defining whether or not imputation will be conducted for missing outcome and/or pre-existing immunity data.
This indicator (1=Yes, 0=No) determines whether TranStat will apply the hybrid Expectation-Maximization (EM), Monte Carlo EM algorithm (Yang et al. 2009) to account for missing outcome (status or onset time) and/or pre-existing immunity status information. This switch turns on the EM algorithm. The choice whether to use the MCEM component of the algorithm is regulated by A.4.8.2.z.

Format

perform-EM-algorithm
Binary indicator for whether or not to use the EM-MCEM algorithm to account for missing outcome and/or pre-existing immunity status information.

A.4.8.2.aa. For clusters of individuals with at least one member who has missing outcome and/or pre-existing immunity status information, defining the level of missingness at which the MCEM component of EM-MCEM algorithm is used in place of the EM.
For clusters of individuals where at least one individual has missing outcome or pre-existing immunity related information, we can calculate the possible number of combinations of values for the missing information. The number of possible combinations for a cluster can be quite larger, so the MCEM component of the EM-MCEM algorithm can be used to approximate the model fit that would be achieved using the full EM algorithm. A non-negative integer (includes 0) is recorded in this section of the config file to indicate the cutoff for the number of possible combinations above which TranStat will start using the MCEM. If the user would like to always use the MCEM for every cluster with missing information, enter a low valued integer. On the other hand, to always use the EM algorithm for every cluster with missing information, enter an integer with a larger value.

Format

min-number-of-possible-status-to-use-mcem
Non-negative integer for the cut-off for the number of combinations above which TranStat will use the MCEM component of the EM-MCEM algorithm.

A.4.8.2.bb. Defining whether the EM-MCEM algorithm will use community-specific weighting

This indicator (1=Yes, 0=No) provides the analyst with the ability to turn on or off the community-specific weighting feature of the EM-MCEM algorithm (see Yang et al. 2009) for more details.

Format

use-community-specific-weighting
Indicator for switching the community-specific weighting on an off.

A.4.8.2.cc. For the MCEM component of the EM-MCEM algorithm, defining the number of Monte Carlo Markov chain samples to draw for calculating the initial starting point for chains that will draw the MCEM samples for the analysis.

Format

number-of-base-mcmc-samples
Non-negative integer for the base number of MCMC samples to draw.

A.4.8.2.dd. For the MCEM component of the EM-MCEM algorithm, defining the number of burnin samples (i.e., after the base sampling) to run for each chain before beginning to draw MCEM samples for the analysis.

Format

number-of-burnin-mcmc-samples
Non-negative integer for the number of burnin MCMC samples to draw during each burnin iteration.

A.4.8.2.ee. For the MCEM component of the EM-MCEM algorithm, defining the number of iterations that the number of burnin samples (Section A.4.8.2.cc) are drawn

Format

number-of-burnin-mcmc-iterations
Non-negative integer for the number of burnin iterations to conduct.

A.4.8.2.ff. For the MCEM component of the EM-MCEM algorithm, defining the number of additional MCEM samples to draw for calculating the Monte Carlo error.

Format

number-of-samplings-for-mc-error
Non-negative integer for the number of extra MCEM samples to draw to calculate the MC component of the error.

A.4.8.2.gg. For the MCEM component of the EM-MCEM algorithm, defining whether or not to use an added bootstrapping procedure in the estimation of the Monte Carlo error.

Format

use-bootstrap-for-mc-error

Indicator (1=Yes, 0=No) for whether or not to use the extra bootstrapping procedure in the estimation of the MC error.

A.4.8.2.hh. A switch for indicating whether or not to skip (1 = skip, 0 = don't skip) estimation of the average variance for the Monte Carlo error.

Format

do-not-calculate-average-variance-for-mc-error
Indicator (1=Yes, 0=No) for whether or not to skip estimation of the average variance for the MC error.

A.4.8.2.ii. A switch to have TranStat perform extra checks related to the nature of the patterns in missing outcome (status and/or onset time) and/or pre-existing immunity information.

Format

check-missingness
Indicator (1=Yes, 0=No) for whether or not to conduct the additional checks related to the missingness patterns for outcome and/or pre-existing immunity information.

A.4.8.2.jj. A switch to have TranStat perform extra checks related to the nature of the mixing/contact patterns within the observed data (i.e., information contained in p2p_contact.dat and c2p_contact.dat).

Format

check-mixing
Indicator (1=Yes, 0=No) for whether or not to conduct the additional checks related to the mixing patterns of the study population.

A.4.8.2.kk. A switch to have TranStat report the overall runtime for the analysis.

Format

check-runtime
Indicator (1=Yes, 0=No) for whether or not to report the analysis runtime.

A.4.8.2.ll. For simulation studies, the analyst can define how infective asymptomatic infections are relative to their symptomatic counter parts

Format

relative-infectivity-of-asymptomatic-case-for-simulation
A zero or positive value should be entered here when conducting simulations studies. Zero denotes that asymptomatic infections are not infectious, and a value of 1.0 or greater indicates that the asymptomatic infections are as or more infectious, respectively, than symptomatic infections.

A.4.8.2.mm. For inference, the analyst can define how infective asymptomatic infections are relative to their symptomatic counter parts
Currently, TranStat does not directly estimate the relative infectiousness of asymptomatic versus symptomatic infections, but there are ways to structure the model, such that this parameter could be estimated. Here the user can define this parameter as a fixed value.

Format

relative-infectivity-of-asymptomatic-case-for-estimation
A zero or positive value should be entered here when conducting inference. Zero denotes that asymptomatic infections are not infectious, and a value of 1.0 or greater indicates that the asymptomatic infections are as or more infectious, respectively, than symptomatic infections.

A.4.8.2.nn. First indicator that helps define the structure of the C2P contact file.
Here the user defines whether or not every member of the study population share the same community-to-person contact history. If this indicator is set to 1, then the c2p_contact.dat file should be structured in the manner detailed in Section A.4.3.2. If this indicator is set to 0, then the c2p_contact.dat file should be structured in the manner detailed in either Section A.4.3.3 (Section A.4.8.2.nn = 1) or Section A.4.3.4 (Section A.4.8.2.nn=0).

Format

members-share-common-c2p-contact-history-across-communities
Indicator (1=Yes, 0=No).

A.4.8.2.oo. Indicator that helps define the structure of the C2P and P2P contact files.
Here the user defines whether or not every member of the same cluster shares the same community-to-person and person-to-person contact histories. Sharing the same person-to-person contact history is equivalent to assuming random mixing within the cluster for the period defined by each P2P contact listed in the p2p_contact.dat file. If this indicator is set to 1, then the c2p_contact.dat file should be structured in the manner detailed in Section A.4.3.3. If this indicator is set to 0 and the value for Section A.4.8.2.mm is 0, then the c2p_contact.dat file should be structured in the manner detailed in Section A.4.3.4. If this indicator is set to 1, then the p2p_contact.dat file should be structured in the manner detailed in Section A.4.4.2, otherwise the structure detailed in Section A.4.4.3 should be used.

Format

members-share-common-contact-history-within-communities
Indicator (1=Yes, 0=No).

A.4.8.2.pp. Indicator for whether or not TranStat should generate its own C2P contact file
This is a convenience feature, allowing the user to have TranStat generate c2p_contact.dat files under the conditions set by the indicators in Sections A.4.8.2.mm and A.4.8.2.nn.

Format

automatically-generate-c2p-contact-file
Indicator (1=generate c2p file, 0=don't generate c2p file).

A.4.8.2.qq. Indicator for whether or not TranStat should generate its own P2P contact file

This is a convenience feature, allowing the user to have TranStat generate p2p_contact.dat files under the conditions set by the indicators in Section A.4.8.2.nn.

Format

automatically-generate-p2p-contact-file
Indicator (1=generate c2p file, 0=don't generate c2p file).

A.4.8.2.rr. Indicator for whether or not TranStat should use the C2P offset weights supplied in the c2p_contact.dat file.

Format

use-c2p-offset
Indicator (1=Yes, 0=No).

A.4.8.2.ss. Indicator for whether or not TranStat should use the P2P offset weights supplied in the p2p_contact.dat file.

Format

use-p2p-offset
Indicator (1=Yes, 0=No).

A.4.8.2.tt. For case-ascertained data (applies to the majority of dataset analyzed by TranStat to date), this indicator allows the user to turn on or off the model's adjustment for left truncation, which is a likely source of ascertainment bias

Format

adjust-for-selection-bias
Indicator (1=Yes, 0=No).

A.4.8.2.uu. Indicator for whether or not to apply standard adjustments for right censoring of observations

Format

adjust-for-right-censoring
Indicator (1=Yes, 0=No).

A.4.8.2.vv. Indicator for whether or not TranStat should re-assign index case status based upon their onset date relative to the user-specified cluster index case(s)
When conducting an appropriate analysis of a case-ascertained study population, TranStat will automatically re-assign index case status to individuals whose onset time is before that of their cluster's user-specified index case (this also means that the user-specified index case will be assign to a non-index status). This switch turns off that function. It is primarily meant to be used for analyses that impute or simulated onset times.

Format

prefix-index-cases
Indicator (1=Yes, 0=No).

A.4.8.2.ww. Here the user defines the average duration of exposure to each of the defined community sources of infection
For each c2p probability listed in Section A.4.8.2.c, the user should list the average duration of exposure experience by the members of the study population with any exposure of that type. One line should be listed per c2p type, and they should be listed in the order of their values from the c2p_contact.dat file, i.e., c2p_mode=0 should be listed on the first line, c2p_mode=1 on the next, etc… The CPI is a deterministic calculation involving the estimated value for the c2p probability and the duration entered here, so this section can also be used to estimate the CPI for other durations of exposure other than the average experienced by the study population. Doing this will not affect the model fit.

Format

epidemic-duration-for-calculating-CPI
One line per C2P probability type. Each line should contain one value representing the average duration of exposure of the study population to that type of exposure (only for those who are exposed to that type of C2P)

A.4.8.2.xx. Defining artificially truncated infectious periods
Processes that are not directly related to the transmission process (for example, treatment, hospitalization, or death) can artificially shorten the effective infectious period of members of the study population. If this artificial truncation for all individuals who might transmit infection via a particular type of P2P contact, then this section can be used to indicate that truncation. Even if no truncation occurs, the analyst must complete fill out this section. This section can also be used to define left truncation of the infectious period.

For each P2P probability defined in Section A.4.8.2.d and in order of the values assigned using the p2p_mode column in the p2p_contact.dat file, a line of text should be defined containing the following: the first day of infectiousness (relative to symptom onset) for that P2P contact type, a space, and the last day of infectiousness (again relative to symptom onset) for that P2P contact type. If no left or right truncation of the infectious period is defined, then the value for the first and last day of infectiousness should match those listed in the first two rows of Section A.4.8.2.b, respectively.

Format

effective-lower-upper-bounds-of-infectious-days-relative-to-symptom-onset-day
One line per P2P probability type. Each line should contain the following: the first day of infectiousness (relative to symptom onset) for that P2P contact type, a space, and the last day of infectiousness (again relative to symptom onset) for that P2P contact type.

A.4.8.2.yy. Defining the average number of contacts per infectious individual, by type of p2p contact.
Estimation of the basic / effective reproductive number requires knowledge of the average number of p2p_contacts that infectious member of the study population is expected to have with susceptible members of the study population per time step (for example, per day). This section first defines the number of different versions of the reproductive number that the user would like to estimate (more than one indicates that the user would like to calculate a number of versions

based upon uncertainty in the average number of p2p-contacts per time step). The number of subsequent lines of text correspond to the product of the number of versions of the reproductive number and the number of p2p transmission probabilities detained in Section A.4.8.2.d. Line of text should be listed in the following order: first p2p transmission probability (p2p_mode=0 from the p2p_contact.dat file), the second p2p transmission probability, …, last p2p transmission probability, and then repeat this sequence for each of the subsequent versions of the reproductive number. Each line should contain the average number of contacts for the p2p transmission probability, followed by a space, and followed by the relative weight of that probabilities contribution to that version of the reproductive number (i.e., weights of 1.0 for every p2p transmission probability contributing to a particular version of the reproductive number will result in each probability contributing equally to the estimated reproductive number).

Format

multiplier-for-calculating-R0
This section first defines the number of different versions of the reproductive number that the user would like to estimate (more than one indicates that the user would like to calculate a number of versions based upon uncertainty in the average number of p2p-contacts per time step). This line must end with a colon.
The number of subsequent lines of text correspond to the product of the number of versions of the reproductive number and the number of p2p transmission probabilities detained in Section A.4.8.2.d. Line of text should be listed in the following order: first p2p transmission probability (p2p_mode=0 from the p2p_contact.dat file), the second p2p transmission probability, …, last p2p transmission probability, and then repeat this sequence for each of the subsequent versions of the reproductive number. Each line should contain the average number of contacts for the p2p transmission probability, followed by a space, and followed by the relative weight of that probabilities contribution to that version of the reproductive number (i.e., weights of 1.0 for every p2p transmission probability contributing to a particular version of the reproductive number will result in each probability contributing equally to the estimated reproductive number).

A.4.8.2.zz. Defining time-steps for a time-varying R0

Format
 # serial-division-of-epidemic-for-calculating-time-varying-R0
Indicator for whether (1) or not (0) R0 is time-varying. This line must end with a colon.
This line should contain three space-delimited integers: Time step indicating location of the earliest change from one R0 segment to the next; time step for the last of these transitions; and (only for situations where there are more than 2 types of P2P contact) the length of each segment.

A.4.8.2.aaa. Indicator for whether or not to perform assessments of the goodness of fit of the model.
TranStat does implement some procedures that will output information that the user can subsequently use to assess model fit, but interpretation of this output does require additional expertise. A value of 1 will tell TranStat to generate the goodness of fit output files, which the user will have to post-process.

Format

goodness-of-fit
Indicator (1=Yes, 0=No)

A.4.8.2.bbb. Indicator for whether or not to perform a permutation test for the null hypothesis that every P2P is equal to 0 (i.e., that there is no person-to-person transmission)
Given that 0 is a boundary value for these P2P probabilities, the appropriate statistical test to address this null hypothesis is a form of permutation test. This indicator turns on (value of 1) or off (value of 0) this test.

Format

perform-statistical-test
Indicator (1=Yes, 0=No)

A.4.8.2.ccc. An indicator for whether or not to estimate the variance.
This is a convenience function to allow the user to only fit the mean model, and ignore estimation of parameter uncertainty. This function is rarely used.

Format

do-not-estimate-variance
Indicator (1=Yes, 0=No)

A.4.8.2.ddd. An indicator for whether or not to output estimates to a file.
This convenience function can reduce the I/O burden for large analysis runs, but it is rarely used in practice.

Format

do-not-output-estimates
Indicator (1=Yes, 0=No)

A.4.8.2.eee. An indicator for whether or not to simplify the format of the output file to a single space-delimited line.
This convenience function can reduce the I/O burden for large analysis runs, but it is rarely used in practice.

Format

simplify-output
Indicator (1=Yes, 0=No)

A.4.8.2.fff. An indicator for whether or not to save SAR output in a separate file with suffix '_SAR'.
This convenience function can reduce the I/O burden for large analysis runs, but it is rarely used in practice.

Format

simplify-output-SAR
Indicator (1=Yes, 0=No)

A.4.8.2.ggg. An indicator for whether or not to save R0 output in a separate file with suffix '_R0'.
This convenience function can reduce the I/O burden for large analysis runs, but it is rarely used in practice.

Format

simplify-output-R0
Indicator (1=Yes, 0=No)

A.4.8.2.hhh. An indicator for whether or not to suppress the output that is typically generated in the command prompt terminal during each analysis run.
This convenience function turns the typical model run output seen in the command terminal off.

Format

run-transtat-silently
Indicator (1=Yes, 0=No)

A.4.8.2.iii. An indicator for whether or not to generate write a log of any errors to a file.
This convenience function turns the typical model run output seen in the command terminal off.

Format

write-error-log
Indicator (1=Yes, 0=No)

A.4.8.3. Sample configuration file (the initial paragraph is a file header describing the transmission model specified by this config.file). This is the config.file that accompanies Toy Example 1 of the SISMID 2017, Module 5, Lecture 6 Case Study.

```
# input-path

# output-path

# min-max-days-and-probs-of-incubation-period
1
1  3
0.33  0.33  0.33


# primary-lower-upper-bounds-and-probs-of-infectious-days-relative-to-symptom-onset-day
1
0  6
1.0  1.0  1.0  0.8  0.6  0.4  0.2


# number-of-c2p-transmission-probabilities
1


# number-of-p2p-transmission-probabilities
1


# number-of-pathogenicity-groups
0


# number-of-preseason-immunity-groups
0


# number-of-time-independent-covariates
1


# number-of-time-dependent-covariates
0


# covariates-affecting-susceptibility-for-c2p-transmission
0:

# covariates-affecting-susceptibility-for-p2p-transmission
1: 1

# covariates-affecting-infectiousness-for-p2p-transmission
```



```
0:

# interactions-for-p2p-transmission
0:

# covariates-affecting-pathogenicity
0:

# covariates-affecting-preseason-immunity
0:

# equal-parameters
3:
1: 1
1: 2
1: 3

# fixed-parameters
0:

# perform-simulation
0

# proportion-with-ambiguity-about-preimmunity-and-escape-status
0

# optimization-choice
0

# converge-criteria
1:
b1  1.0e-4
p1  1.0e-4
or1  1.0e-4

# initial-estimates
1:1
b1  1.0e-1
p1  1.0e-1
or1  1.0e-4

# search-bounds
1:
b1  1.000000e-8  1.000000e-1
p1  1.000000e-8  1.000000e-1
or1  1.000000e-8  150

# perform-EM-algorithm
0

# min-number-of-possible-status-to-use-mcem
```

0

use-community-specific-weighting
0

number-of-base-mcmc-samples
0

number-of-burnin-mcmc-samples
0

number-of-burnin-mcmc-iterations
0

number-of-samplings-for-mc-error
0

use-bootstrap-for-mc-error
0

do-not-calculate-average-variance-for-mc-error
0

check-missingness
0

check-mixing
0

check-runtime
0

relative-infectivity-of-asymptomatic-case-for-simulation
0

relative-infectivity-of-asymptomatic-case-for-estimation
0

members-share-common-c2p-contact-history-across-communities
0

members-share-common-contact-history-within-communities
0

automatically-generate-c2p-contact-file
0

automatically-generate-p2p-contact-file
0

use-c2p-offset

```
0

# use-p2p-offset
0

# adjust-for-selection-bias
1

# adjust-for-right-censoring
0

# prefix-index-cases
1

# epidemic-duration-for-calculating-CPI
10

# effective-lower-upper-bounds-of-infectious-days-relative-to-symptom-onset-day
0:

# multiplier-for-calculating-R0
1:
3.0 1.0

# serial-division-of-epidemic-for-calculating-time-varying-R0
0:

# goodness-of-fit
0

# perform-statistical-test
0

# do-not-estimate-variance
0

# do-not-output-estimates
0

# simplify-output
0

# simplify-output-SAR
0

# simplify-output-R0
0

# run-transtat-silently
0
```

```
# write-error-log
0
```