# Real-time analysis of infectious disease outbreaks using TranStat

Jonathan D. Sugimoto, PhD (jsugimot@ufl.edu)

Yang Yang, PhD (yangyang@phhp.ufl.edu)

# Lecture 5 Outline

- Overview of TranStat
- Basic Description of the statistical model implement by TranStat
- Case Studies
  - Case study 1: Independent cluster data
  - Case study 2: Dependent cluster data
  - Case study 3: Large population surveillance data
  - Case study 4: Accounting for missing outcome information
- Summary
- Exercise for Lecture 10

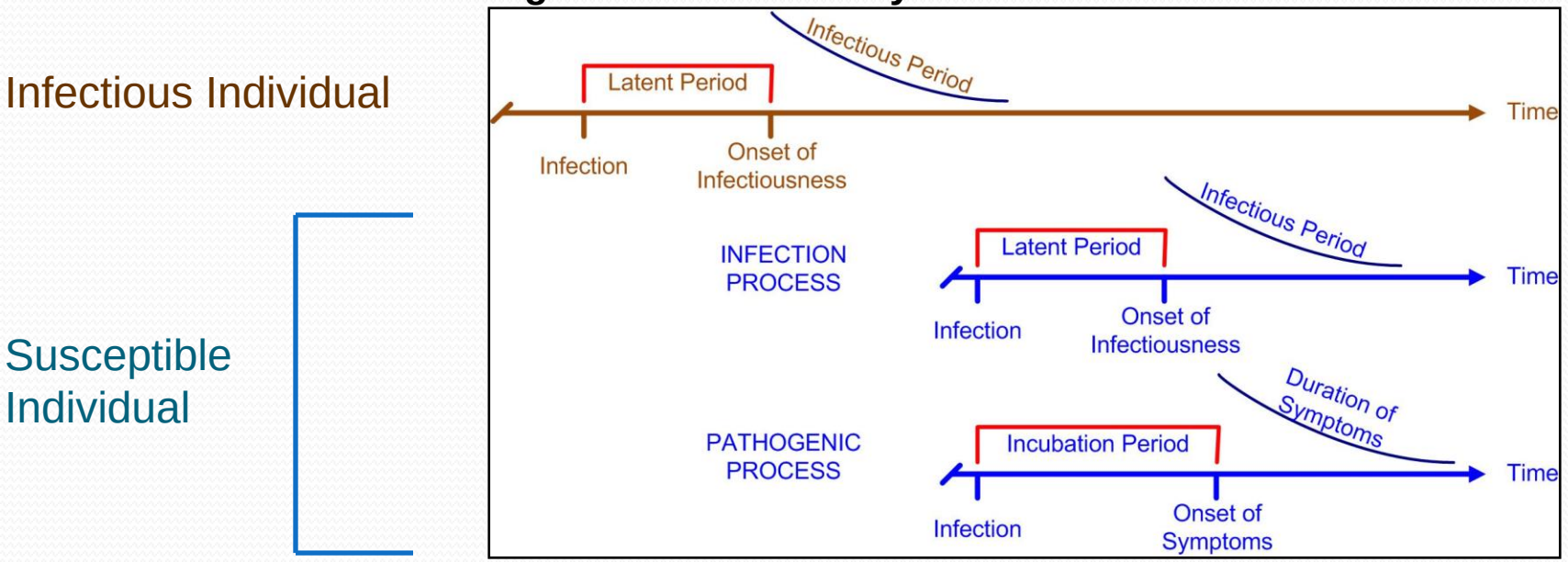
# Motivation

To enable field personnel and researchers to analyze data from local outbreaks of infectious diseases, with the aim of…

- Detecting individual-to-individual (person-to-person) transmission of pathogens
- Evaluating the transmissibility of pathogens
- Evaluating the effects of risk factors and interventions on transmission
- Providing basic summary statistics, such as epidemic curves and case fatality ratios (version 1 only)
- Performing simulation studies, for example, to perform power calculations for study design purposes

# Basic Concepts: Natural History of Infection and Disease

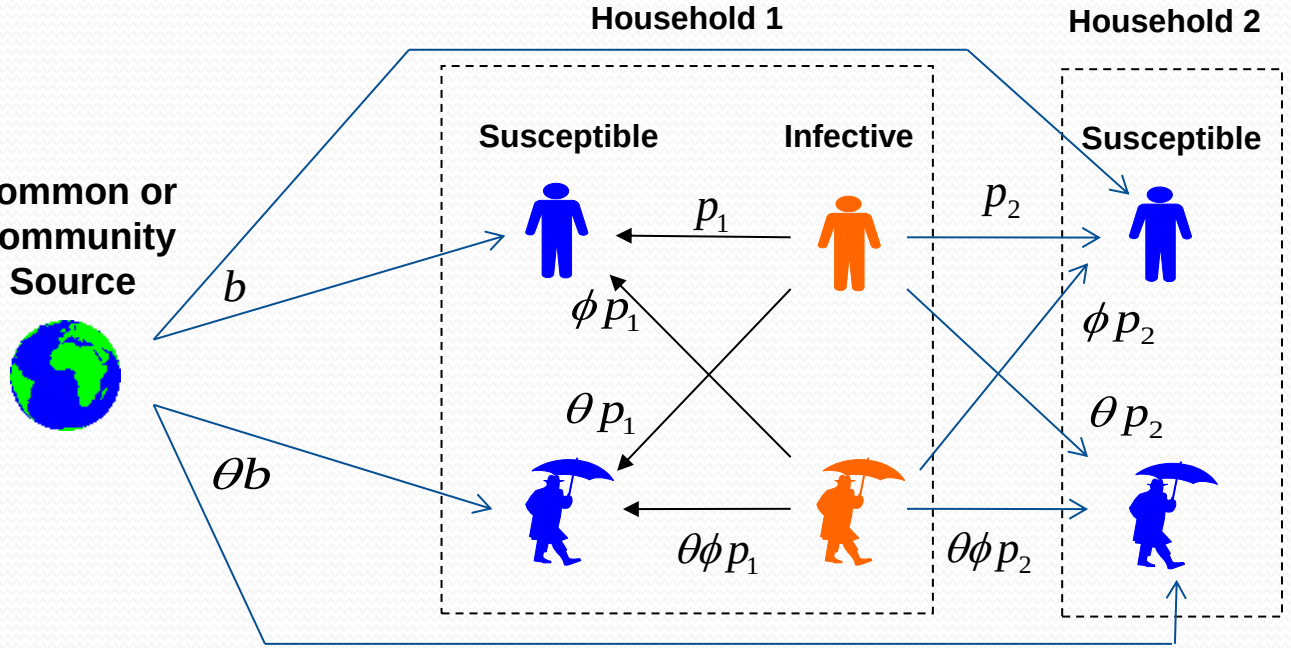
**Figure. Natural History of Infection and Disease**




- Infection depends upon exposure to an infectious individual (see next slide for more details)
- Both concurrently occurring processes are often (or are assumed to be) strongly correlated, for example, onset of symptoms may be assumed to indicate onset of infectiousness
- Individuals do not necessarily complete each process in its entirety, *e.g.*, an individual may become infectious, but never exhibit clinically-apparent symptoms (infectious asymptomatic infection)

# Basic Concepts: Exposure

**Figure. Population and Contact Structure**



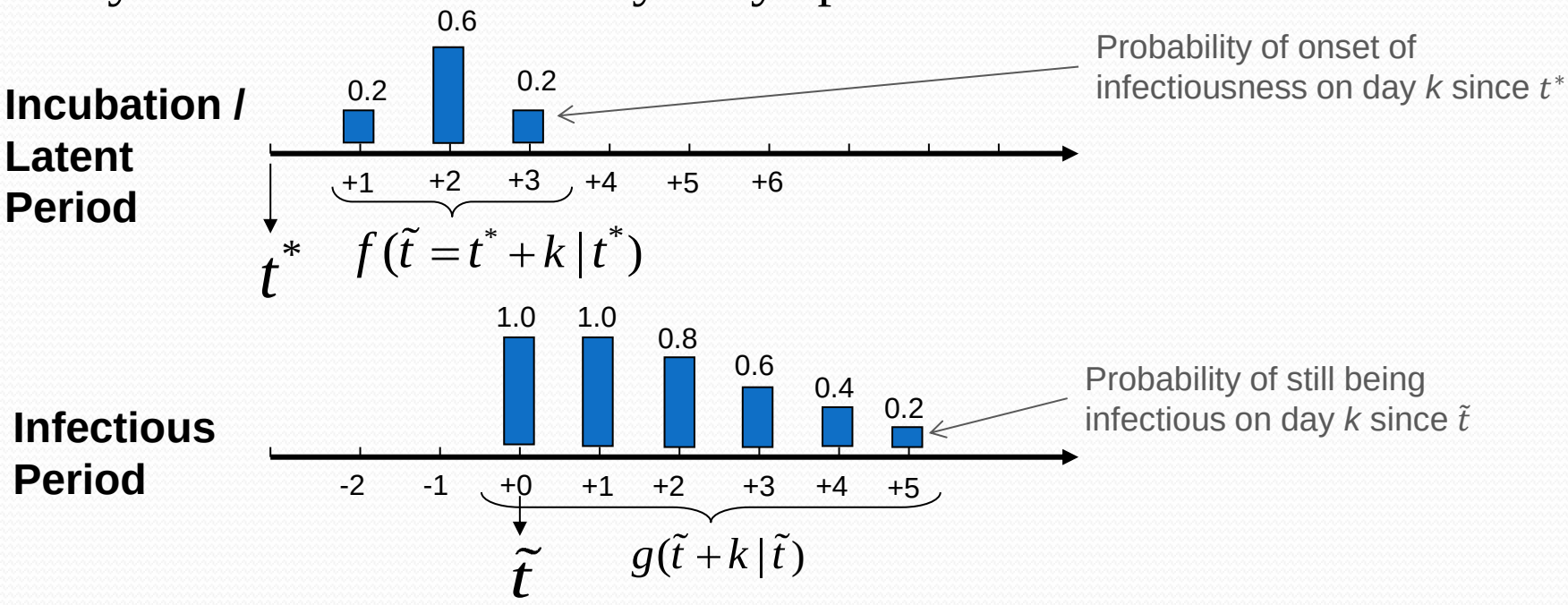

- Contact = exposure to a specific source of infection for a defined period of time (typically, a day)
- 'Household' = general term for clusters of individuals who are more likely to mix with each other than with other members of the population. Multiple types of households may be defined.
- Types of contact and associated transmission probabilities
  - P2P, or person-to-person, exposure to a specific individual: within household, $p_1$, and between household (for example, household in the same neighborhood), $p_2$
  - C2P, or community-to-person exposure to non-specific sources of infection: $b$
- $\theta$ and $\phi$ denote covariate effects (risk-factors or interventions) on susceptibility and infectiousness, respectively.

# Model: Data Inputs

- Individual-level information
  - Household (cluster) membership
  - Covariates: *e.g.*, age or vaccination status
  - Outcome-related information: infection and symptomatic status, onset times, and laboratory test results.
  - Information about any pre-existing immunity to infection
  - Indication of whether or not data is missing for each of the outcome and pre-existing immunity related data inputs
- Household or Cluster level information
  - Population and/or contact structure
  - Beginning and end of observation period for each cluster

# Model: Incubation/Latent and Infectious Period Distributions (assumed known)

$t^*$: day of infection  $\tilde{t}$: day of symptom onset




- These are sample distributions for the incubation/latent and infectious periods.
- This example assumes that onset of symptoms indicates onset of infectiousness, *i.e.*, incubation=latent period.
- TranStat inputs: Minimum and maximum values for *k* and the daily probability distribution (blue bars)

# Likelihood

- $T = \begin{cases} \text{onset of infection}, & \text{infected} \\ \text{end of follow} - \text{up}, & \text{otherwise} \end{cases}$
- Probability that *j* infects *i* during day *t*:
  $logit(p_{ijt}) = logit(p) + \boldsymbol{X_i}\beta_S + \boldsymbol{X_j}\beta_I + \boldsymbol{X_{ij}}'\beta_{SI}, j \in \mathcal{H}_i$
- An important example of interaction:
  - Let $r_i$ be the vaccination status and the only covariate for person *i*
  - $logit(p_{ijt}) = logit(p) + r_i\theta + r_j\phi + r_i r_j\psi$
  - $VE_S = 1 - \theta, VE_I = 1 - \phi, VE_T = 1 - \psi$

# Likelihood (continued)

- Probability that the common/community source infects *i* on day *t*:

$$logit(b_{it}) = logit(b) + \boldsymbol{X_i}\alpha_S$$

- Probability of *i* escaping infection on day *t*:

$$e_{it} = (1 - b_{it}) \prod_{j=1}^{N} \left(1 - p_{ijt} g(t|\tilde{t}_j)\right)$$

- Probability of escaping infection up to day *t*:

$$Q_{it} = \prod_{\tau=1}^{t} e_{it}$$

- Likelihood contribution by *i*:

$$L_i = \begin{cases} Q_{iT}, & \text{infected} \\ \sum_t f\left(\tilde{t}_j \middle| t\right) Q_{i(t-1)}(1 - e_{it}), & \text{otherwise} \end{cases}$$

# Some Statistical Adjustments

- Selection bias: a household is observed only upon ascertainment of an index case
  - Probability of no symptom onset on day $\tilde{t}_{idx}$:

$$L_i^m = \begin{cases} L_i, & i \text{ is index} \\ Q_{i\tilde{t}_{idx}} + \sum_{t<\tilde{t}_{idx}} \Pr(\tilde{t}_i > \tilde{t}_{idx}|t)\, Q_{i(t-1)}(1-e_{it}), & \text{not index} \end{cases}$$

  - Maximize the conditional likelihood, $\prod_i L_i/L_i^m$

- Right censoring: showing no symptoms by day *T* does not necessarily mean that *i* escaped infection.

$$L_i = Q_{iT} + \sum_{t<T} \Pr(\tilde{t}_i > T|t)\, Q_{i(t-1)}(1-e_{it}), \qquad \text{not index}$$

# Other Statistical Features

- Goodness of fit: comparing observed with expected frequency of symptom onset per person-day
- Hypothesis test to detect person-to-person transmission (only available in TranStat version 1) (Yang et al. Annals of Applied Stat, 2007)
  - $H_0: p = 0 \; vs. \; H_1: p \neq 0$
  - Test statistic: $\lambda = -2log \frac{sup_{\boldsymbol{b}} L_o(\boldsymbol{b}|\boldsymbol{t})}{sup_{\boldsymbol{b},\boldsymbol{p}} L(\boldsymbol{b},\boldsymbol{p}|\boldsymbol{t})}$
  - Under $H_0$, permute the symptom onset dates.

# TranStat Version 1

- Can fit simple models with $b$, $p_1$, and $p_2$, but no covariates
- GUI available
- Data input and basic editing functions available
- Sample datasets provided
- No longer under development, so bugs are still present

# TranStat Version 3

- Any number of *b*'s and *p*'s
- Covariate adjustment
- Flexible contact structure
- Accounts for unobserved pre-existing immunity and/or asymptomatic infection
- Accounts for missing data related to infection or symptomatic status, and missing onset times.
- GUI not available

# Input Files for TranStat 3

DO NOT INCLUDE COLUMN TITLES IN ANY TRANSTAT INPUT FILE!

- Household / Cluster profile: “community.dat”

| Household / Cluster ID | Start Observation | End Observation |
|---|---|---|
| 1 | 1 | 45 |
| … | … | … |
| H | 34 | 56 |

- Population profile: “pop.dat”

| Person ID | Cluster ID | Pre-existing Immune Status | Infection Status | Symptomatic Status | Onset Time | Index Case Indicator | Pathogenicity Type | Pre-existing Immunity Type | Ignore Indicator |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 1 | 34 | 1 | 0 | 0 | 0 |
| … | … | … | … | … | … | … | … | … | … |
| N | C | 1 | 0 | 0 | -1 | 0 | 0 | 0 | 0 |

# Input Files for TranStat 3 (continued)

- Time independent covariates: “time_ind_covariate.dat”
  - One line per individual
  - One column per covariate
  - No missing information

| Person ID | Age | Vaccination Status | Gender |
|---|---|---|---|
| 1 | 34 | 0 | 0 |
| … | … | … | … |
| N | 103 | 1 | 1 |

- Time dependent covariates: “time_dep_covariates.dat”
  - One line per individual per time period (a set of one or more contiguous time units)
  - One column per covariate
  - No missing information

| Person ID | Start Time (day) | End Time (day) | Antiviral Prophylaxis |
|---|---|---|---|
| 1 | 1 | 3 | 0 |
| … | … | … | … |
| N | 45 | 56 | 1 |

# Input Files for TranStat 3 (continued)

- C2P contact file: “c2p_contact.dat”
  - C2P contacts can be indexed in three manners
    - no ID, which assumes the same contact history for all individuals

| Start Time (day) | End Time (day) | Type of C2P Contact | Weight | Ignore C2P Contact Indicator |
|---|---|---|---|---|
| 1 | 66 | 0 | 0 | 0 |
| … | … | … | … | … |
| 28 | 28 | 1 | 0.85 | 1 |

    - by cluster ID, which assumes the same contact history for all members of a cluster
    - by person ID, which specifies a separate contact history for each individual

| Cluster or Person ID | Start Time (day) | End Time (day) | Type of C2P Contact | Weight | Ignore C2P Contact Indicator |
|---|---|---|---|---|---|
| 1 | 1 | 66 | 0 | 0 | 0 |
| … | … | … | … | … | … |
| C or N | 28 | 28 | 1 | 0.85 | 1 |

  - Contact types are numbered using consecutive non-negative integers, beginning with 0

# Input Files for TranStat 3 (continued)

- P2P contact file: “p2p_contact.dat”
  - P2P contacts can be indexed in three manners
    - by cluster ID, which assumes the same contact history between all members of a cluster (requires indexing c2p_contact.dat by cluster ID)

| Cluster ID | Start Time (day) | End Time (day) | Type of P2P Contact | Weight | Ignore P2P Contact Indicator |
|---|---|---|---|---|---|
| 1 | 1 | 66 | 0 | 0 | 0 |
| … | … | … | … | … | … |
| C | 28 | 28 | 1 | 0.85 | 1 |

    - by person ID, which specifies a separate contact history between each individual

| Start Time (day) | End Time (day) | Person ID: Infective | Person ID: Susceptible | Type of P2P Contact | Weight | Ignore P2P Contact Indicator |
|---|---|---|---|---|---|---|
| 1 | 66 | 1 | 4 | 0 | 0 | 0 |
| … | … | … | | … | … | … |
| 28 | 28 | N | 66 | 1 | 0.85 | 1 |

  - Contact types are numbered using consecutive non-negative integers, beginning with 0

# Input Files for TranStat 3 (continued)

- Imputation control file: “impute.dat”
  - Include one row per individual for whom at least one outcome or pre-existing immunity related value is missing

| Person ID | Possible Pre-Existing Immunity | Possible Escape | Possible Symptomatic Infection | Start Time for Imputing Symptomatic Infection Onset Time | Stop Time for Imputing Symptomatic Infection Onset Time | Possible Asymptomatic Infection | Start Time for Imputing Asymptomatic Infection Onset Time | Stop Time for Imputing Asymptomatic Infection Onset Time |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | -1 | -1 | 0 | -1 | -1 |
| … | … | … | … | … | … | … | … | … |
| N | 0 | 0 | 1 | 1 | 10 | 1 | 1 | 10 |

# Configuration File

- Natural history of disease, *i.e.*, incubation and infectious periods.
- Profile of parameters to be estimated
  - Numbers of C2P and P2P contact types
  - Numbers of time-independent and time-dependent covariates
- Covariates effects on…
  - Susceptibility due to exposure through…
    - C2P contact
    - P2P contact
  - Infectiousness
  - Interaction between C2P and P2P transmission
- Define equivalence classes of parameters
- Specify which parameters have fix values

# Configuration File (continued)

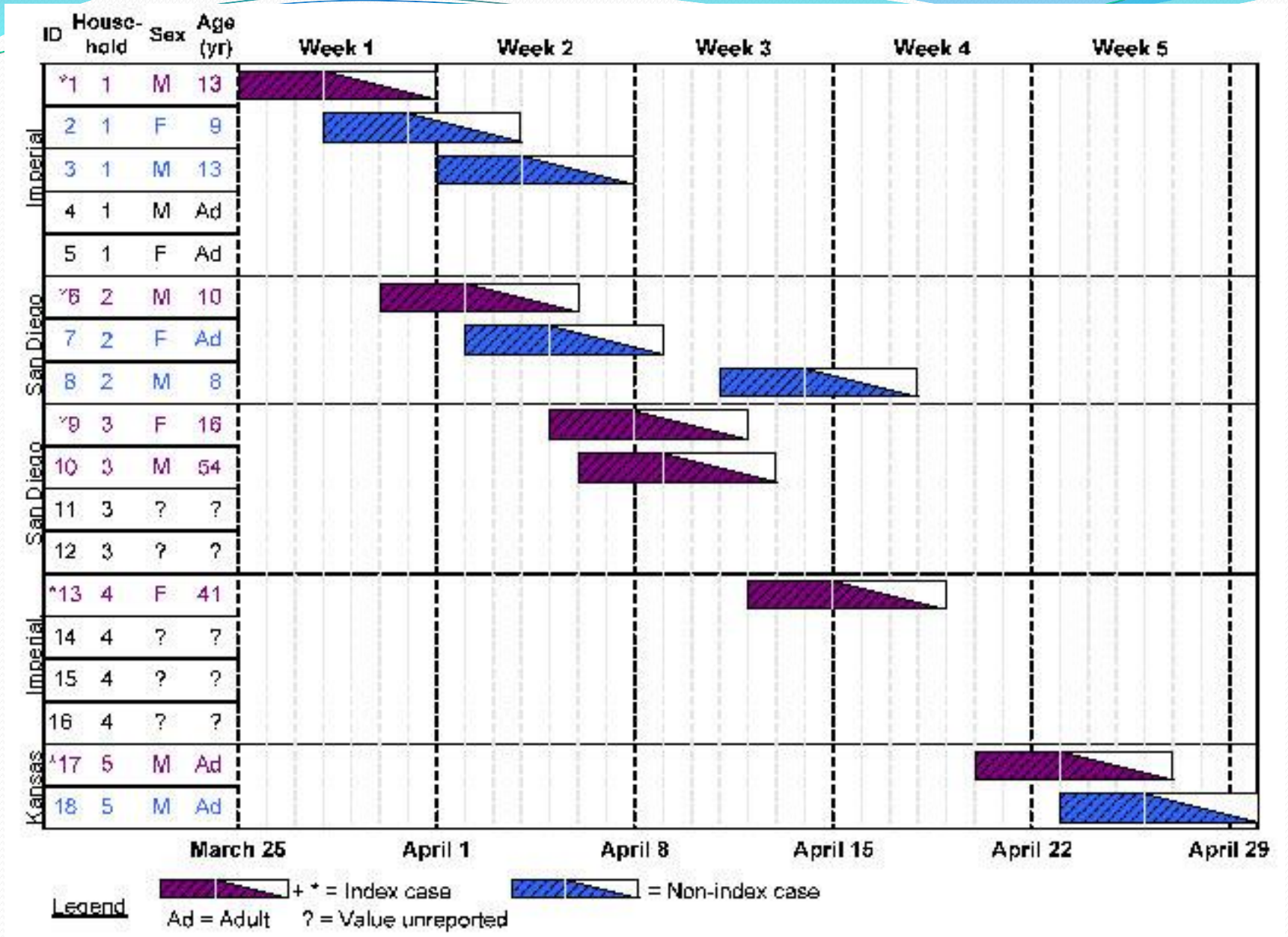
- How should TranStat handle C2P and P2P contact files.
  - Community members share contact history?
  - Community members share risk history (same covariates)?
  - Auto-generate the C2P/P2P contact files?
- Choose whether to adjust for selection bias and/or right censoring
- Choose whether to calculate/perform goodness-of-fit, case fatality ratio (deprecated), hypothesis test (under development), etc.
- Controlling optimization
- Controlling output
- Controlling data augmentation/multiple imputation procedures for missing data for variables related to outcome or pre-existing immunity

# TranStat Output

- Estimates file: “estimates.txt”
  - Outputs estimates, standard errors, and 95% confidence intervals for *b*’s, *p*’s, covariate effects, CPI, SAR’s, $R_0$ (or local R), variance-covariance matrix
  - $CPI_x = 1-(1-b_x)^D$, where *D* is the average duration of exposure to the common source
  - $SAR_x = 1 - \sum_{t=0}^{Z}(1 - g(t|\tilde{t}_j)p_x)$, where *Z* denotes maximum length of the infectious period
- Error file: “error.txt” – list of errors encountered during the estimation process

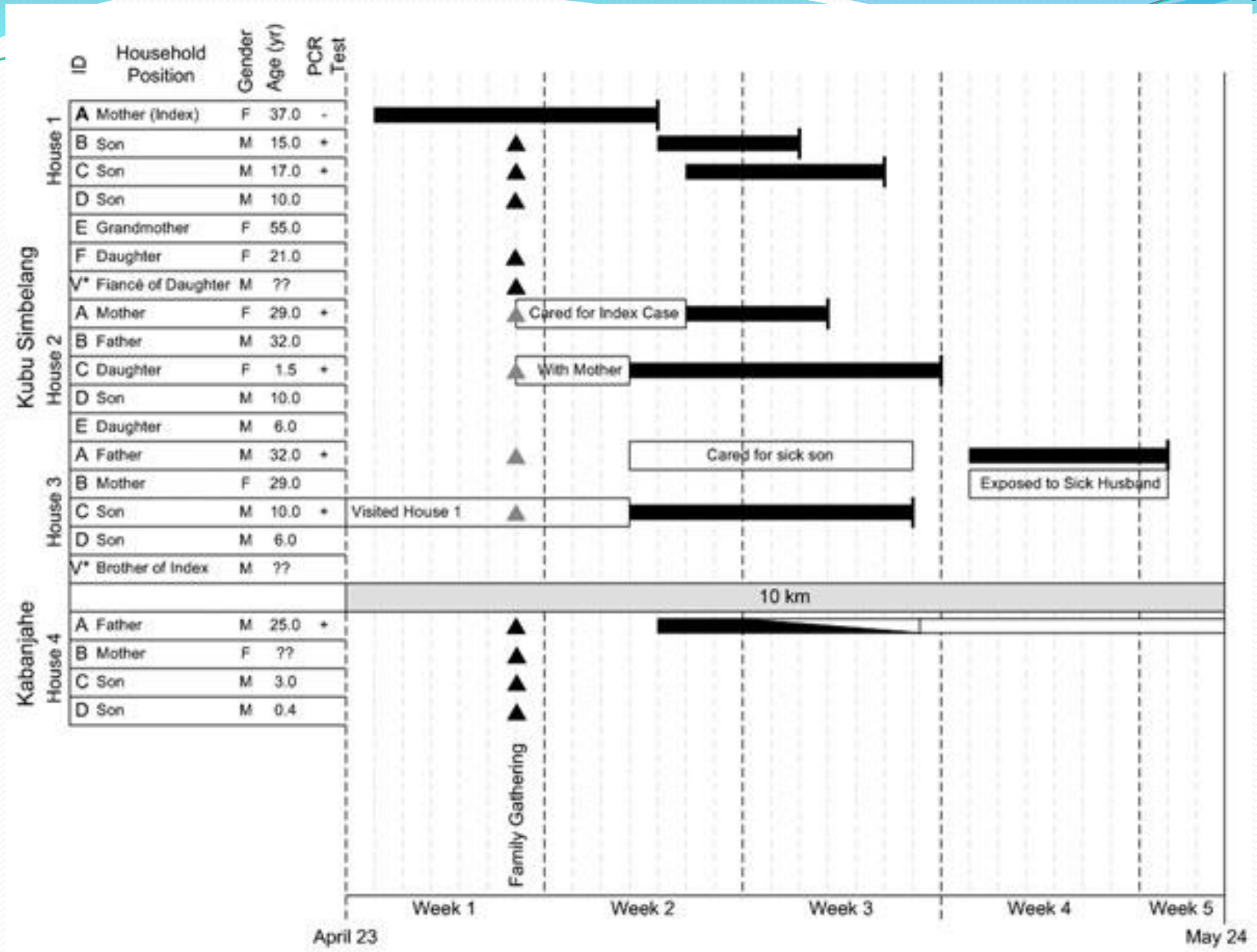
# Case Study 1: US household outbreaks of Influenza A(H1N1) 2009

- Household structure is known => can model within-household transmission.
- Households not in the same neighborhood => can not model inter-household transmission.
- Households can be regarded as independent mini-communities.
- People in the same households share the same history of contact and exposure.


ID
House-hold
Sex
Age (yr)
Week 1
Week 2
Week 3
Week 4
Week 5
Imperial
San Diego
San Diego
Imperial
Kansas
*1 1 M 13
2 1 F 9
3 1 M 13
4 1 M Ad
5 1 F Ad
*6 2 M 10
7 2 F Ad
8 2 M 8
*9 3 F 16
10 3 M 54
11 3 ? ?
12 3 ? ?
*13 4 F 41
14 4 ? ?
15 4 ? ?
16 4 ? ?
*17 5 M Ad
18 5 M Ad
March 25
April 1
April 8
April 15
April 22
April 29
Legend
+ * = Index case
= Non-index case
Ad = Adult
? = Value unreported

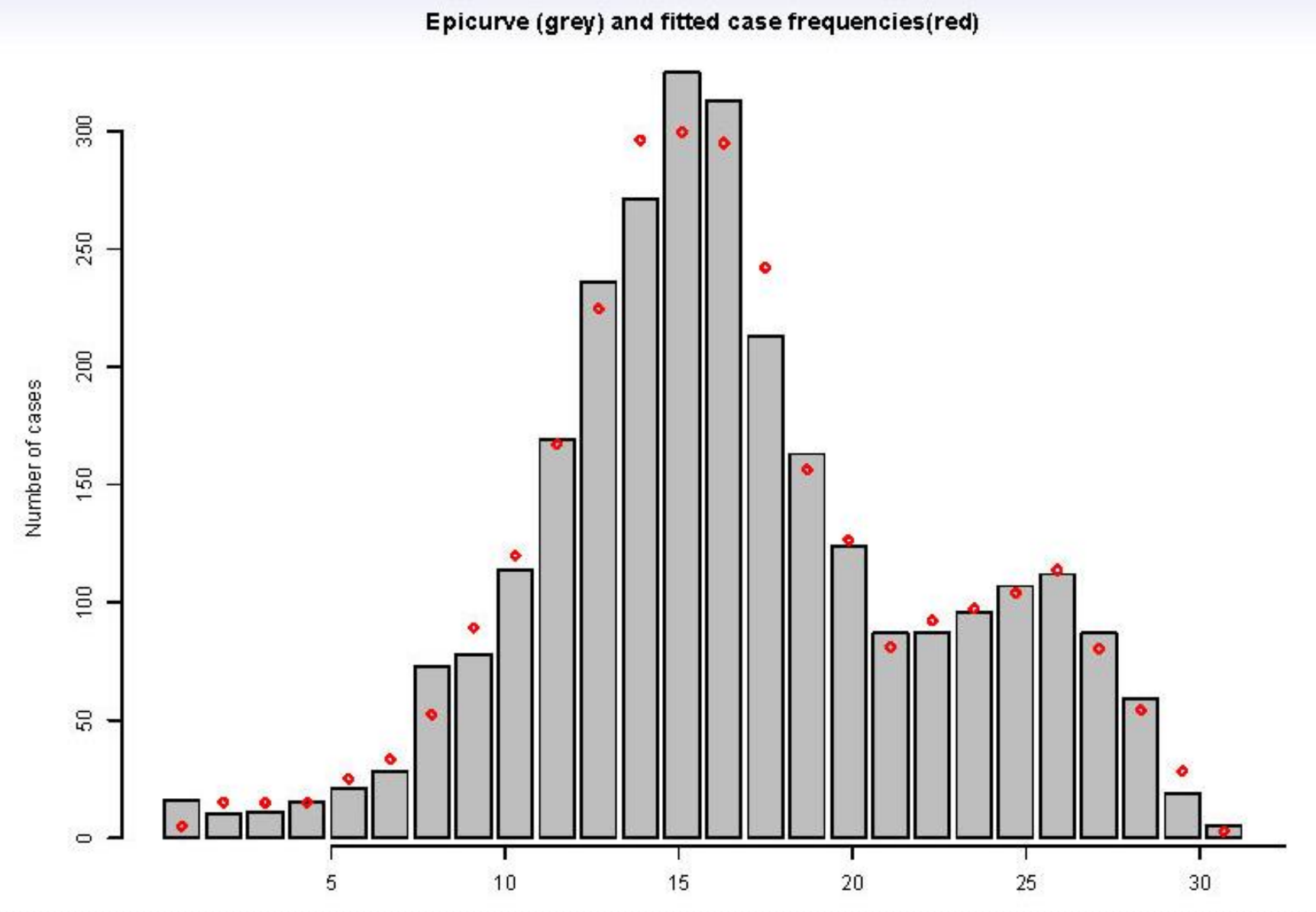
# Case Study 2: Indonesian household outbreaks of avian influenza A(H5N1)

- An outbreak caused by a family gathering of multiple households.
- Transmission occurred both within and between households.
- In TranStat, clusters that have cross-transmission should be considered as a single community.
- Individual level contact and risk history.


ID
Household Position
Gender
Age (yr)
PCR Test
Kubu Simbelang
House 1
A Mother (Index) F 37.0 -
B Son M 15.0 +
C Son M 17.0 +
D Son M 10.0
E Grandmother F 55.0
F Daughter F 21.0
V* Fiancé of Daughter M ??
House 2
A Mother F 29.0 +
B Father M 32.0
C Daughter F 1.5 +
D Son M 10.0
E Daughter M 6.0
House 3
A Father M 32.0 +
B Mother F 29.0
C Son M 10.0 +
D Son M 6.0
V* Brother of Index M ??
Kabanjahe
House 4
A Father M 25.0 +
B Mother F ??
C Son M 3.0
D Son M 0.4
Cared for Index Case
With Mother
Cared for sick son
Exposed to Sick Husband
Visited House 1
10 km
Family Gathering
Week 1
Week 2
Week 3
Week 4
Week 5
April 23
May 24

# Case Study 3: Influenza A(H1N1) 2009 outbreak in Mexico

- People: 2,895 confirmed cases
- Time: March 11–? We use the data up to May 15.
- Case numbers are aggregated by day.
- Contact structure is unknown.
- $R_0$ is estimable:
  - Distribution of serial interval based on all possible transmission networks
  - Chain binomial model


Epicurve (grey) and fitted case frequencies(red)
Number of cases
0
50
100
150
200
250
300
5
10
15
20
25
30

# Case Study 3 (continued)

- For large population, chain binomial model becomes Poisson
- On day t, observe number of susceptibles *S(t)*, infectives *I(t)*, and new infections *X(t)*.

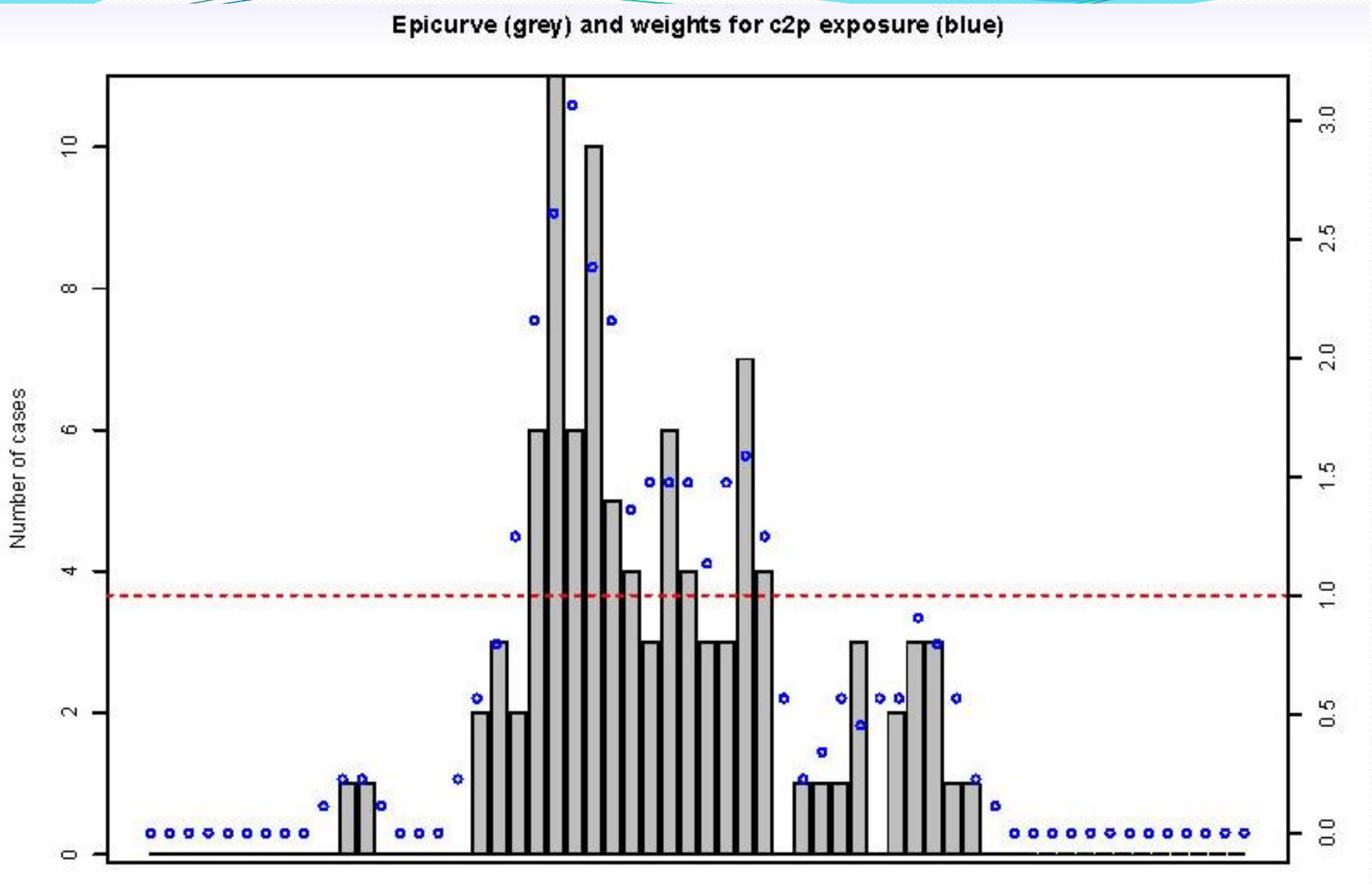
$$\binom{S(t)}{X(t)}\{1-(1-p)^{I(t)}\}^{X(t)}(1-p)^{I(t)S(t+1)}$$

$$\rightarrow \frac{(\lambda I(t))^{X(t)}}{X(t)!}\exp\{-\lambda I(t)\}$$

- $\hat{\lambda}=\frac{\sum_{t=1}^{T}X(t)}{\sum_{t=1}^{T}I(t)}\rightarrow\lambda$ a.s., and $\hat{R}_0=D\hat{\lambda}\rightarrow R_0$ a.s.
- To use TranStat, create D-1 uninfected people for each observed case. D = 100 is sufficiently large.

# Case study 4: Influenza A(H1N1) 2009 household outbreaks in Los Angeles

- A total of 58 households with $\geq$ 1 cases, non-random sample.
- 60 index cases and 37 secondary cases.
- All index cases were laboratory confirmed with either pandemic H1N1 or seasonal influenza A.
- Outbreaks started from April 22 to May 19, 2009.
- Ages are known for all, and seasonal flu vaccine and antiviral treatment are known for part of the surveyed population.
- Missing information:
  - asymptomatic infection
  - pre-existing immunity: Assumed to be non-existent in this population, because this strain of influenza A was first described in humans during the spring of 2009.
- EM-MCEM (Yang et al., Biometrics 2012)


Epicurve (grey) and weights for c2p exposure (blue)
Number of cases
0
2
4
6
8
10
0.0
0.5
1.0
1.5
2.0
2.5
3.0

# Summary

- Current progress
  - Simulation component for statistical power calculations
  - More efficient algorithm for multiple imputation of outcome and pre-existing immunity related quantities
- Future improvements
  - GUI
  - Permutation test for P2P transmission
  - Multiple imputation for covariates
  - More statistical models/methods
    - Final-size models (Longini and Koopman, Biometrics, 1982; Addy et al., Biometrics, 1991)
    - Branching process
    - Bayesian approaches

# Exercise

- In Case Study 4, we fit a model with one p and one b.
- Exercise: The following are independent tasks.
  a) Estimate the effect of the binary covariate (age-group) in the time independent covariate file for Case Study 4. Estimate a single effect of age-group on susceptibility. Report the covariate effect, *p*, *b*, and pathogenicity proportion, along with 95% confidence intervals.
  b) Using the Case Study 4 data, fit a model with only a single *p* and *b*, but exclude households with Cluster ID's = 2, 17, and 18.
  c) Bonus: The members of the population for Case Study 4 with the following Person IDs were determined to be potential asymptomatic infections: 21, 45, 93, 98, 200, 206, 215, and 269. The laboratory-test results for these 8 individuals are missing, so we must impute both their outcome status (either escaped or asymptomatic infection), as well as their onsets dates. Assume that asymptomatic infections are 67% as infectious as symptomatic infections. Fit a model that includes a *p*, *b*, and the pathogenicity proportion, but also accounts for the unknown outcome status of these 8 individuals. Allow imputation of onset dates for these 8 individuals from day 1 to 45.

# Real-time analysis of infectious disease outbreaks using TranStat

EXERCISE REVIEW

# Exercise Answer Key

- In Case Study 4, we fit a model with one p and one b.
- Exercise for Lecture 10 (additional assistance can be found by consulting Tutorial A on the TranStat website). The following are independent tasks.
  a) Estimate the effect of the binary covariate (age-group) in the time independent covariate file for Case Study 4. Estimate a single effect of age-group on susceptibility. Report the covariate effect, *p*, *b*, and pathogenicity proportion, along with 95% confidence intervals.
  b) Using the Case Study 4 data, fit a model with only a single *p* and *b*, but exclude households with Cluster ID's = 2, 17, and 18.
  c) Bonus: The members of the population for Case Study 4 with the following Person IDs were determined to be potential asymptomatic infections: 21, 45, 93, 98, 200, 206, 215, and 269. The laboratory-test results for these 8 individuals are missing, so we must impute both their outcome status (either escaped or asymptomatic infection), as well as their onsets dates. Assume that asymptomatic infections are 67% as infectious as symptomatic infections. Fit a model that includes a *p*, *b*, and the pathogenicity proportion, but also accounts for the unknown outcome status of these 8 individuals.

# Exercise Answer Key

a) Estimate the effect of the binary covariate (age-group) in the time independent covariate file for Case Study 4. Estimate a single effect of age-group on susceptibility. Report the covariate effect, *p*, *b*, and pathogenicity proportion, along with 95% confidence intervals.

**Approach**

1. Save a copy of the default config.file and pop.dat.
2. Make the following changes to the config.file. The new values are shown in bold-face font.
3. Run TranStat with this updated version of the config.file and the default versions of the pop.dat, community.dat, c2p_contact.dat, p2p_contact.dat, and time_ind_covariated.dat.

**Changes to config.file**

…
# number-of-pathogenicity-groups
0 -> **1**
…
# number-of-time-independent-covariates
0 -> **1**
…
# covariates-affecting-susceptibility-for-c2p-transmission
0: -> **1: 1**

# covariates-affecting-susceptibility-for-p2p-transmission
0: -> **1: 1**
…
# equal-parameters
2 -> **4**
1: 1
1: 2
**1: 3**
**2: 4 5**
…

…
# converge-criteria
1
b 1e-5
p 1e-5
**u 1e-5**
**or1 1e-5**

# initial-estimates
5:0

# search-bounds
1
b 0.0000000001 0.5
p 0.000000001 0.5
**u 0.000000001 1.5**
**or1 0.00000001 150**

# Exercise Answer Key

b) Using the Case Study 4 data, fit a model with only a single *p* and *b*, but exclude households with Cluster ID's = 2, 17, and 18.

Approach

1. For each individual member of the clusters 2, 17, and 18 (individuals: 16-22 and 105-113), change the value for the Ignore indicator in the pop.dat (last column on the right) from the default value of 0 to the value of 1.
2. Run TranStat with this updated version of the pop.dat and the default versions of the config.file, community.dat, c2p_contact.dat, p2p_contact.dat, and time_ind_covariated.dat.

# Exercise Answer Key

c) Bonus: The members of the population for Case Study 4 with the following Person IDs were determined to be potential asymptomatic infections: 21, 45, 93, 98, 200, 206, 215, and 269. The laboratory-test results for these 8 individuals are missing, so we must impute both their outcome status (either escaped or asymptomatic infection), as well as their onsets dates. Assume that asymptomatic infections are 67% as infectious as symptomatic infections. Fit a model that includes a *p*, *b*, and the pathogenicity proportion, but also accounts for the unknown outcome status of these 8 individuals.

**Approach**

1. I have already constructed the impute.dat file for you.
2. Using the default config.file, make the following changes.
3. Run TranStat with this updated version of the config.file and the default versions of the pop.dat, community.dat, c2p_contact.dat, p2p_contact.dat, and time_ind_covariated.dat.

**Changes to config.file**

…
# number-of-pathogenicity-groups
0 -> **1**
…
# equal-parameters
2 -> **3**
1: 1
1: 2
**1: 3**
…
# converge-criteria
1
b 1e-5
p 1e-5
**u 1e-5**
…
# search-bounds
1
b 0.0000000001 0.5
p 0.000000001 0.5
**u 0.000000001 1.5**

# perform-EM-algorithm
0 -> **1**
…
# relative-infectivity-of-asymptomatic-case-for-estimation
0 -> **0.67**
…