Chain Binomial Model in TranStat

Yang Yang

Department of Biostatistics & Emerging Pathogens Institute
University of Florida
Data Hierarchy

- Number of infections and susceptibles at the beginning and the end of epidemic (Final-size binomial models).
  - Regression models
    - Generalized Estimating Equations
    - Mixed effects logistic regression
  - Iterative binomial models
- Disease onset times plus disease natural history.
  - Chain-binomial likelihood models, e.g., Rampey et. al.(1992), Yang et. al. (2006, 2008).
  - Transmission-network-based survival models, e.g., Kenah (2010)
- Contact structure and/or exposure history
Data Hierarchy (continue)

- Laboratory confirmation of a single pathogen.
  - Chain-binomial likelihood model with EM-MCEM or Bayesian model, e.g., Cauchemez et al. (2004), Yang et al. (2008, 2012).

- Laboratory confirmation of multiple co-circulating pathogens.
  - Bayesian transmission models, e.g., Auranen et al. (2000), Yang et al. (2010, 2019).
Contact-based modeling

- A contact is specific to
  - A time unit (e.g., a day)
  - A setting of mixing (e.g., household)

- Types of contacts (source of transmission)
  - Person-to-person (P2P):
    Close contact with specific infectious individuals in mixing groups, e.g., households, neighborhoods, schools, hospitals.
  - Common-source-to-person (C2P) contact:
    Contact with unobserved nonspecific individuals, zoonotic sources, or environmental reservoir.
Transmission Patterns and Parameters of Interest

Household 1

Susceptible

Infective

Household 2

Susceptible

Common Source

b

θb

θ p_1

φ p_1

p_1

p_2

θ p_2

φ p_2

θφ p_1

θφ p_2
Time of Infection

<table>
<thead>
<tr>
<th>Incubation period</th>
<th>Infectious period</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.6</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>0.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Onset time of symptoms and infectiousness

\[ \Pr(\tilde{t}|t) = f(\tilde{t} - t) \quad l_{\text{min}} \leq \tilde{t} - t \leq l_{\text{max}} \]

\[ \Pr(t|\tilde{t}) = g(t - \tilde{t}) \]

\( f(\tilde{t} - t) \): Probability of symptom onset on day \( \tilde{t} \) given infection on day \( t \).

\( g(t - \tilde{t}) \): Probability of host being infective on day \( t \) given symptom onset on day \( t \).
Chain binomial model (continue)

Probability that the common source or person $j$ infects $i$ in day $t$:

$$\logit(b_{it}) = \logit(b) + x_{it}'\beta_b$$

$$\logit(p_{ijt}) = \begin{cases} 
\logit(p_1) + x_{ijt}'\beta_{p_1}, & H_i = H_j \\
\logit(p_2) + x_{ijt}'\beta_{p_2}, & H_i \neq H_j, N_i = N_j \\
0, & \text{otherwise.} 
\end{cases}$$

An example of covariate adjustment:

- let $r_i$ be vaccine status of person $i$ and be the only covariate.
- $$\logit(p_{ijt}) = \logit(p_k) + r_i\theta + r_j\phi + r_i r_j\psi$$
- $$\text{VE}_S = 1 - e^\theta, \text{VE}_I = 1 - e^\phi \text{ and } \text{VE}_T = 1 - e^{\theta+\phi+\psi}.$$
Chain binomial model (continue)

Probability of escaping infection on day $t$, $t = 1, \ldots, T$:

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

Probability of escape up to day $t$: $Q_{it} = \prod_{\tau=1}^{t} e_{i\tau}$

Probability of escape before and infection on day $t$:

$U_{it} = Q_{i,t-1}(1 - e_{it})$

let $t_i = \tilde{t}_i - l_{\text{max}}$, $\overline{t}_i = \tilde{t}_i - l_{\text{min}}$. Likelihood contributed by $i$ is:

$$L_i = \begin{cases} Q_{iT}, & \text{escaped} \\ U_{it}, & \text{if known to be infected on } t \\ \sum_{t_i = \overline{t}_i}^{\tilde{t}_i} U_{it} \times f(\tilde{t}_i - t), & \text{if symptom onset on } \tilde{t}_i \end{cases}$$

Secondary attack rate: $\text{SAR}_k = 1 - \prod_{d=0}^{D-1} [1 - p_k g(d)]$, $k = 1, 2$.

Effective reproductive number: $R_0 = \sum_k n_k \times \text{SAR}_k$. 
Missing Data Patterns

(a) Univariate Nonresponse

(b) Multivariate Two Patterns

(c) Monotone

(d) General

(e) File Matching

(f) Factor Analysis
Mechanisms of Generating Missing Data

Suppose data \((X_i, Y_i, M_i)\) is generated (iid across all \(i\)) via
\[
f(X_i, Y_i, M_i | \theta, \phi) = f(X_i, Y_i | \theta)f(M_i | X_i, Y_i, \phi), \ i = 1, \ldots, N.
\]
Suppose \(X_i\) is always observed, and \(Y_i\) is sometimes observed as indicated by \(M_i\) (1=missing, 0=observed). Can we estimate \(\mathbb{E}(Y|X = x)\) by
\[
\frac{1}{\sum_{i=1}^{n} 1(M_i = 0, X_i = x)} \sum_{i=1}^{n} 1(M_i = 0, X_i = x) Y_i?
\]

- **Missing completely at random (MCAR):**
  
  \(f(M_i | X_i, Y_i, \phi) = f(M_i | \phi).\) Clearly, \(\mathbb{E}(Y_i|X_i, M_i = 0) = \mathbb{E}(Y_i|X_i)\)
because \(f(Y_i|X_i, M_i) = f(Y_i|X_i).\)

- **Missing at random (MAR):** \(f(M_i | X_i, Y_i, \phi) = f(M_i | X_i, \phi).\)
  We still have \(\mathbb{E}(Y_i|X_i, M_i = 0) = \mathbb{E}(Y_i|X_i)\) because
  \[
  f(Y_i | X_i, M_i) = \frac{f(Y_i, X_i, M_i)}{f(X_i, M_i)} = \frac{f(X_i, Y_i)f(M_i | X_i, Y_i)}{f(X_i)f(M_i | X_i)} = f(Y_i | X_i).
  \]

- **Missing not at random (MNAR):** \(f(M_i | X_i, Y_i, \phi)\) cannot be further reduced.
Methods for handling Missing Data

- Using completely observed data only. Valid under MCAR, but not necessarily under MAR, e.g., using 
  \[
  \frac{1}{\sum_{i=1}^{n} \mathbf{1}(M_i=0)} \sum_{i=1}^{n} \mathbf{1}(M_i = 0)Y_i
  \]
  to estimate \( \mathbb{E}(Y) \).

- Weighting by inverse probability of not missing. Valid under both MCAR and MAR.

- One-time imputation (hot deck imputation, mean imputation, carry-forward) and multiple imputation.

- Model-based approaches to integrate out missing values, e.g., Expectation-Maximization algorithm, Markov chain Monte Carlo.
Likelihood Model with Data Augmentation

Augmenting data with unobserved quantities (Yang, Longini and Halloran, 2007)

- Pairwise transmission outcome $Y_{ji}(t)$ (1: transmission, 0: escape).
- $Y_{ji}(t)$ is defined only if $Y_{ji}(\tau) = 0$ for all $\tau < t$.
- $Y_{ji}(t)$ is not observed when $j$ is infectious and $t_i \leq t \leq \bar{t}_i$.
- $Y_{ji}(t)$ is independent of $Y_{ki}(t)$ for the same day $t$.
- More convenient to work with
  
  $$Z_{ji}(t) = Y_{ji}(t) \prod_{k \in D_i, \tau < t} \left(1 - Y_{ki}(\tau)\right)$$
  and
  $$\bar{Z}_{ji}(t) = \left(1 - Y_{ji}(t)\right) \prod_{k \in D_i, \tau < t} \left(1 - Y_{ki}(\tau)\right),$$
  where $D_i$ is the collection of potential infective sources for person $i$. 
### Exposure Outcome

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Outcome</th>
<th>Expected Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i \leftarrow j$</td>
<td>(1: transmission, 0: escape)</td>
<td>$\tilde{Z}_{21}(\tilde{t}_1 - 1)</td>
</tr>
<tr>
<td>$1 \leftarrow 2$</td>
<td>$0$</td>
<td>$Z_{21}(\tilde{t}_1 - 1)</td>
</tr>
<tr>
<td>$1 \leftarrow c$</td>
<td>$0$</td>
<td>$\tilde{Z}_{c1}(\tilde{t}_1 - 1)</td>
</tr>
<tr>
<td>$1 \leftarrow c$</td>
<td>$1$</td>
<td>$Z_{c1}(\tilde{t}_1 - 1)</td>
</tr>
</tbody>
</table>
Likelihood Model with Data Augmentation (continue)

Let $Z_i(t) = \max_{j \in D_i} Z_{ji}(t)$ indicates if $Z_{ji}(t) = 1$ for any $j$ on day $t$. The likelihood of the augmented data is

$$L_i(b, p, \theta, \phi | \tilde{t}_j, Z_{ji}(t), \tilde{Z}_{ji}(t), j \in D_i, t \leq T)$$

$$= \prod_{t=1}^{T} \left\{ g(\tilde{t}_i | t)^{Z_i(t)} \prod_{j \in D_i} (p_{ji}(t))^{Z_{ji}(t)} (1 - p_{ji}(t))^{\tilde{Z}_{ji}(t)} \right\},$$

The log-likelihood is

$$\log(L_i(b, p, \theta, \phi | \tilde{t}_j, Z_{ji}(t), \tilde{Z}_{ji}(t), j \in D_i, t \leq T))$$

$$\propto \sum_{t=1}^{T} \sum_{j \in D_i} \left\{ Z_{ji}(t) \log(p_{ji}(t)) + \tilde{Z}_{ji}(t) \log(1 - p_{ji}(t)) \right\},$$
The E-M algorithm: Define the following events

- $S_i(t)$: $i$ has symptom onset on day $t$.
- $I_i(t)$: $i$ is infected on day $t$.
- $I_{ji}(t)$: $j$ infects $i$ on day $t$.

whose probabilities are given by

\[
\Pr[I_{ji}(t)] = \hat{Q}_i(t-1)p_{ji}(t)
\]
\[
\Pr[I_i(t)] = \hat{Q}_i(t-1)\{1 - \hat{e}_i(t)\},
\]
\[
\Pr[S_i(\tilde{t}_i)] = \sum_{\tau=\tilde{t}_i}^{\tilde{t}_i} f(\tilde{t}_i - \tau) \times \Pr[I_i(\tau)],
\]
Likelihood Model with Data Augmentation (continue)

The conditional distributions of $Z_{ji}(t)$ and $\tilde{Z}_{ji}(t)$ are

$$\Pr(Z_{ji}(t) = 1| b, p, \theta, \phi, \tilde{t}_i) = \begin{cases} \frac{\Pr[I_{ji}(t)] \times f(\tilde{t}_i-t)}{\Pr[S_i(\tilde{t}_i)]}, & \tilde{t}_i \leq t \leq \bar{t}_i, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$\Pr(\tilde{Z}_{ji}(t) = 1| b, p, \theta, \phi, \tilde{t}_i) = \begin{cases} \frac{f(\tilde{t}_i-t) \times \left\{ \Pr[I_i(t)] - \Pr[I_{ji}(t)] \right\}}{\Pr[S_i(\tilde{t}_i)]} + \sum_{\tau=t+1}^{\tilde{t}_i} \frac{f(\tilde{t}_i-\tau) \times \Pr[I_i(\tau)]}{\Pr[S_i(\tilde{t}_i)]}, & \tilde{t}_i \leq t \leq \bar{t}_i, \\ 1, & t < \tilde{t}_i, \\ 0, & \text{otherwise}. \end{cases}$$
EM-MCEM: Account for Missing Outcomes

- Independent clusters (households) each of size $n_h$, $h = 1, \ldots, H$.
- Let $\tilde{t}_i$ be infectiousness onset time if $i$ is infected. This is also the symptom onset time if $i$ is a symptomatic case.
- Partition the population into four final states according to $z_i$ (preseason immune status), $y_i$ (infection status), and $s_i$ (symptom status):
  1. prior immunity ($z_i = 1, y_i = 0, s_i = 0$)
  2. susceptible, but escaped infection ($z_i = 0, y_i = 0, s_i = 0$)
  3. symptomatic infection ($z_i = 0, y_i = 1, s_i = 1$), $\tilde{t}_i$ observed.
  4. asymptomatic infection ($z_i = 0, y_i = 1, s_i = 0$), $\tilde{t}_i$ not observed.
- $u_i = (z_i, y_i, s_i, \tilde{t}_i)$ is the complete individual data.
\begin{itemize}
  \item $z_i \sim \text{Bernolli}(\alpha)$, $\alpha$: proportion of pre-immunity.
  \item $s_i \sim \text{Bernolli}(\phi)$, $\phi$: probability of symptoms if infected.
  \item Covariates are adjusted for via logistic regressions:
    \begin{align*}
      \logit(\alpha_i) &= \logit(\alpha) + \mathbf{x}_i^\top \beta_\alpha, \\
      \logit(\phi_{it}) &= \logit(\phi) + \mathbf{x}_{it}^\top \beta_\phi, \\
      \logit(b_{it}) &= \logit(b) + \mathbf{x}_{it}^\top \beta_b, \\
      \logit(p_{ijt}) &= \logit(p) + \mathbf{x}_{ijt}^\top \beta_p.
    \end{align*}
  \item Probabilities of escaping infection on and up to day $t$ are
    
    $e_{it} = (1 - b_{it}) \prod_{j:c_j = c_i} \left(1 - \theta^1 - s_j p_{ijt} g(t - \tilde{t}_j + 1)\right)$ and
    
    $Q_{it} = \prod_{t=1}^{T} e_{il}$
  \item Let $\psi = \{b, p, \alpha, \phi, \beta_\alpha, \beta_\phi, \beta_b, \beta_p\}$. The likelihood contributed by individual $i$:
    \begin{align*}
      L(i)(\psi|\mathbf{u}_i) &= \alpha_i^{z_i} (1 - \alpha_i)^{1 - z_i} \times \\
      &\left\{Q_{iT}^{1 - y_i} \left(\sum_{t} f(\tilde{t}_i - t) Q_{i(t-1)} (1 - e_{it})(\phi_{it})^{s_i}(1 - \phi_{it})^{1 - s_i}\right)^{y_i}\right\}^{1 - z_i}.
    \end{align*}
\end{itemize}
For each household $h$ define collections of observed and missing data

- $O_h = \{u_i : c_i = h \text{ and } u_i \text{ is completely observed}\}$,
- $U_h = \{u_i : c_i = h \text{ and } u_i \text{ is not completely observed}\}$,

For the population, define $O = \{O_h : h = 1, \ldots, H\}$ and $U = \{U_h : h = 1, \ldots, H\}$.

Assuming independence between households, the household-level and population-level likelihoods based on the complete data are

$$L_h(\psi | O_h, U_h) = \prod_{i : c_i = h} L(i)(\psi | u_i)$$

and

$$L(\psi | O, U) = \prod_{h=1}^{H} L_h(\psi | O_h, U_h).$$
• Imputation of $U_h$ is performed at the household level.
• Let $\{U^*_{hk} : k = 1, \ldots, \delta_h\}$ be the collection of all possible realizations of $U_h$ for household $h$, $h = 1, \ldots, H$.
• $\delta_h = 1$ if $U_h$ is empty. For non-empty $U_h$, we expect $\delta_h > 1$.
  Example: If $\tilde{t}_i$ of individual $i$ of household $h$ is the only unobserved quantity in that household, then the range of $U_h$ is determined by $1 + l \leq \tilde{t}_i \leq T$ and $\delta_h = T - l$, where $l$ here is the minimum duration of the latent period.
• $\delta_h$ could be too large for the EM algorithm to enumerate all possibilities.
• We propose to use EM whenever affordable, and Monte Carlo EM otherwise. The MCEM is based on importance sampling (Levine and Casella, 2001).
Choose $J$ to partition the households into three groups:
$\Delta_{OBS} = \{ h : \delta_h = 1 \}$, $\Delta_{EM} = \{ h : 1 < \delta_h < J \}$, and
$\Delta_{MCEM} = \{ h : \delta_h \geq J \}$.

Choose a large integer $K$ to be the number of importance samples for the MCEM algorithm.

Choose an initial value $\psi^{(0)}$ for $\psi$. For household $h \in \Delta_{MCEM}$, draw $K$ samples of $U_h$ from $\Pr(U_h | O_h, \psi^{(0)})$ using a MCMC algorithm, and let these samples be $\hat{U}_{hk}$, $k = 1, \ldots, K$.

Set $\hat{\psi}^{(0)} = \psi^{(0)}$
At iteration $r \geq 0$,

- **i-** Update the conditional probabilities for all $h \in \Delta_{EM}$:

\[
\lambda_{hk}^{(r)} = \frac{L_h(\hat{\psi}^{(r)}|O_h, U_{hk}^*)}{\sum_{l=1}^{\delta_h} L_h(\hat{\psi}^{(r)}|O_h, U_{hl}^*)}, 
\]

$k = 1, \ldots, \delta_h$,

and the importance weights for all $h \in \Delta_{MCEM}$:

\[
\omega_{hk}^{(r)} = \frac{L_h(\hat{\psi}^{(r)}|O_h, \hat{U}_{hk})}{L_h(\psi^{(0)}|O_h, \hat{U}_{hk})}, 
\]

$k = 1, \ldots, K$.

- **ii-** Maximize

\[
\Omega(\psi, \hat{\psi}^{(r)}) = \sum_{h \in \Delta_{OBS}} \ln L_h(\psi|O_h) + \sum_{h \in \Delta_{EM}} \sum_{k=1}^{\delta_h} \lambda_{hk}^{(r)} \ln L_h(\psi|O_h, U_{hk}^*) 
\]

\[
\quad + \sum_{h \in \Delta_{MCEM}} \frac{1}{\omega_h} \sum_{k=1}^{K} \omega_{hk}^{(r)} \ln L_h(\psi|O_h, \hat{U}_{hk}) 
\]

with regard to $\psi$ to find $\hat{\psi}^{(r+1)}$, where $\omega_h^{(r)} = \sum_{k=1}^{K} \omega_{hk}^{(r)}$.

Repeat this step until convergence in the estimates of $\psi$, and denote the final estimate by $\hat{\psi}$.\]
Variance estimation

- For point estimation, one maximizes at each EM iteration

\[ E_{U|O,\hat{\psi}^{(r)}} \ln L(\psi|O, U) = \sum_{h=1}^{H} E_{U_h|O_h,\hat{\psi}^{(r)}} \ln L(\psi|O_h, U_h). \]

- Variance estimation requires evaluation of

\[ E_{U|O,\hat{\psi}} \left( \frac{d \ln L(\psi|O, U)}{d \psi} \right) \left( \frac{d \ln L(\psi|O, U)}{d \psi} \right)^\tau \neq \left( E_{U|O,\hat{\psi}} \frac{d \ln L(\psi|O, U)}{d \psi} \right) \times \left( E_{U|O,\hat{\psi}} \frac{d \ln L(\psi|O, U)}{d \psi} \right)^\tau. \]

- we generate \( K \) new importance samples based on any parameter value \( \tilde{\psi} \approx \hat{\psi} \) for all households in \( \Delta_{EM} \cup \Delta_{MCEM} \).

- If \( K \) is sufficiently large, the algorithm should converge to the same final estimates \( \hat{\psi} \).
• Denote the new importance samples by \( \tilde{U}_{hk}, k = 1, \ldots, K, \)
\( h \in \Delta_{EM} \cup \Delta_{MCEM}. \)

• Let \( \tilde{U}.k = \{ \tilde{U}_{hk} : h = 1, \ldots, H \} \). and
\( \tilde{\omega}_k = \frac{L(\hat{\psi}|O, \tilde{U}.k)}{L(\hat{\psi}|O, \tilde{U}.k)} \), \( k = 1, \ldots, K. \)

• Ignoring the MC error, the covariance matrix of \( \hat{\psi} \) is estimated by (Louis, 1982):

\[
\begin{align*}
\hat{V}^{-1}(\hat{\psi}, \tilde{\psi}, \tilde{U}) & = \left( \frac{1}{\sum_{k=1}^{K} \tilde{\omega}_k} \right) \left( \frac{1}{\sum_{k=1}^{K} \tilde{\omega}_k} \right) \tau \\
& - \frac{1}{\sum_{k=1}^{K} \tilde{\omega}_k} \sum_{k=1}^{K} \tilde{\omega}_k \left\{ \frac{d^2 \ln L(\psi|O, \tilde{U}.k)}{d\psi^2} + \frac{d \ln L(\psi|O, \tilde{U}.k)}{d\psi} \left( \frac{d \ln L(\psi|O, \tilde{U}.k)}{d\psi} \right)^\tau \right\} \\
& \quad | \psi = \hat{\psi}. \end{align*}
\]
TranStat: an efficient tool for outbreak analysis

- Any number of $b$’s and $p$’s.
- Adjust for any number of time-dependent and time independent covariates.
- Flexible common-source-to-person (c2p) and person-to-person (p2p) contact structures.
- Account for preseason immunity and asymptomatic infections
- Assess goodness of fit.
- Written in C, and optimized for computational efficiency.
Input file I: "pop.dat", Population profile

- One line per person.
- Both individual ID and community ID are numbered 0, 1, 2, \cdots.
- Weight can be used for epidemic curve data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>45</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
</tbody>
</table>

† Abbreviations: comm.=community, imm=immunity, inf=infection, stat=status, sym=symptom, cens=censoring, grp=group, wt=weight.
Input file II: "community.dat", Community profile

- For prospective design, starting and stopping days generally cover the duration of epidemic.
- For case-ascertained design, the starting day should be a few days before the symptom onset day of the index case, at least covering the maximum duration of the incubation period.

<table>
<thead>
<tr>
<th>Community ID</th>
<th>Start Day</th>
<th>Stop Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>120</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Input file III: "time_ind_covariate.dat"

**time-independent covariates**

- One line per individual.
- put as many covariates as you want, but remember the order.

<table>
<thead>
<tr>
<th>Individual ID</th>
<th>Age</th>
<th>Gender</th>
<th>Pre-season HI Titer</th>
<th>⋮</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>1</td>
<td>20</td>
<td>⋮</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>0</td>
<td>20</td>
<td>⋮</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>0</td>
<td>40</td>
<td>⋮</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>1</td>
<td>80</td>
<td>⋮</td>
</tr>
<tr>
<td>...</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>
Input file IV: "time_dep_covariate.dat"

time-dependent covariates

- One line per individual per time unit.
- Variables are numbered after time-independent ones.

<table>
<thead>
<tr>
<th>Individual ID</th>
<th>Start Day</th>
<th>Stop Day</th>
<th>Antiviral Treatment</th>
<th>Viral Shedding</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>0</td>
<td>13</td>
<td>45</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>26</td>
<td>45</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>
Input file V: "c2p_contact.dat"
common source to person contact history

- Community-specific or individual-specific format, depending on whether all individuals in the same community share the same c2p contact profile.
- Contact modes are numbered 0, 1, 2, \ldots.
- Offsets reflect variation in the infectivity level of the common source, e.g., \log(\text{daily number of infectious people}).

<table>
<thead>
<tr>
<th>Community/Individual ID</th>
<th>Start Day</th>
<th>Stop Day</th>
<th>Contact Mode</th>
<th>Offset</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>22</td>
<td>0</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>23</td>
<td>30</td>
<td>1</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>30</td>
<td>0</td>
<td>1.0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>30</td>
<td>1</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Input file VI: "p2p_contact.dat"
person-to-person contact history

- Only need to list contacts during the infectious periods of cases

### Community-specific

<table>
<thead>
<tr>
<th>Community ID</th>
<th>Individual ID</th>
<th>Start Day</th>
<th>Stop Day</th>
<th>Contact Mode</th>
<th>Offset</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>5</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>20</td>
<td>26</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>6</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

... 

### Individual-specific

<table>
<thead>
<tr>
<th>Start Day</th>
<th>Stop Day</th>
<th>Person 1 ID</th>
<th>Person 2 ID</th>
<th>Contact Mode</th>
<th>Offset</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

...
Input file VII: "impute.dat", imputation profile

- Four possibilities: preseason immunity, non-infection, symptomatic infection, asymptomatic infection.
- For symptomatic infection, give first and last possible days of illness onset.
- For asymptomatic infection, give first and last possible days of peak infectivity.

<table>
<thead>
<tr>
<th>Person ID</th>
<th>Preseason Immunity</th>
<th>Non-Infected</th>
<th>Sym. Inf.</th>
<th>First Possible Day</th>
<th>Last Possible Day</th>
<th>Asym. Inf.</th>
<th>First Possible Day</th>
<th>Last Possible Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>274</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>12</td>
<td>41</td>
</tr>
<tr>
<td>374</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>375</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>436</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>12</td>
<td>38</td>
</tr>
<tr>
<td>531</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>12</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Input file VIII: "config.file"
Model configuration file

- Natural history of disease, i.e., incubation and infectious periods.
- Parameters to be estimated, e.g.,
  - Numbers of c2p, p2p contact modes.
  - Numbers of time-independent and time-dependent covariates.
  - Covariates affecting susceptibility (c2p/p2p), infectiousness (p2p), or interaction (p2p).
  - Define equivalence classes of parameters. Parameters in the same class are equal.
  - Specify which parameters have fix values.
Choose EM-MCEM algorithm if there is uncertainty in infection/disease outcome.
  - Number of burn-in runs
  - Number of importance samples
  - Threshold for MCEM activation

Statistical adjustment: selection bias or right censoring.

Simulate epidemics to check goodness of fit or not.

Optimization options.

Many more.
Outputs of TranStat

- Estimates, SD, 95% CIs and p-values of parameters.
  - Secondary attack rate: $\text{SAR}_k = 1 - \prod_t \left(1 - p_{kg(t|\tilde{t})}\right)$.
  - Local reproductive number: $R = \sum_k N_k \text{SAR}_k$.
  - Odd ratios, i.e., $\exp(\alpha_S)$, $\exp(\beta_S)$, $\exp(\beta_I)$ and $\exp(\beta_{SI})$.
  - Two output format: detailed vs. simplified.

- p-value for testing existence of person-to-person transmission.

- Goodness of fit: observed and fitted daily numbers of infections, together with simulated bounds.
Case study 1: Contact-tracing data of COVID-19 in Guangzhou, China

- Timeline: Jan 7 - Feb 18, 2020.
- 195 close contact groups, 215 primary cases, 134 secondary cases, 1964 uninfected contacts.
- Among the 349 cases, 19 (5.4%) were asymptomatic.
- 153 (73%) primary cases and 66 (46%) secondary cases were imported.
- Main goals:
  - Estimate household SAR and non-household SAR;
  - Evaluate age and gender effects on susceptibility and infectivity;
  - Assess infectivity during the incubation (preclinical) period.
Table 1. Demographic compositions of the study population stratified by case type (primary, secondary and non-case) and contact type (household [HH] and non-household). Contact type is determined by relationship with the primary cases of each close contact group. Percentages are enclosed in parentheses. The data-based secondary attack rate is calculated as the number of secondary cases divided by the sum of secondary cases and non-cases.

<table>
<thead>
<tr>
<th>Definition of household</th>
<th>Factor</th>
<th>Category</th>
<th>Primary cases</th>
<th>Secondary cases</th>
<th>Non-cases</th>
<th>Overall</th>
<th>Data-based SAR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>HH</td>
<td>Non-HH</td>
<td>HH</td>
<td>Non-HH</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Close relatives</td>
<td>Age</td>
<td>&lt;20</td>
<td>10 (5)</td>
<td>9 (9)</td>
<td>1 (3)</td>
<td>163 (24)</td>
<td>70 (5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20-59</td>
<td>145 (67)</td>
<td>67 (65)</td>
<td>22 (71)</td>
<td>385 (57)</td>
<td>961 (75)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>≥60</td>
<td>60 (28)</td>
<td>27 (26)</td>
<td>8 (26)</td>
<td>120 (18)</td>
<td>247 (19)</td>
</tr>
<tr>
<td>Sex</td>
<td>Female</td>
<td>107 (50)</td>
<td>57 (55)</td>
<td>17 (55)</td>
<td>341 (50)</td>
<td>627 (49)</td>
<td>1149 (50)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>108 (50)</td>
<td>46 (45)</td>
<td>14 (45)</td>
<td>335 (49)</td>
<td>651 (51)</td>
<td>1154 (50)</td>
</tr>
<tr>
<td>Month †</td>
<td>Jan.</td>
<td>193 (90)</td>
<td>98 (95)</td>
<td>29 (94)</td>
<td>545 (80)</td>
<td>681 (53)</td>
<td>1546 (67)</td>
</tr>
<tr>
<td></td>
<td>Feb.</td>
<td>22 (10)</td>
<td>5 (5)</td>
<td>2 (6)</td>
<td>136 (20)</td>
<td>602 (47)</td>
<td>767 (33)</td>
</tr>
<tr>
<td>HH size</td>
<td>≤6</td>
<td>160 (74)</td>
<td>69 (67)</td>
<td>23 (74)</td>
<td>302 (44)</td>
<td>874 (68)</td>
<td>1428 (62)</td>
</tr>
<tr>
<td></td>
<td>&gt;6</td>
<td>55 (26)</td>
<td>34 (33)</td>
<td>8 (26)</td>
<td>379 (56)</td>
<td>409 (32)</td>
<td>885 (38)</td>
</tr>
<tr>
<td>Origin</td>
<td>Imported</td>
<td>158 (73)</td>
<td>59 (57)</td>
<td>3 (10)</td>
<td>227 (51)</td>
<td>684 (57)</td>
<td>1546 (67)</td>
</tr>
<tr>
<td></td>
<td>Local</td>
<td>57 (27)</td>
<td>44 (43)</td>
<td>28 (90)</td>
<td>117 (26)</td>
<td>206 (58)</td>
<td>2313 (100)</td>
</tr>
</tbody>
</table>

| Residential address    | Age    | <20      | 10 (5)        | 9 (9)          | 1 (3)     | 117 (26) | 116 (8)          | 253 (11) | 6.4 (2.8, 12.2) | 1.7 (0.21, 5.0) |
|                        |        | 20-59    | 145 (67)      | 59 (63)        | 30 (73)   | 260 (58) | 1086 (72)        | 1580 (68) | 18.5 (14.4, 23.2) | 2.7 (1.8, 3.8) |
|                        |        | ≥60      | 60 (28)       | 26 (28)        | 9 (22)    | 67 (15)  | 300 (20)         | 462 (20)  | 28.0 (19.1, 38.2) | 2.9 (1.3, 5.5) |
| Sex                    | Female | 107 (50) | 53 (57)       | 21 (51)       | 227 (51)  | 741 (49) | 1149 (50)        | 18.9 (14.5, 24.0) | 2.8 (1.7, 4.2) |
|                        | Male   | 108 (50) | 40 (43)       | 20 (49)       | 218 (49)  | 768 (51) | 1154 (50)        | 15.5 (11.3, 20.5) | 2.5 (1.6, 3.9) |
| Month †                | Jan.   | 193 (90) | 88 (95)       | 39 (95)       | 362 (81)  | 864 (57) | 1546 (67)        | 19.6 (16.0, 23.5) | 4.3 (3.1, 5.9) |
|                        | Feb.   | 22 (10)  | 5 (5)         | 2 (5)         | 87 (19)   | 651 (43) | 767 (33)         | 5.4 (1.8, 12.2)  | 0.31 (0.04, 1.1) |
| HH size                | ≤6     | 188 (87) | 79 (85)       | 32 (78)       | 309 (69)  | 1191 (79) | 1799 (78)        | 20.4 (16.5, 24.7) | 2.6 (1.8, 3.7) |
|                        | >6     | 27 (13)  | 14 (15)       | 9 (22)        | 140 (31)  | 324 (21) | 514 (22)         | 9.1 (5.1, 14.8)  | 2.7 (1.2, 5.1) |
| Origin                 | Imported | 158 (73) | 56 (60)       | 6 (15)        | 227 (51)  | 684 (57) | 1546 (67)        | 19.6 (16.0, 23.5) | 4.3 (3.1, 5.9) |
|                        | Local  | 57 (27)  | 37 (40)       | 35 (85)       | 117 (26)  | 206 (58) | 2313 (100)       | 13.2 (10.9, 15.7) | 2.4 (1.6, 3.3) |

‡ Secondary cases and non-cases in each CCG were allocated to January or February of 2020 according to the proportion of the primary cases’ infectious periods falling in January vs. that in February.
Table 2. Model-based estimates (and 95% confidence intervals) of secondary attack rates among household and non-household contacts, and model-based estimates of the local reproductive number (local $R$) with and without quarantine. Estimates are reported using two different definitions of household contact (close relatives or individuals sharing the same residential address) and for selected settings of the natural history of disease. This model is not adjusted for age group, epidemic phase or household size.

<table>
<thead>
<tr>
<th>Definition of household</th>
<th>Parameter</th>
<th>Setting</th>
<th>Mean incubation period = 5 days</th>
<th>Mean incubation period = 7 days</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Max infectious period = 13 days</td>
<td>Max infectious period = 22 days</td>
</tr>
<tr>
<td>Close relatives</td>
<td>SAR (%)</td>
<td>Household</td>
<td>12·4 (9·8, 15·4)</td>
<td>15·5 (11·7, 20·2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-household</td>
<td>7·9 (5·3, 11·8)</td>
<td>10·4 (6·7, 15·8)</td>
</tr>
<tr>
<td></td>
<td>Local $R$</td>
<td>With quarantine</td>
<td>0·50 (0·41, 0·62)</td>
<td>0·51 (0·39, 0·66)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No quarantine</td>
<td>0·60 (0·49, 0·74)</td>
<td>0·76 (0·59, 1·00)</td>
</tr>
<tr>
<td>Residential address</td>
<td>SAR (%)</td>
<td>Household</td>
<td>17·1 (13·3, 21·8)</td>
<td>21·2 (15·8, 27·8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-household</td>
<td>7·3 (5·4, 9·9)</td>
<td>9·3 (6·5, 13·1)</td>
</tr>
<tr>
<td></td>
<td>Local $R$</td>
<td>With quarantine</td>
<td>0·50 (0·40, 0·61)</td>
<td>0·50 (0·38, 0·65)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No quarantine</td>
<td>0·59 (0·48, 0·72)</td>
<td>0·74 (0·57, 0·96)</td>
</tr>
</tbody>
</table>
Table 4. Model-based odds ratios (and 95% confidence intervals) for the effects of age group and epidemic phase (Feb. vs. Jan.) on susceptibility and relative infectivity during the illness period compared to the incubation period. Estimates are reported using two different definitions of household contact (close relatives or individuals sharing the same residential address) and for selected settings of the natural history of disease. This model is adjusted for age group, epidemic phase, and household size.

<table>
<thead>
<tr>
<th>Definition of household contact</th>
<th>Parameter</th>
<th>Odd ratio</th>
<th>Mean incubation period = 5 days</th>
<th>Mean incubation period = 7 days</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Max infectious period = 13 days</td>
<td>Max infectious period = 22 days</td>
</tr>
<tr>
<td>Close relatives</td>
<td>Susceptibility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age group &lt;20 vs. ≥60</td>
<td>0.23 (0.11, 0.46)</td>
<td>0.22 (0.11, 0.46)</td>
<td>0.22 (0.11, 0.45)</td>
</tr>
<tr>
<td></td>
<td>Age group 20-59 vs. ≥60</td>
<td>0.64 (0.43, 0.97)</td>
<td>0.64 (0.42, 0.96)</td>
<td>0.63 (0.42, 0.95)</td>
</tr>
<tr>
<td></td>
<td>Feb. vs. Jan.</td>
<td>0.42 (0.17, 1.07)</td>
<td>0.46 (0.19, 1.10)</td>
<td>0.36 (0.12, 1.05)</td>
</tr>
<tr>
<td></td>
<td>Infectivity</td>
<td>0.60 (0.27, 1.36)</td>
<td>0.42 (0.19, 0.91)</td>
<td>0.29 (0.10, 0.88)</td>
</tr>
<tr>
<td>Residential address</td>
<td>Susceptibility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age group &lt;20 vs. ≥60</td>
<td>0.22 (0.11, 0.46)</td>
<td>0.22 (0.11, 0.45)</td>
<td>0.22 (0.11, 0.44)</td>
</tr>
<tr>
<td></td>
<td>Age group 20-59 vs. ≥60</td>
<td>0.67 (0.45, 1.00)</td>
<td>0.67 (0.45, 1.00)</td>
<td>0.66 (0.44, 0.99)</td>
</tr>
<tr>
<td></td>
<td>Feb. vs. Jan.</td>
<td>0.57 (0.23, 1.39)</td>
<td>0.62 (0.27, 1.44)</td>
<td>0.50 (0.18, 1.36)</td>
</tr>
<tr>
<td></td>
<td>Infectivity</td>
<td>0.54 (0.23, 1.26)</td>
<td>0.38 (0.17, 0.84)</td>
<td>0.24 (0.07, 0.79)</td>
</tr>
<tr>
<td></td>
<td>Illness vs. Incubation</td>
<td>0.54 (0.23, 1.26)</td>
<td>0.38 (0.17, 0.84)</td>
<td>0.24 (0.07, 0.79)</td>
</tr>
</tbody>
</table>
Case study 2: Estimate $R_t$ of COVID-19 in Wuhan, China

- 8866 probable cases in 30 provinces of China, 4021 (45%) lab-confirmed.
- 3731 probable cases (1664 lab-confirmed) in Wuhan.
- Mean age was 48 years (SD=16 years).
- Main goals: To estimate time-varying effective reproduction number $R_t$. 

