Outline

- Introduction
- Discrete-time Likelihood model for complete data
- Selection bias, left truncation and right censoring
- Data augmentation
- Summary
Introduction to the problem
Basic data structure

- A community of households.
- Two infection sources.
  - Community at large: zoonotic source, visitors, neighborhoods.
  - Infected household members.
- Infection leads to symptoms (natural history)
- Symptom diary, e.g., headache, sore throat, temperature.
- Lab-confirmation:
  - Viral culture at beginning and end of study or triggered by symptom onset.
  - Antibody titers: 4-fold increase from baseline to the end of study.
- Intervention implemented, e.g., vaccine vs. placebo.
Hierarchy of Information Levels

- Consecutive occurrence of infections is a counting process, observed at different information levels (Rhodes, Halloran and Longini, JRSS B, 1996)
  - How many infections have occurred in $(0, T]$. Final value models (Longini et al, 1982; Addy et al, 1991).
  - Times at which infection or symptom onset occurs. Survival model (Longini and Halloran, 1996).
  - Who contacts whom and/or who infects whom. Discrete-time likelihood models (Rampey et al, 1992; Yang et al, 2006).
  * Sometimes difficult to obtain.
  * Clustering pattern is the bottom line.
• Define contact before define transmission probability and intervention efficacy.
  – The nature of the behavior, e.g., conversation, needle injection, being in the same household.
  – The covariates associated with such behavior.
    * Characteristics of the infectious, e.g., an untreated infectious person.
    * Characteristics of the susceptible.
    * Characteristics of the behavior.
Capacity of Transmission

- Instantaneous risk (hazard) $\lambda(t) = \Pr(t < t^* < t + \delta \mid t^* > t)$.
  - The probability of transmission in $(t, t + \delta)$ given no transmission before $t$.
  - A continuous-time concept.

- Transmission probability: $\Pr(t^* < b \mid t^* > a) = 1 - \exp\left\{\int_a^b \lambda(t) dt\right\}$.
  - The probability of transmission during a time span given no transmission before.
  - Related to cumulative risk, but is the basis for discrete-time methods, e.g., trans. prob. per day, $p = 1 - \exp\{-\lambda\}$. 
Intervention Efficacy

• Efficacy = 1 - Relative Risk = 1 - (risk in intervention / risk in control).

• Risk could be instantaneous risk or transmission probability.
  – Efficacy based on transmission probability is generally smaller than that based on instantaneous risk.

• Various types of efficacies in the context of infectious disease: $VE_S$, $VE_I$, $VE_T$, $VE_P$ (Halloran, Struchiner and Longini, AJE, 1997).

• Distinguish between intervention efficacy and effectiveness.
  – Efficacy is an attribute of the intervention itself.
  – Effectiveness depends on how you distribute the intervention.
Final Size Model

- Longini and Koopman (Biometrics, 1982)
  - $B$: Probability of escaping infection from external source during epidemic.
  - $Q$: Probability of escaping infection from an infectious household member during epidemic.
  - $m_{jk}$: probability that $j$ out of $k$ household members are infected.
    * Household with a single person: $m_{01} = B$ and $m_{11} = 1 - B$
    * Household with two members:
      - $m_{02} = B^2$
      - $m_{12} = 2(1 - B)BQ$
      - $m_{22} = 1 - m_{02} - m_{12} = 2(1 - B)(1 - Q)B + (1 - B)^2$
    * In general, $m_{jk} = \binom{k}{j} m_{jj} B^{k-j} Q^{j(k-j)}$ and $m_{jj} = 1 - \sum_{l<j} m_{lj}$. 
- Maximum likelihood estimation
  * Likelihood: \( L(B, Q) = \prod_{k,j} m_{jk}^{a_{jk}} \), where \( a_{jk} \) is the frequency of households corresponding to \( m_{jk} \).
  * Score function:
    \[
    \frac{\partial \ln L}{\partial B} = \sum_{k,j} a_{jk} \left\{ \frac{1}{m_{jj}} \frac{\partial m_{jj}}{\partial B} + \frac{k - j}{B} \right\}.
    \]
  * Fisher’s information:
    \[
    -E\left( \frac{\partial^2 \ln L}{\partial B^2} \right) = \sum_{k,j} n_k m_{jk} \left\{ \frac{1}{m_{jj}^2} \left( \frac{\partial m_{jj}}{\partial B} \right)^2 - \frac{1}{m_{jj}^2} \frac{\partial^2 m_{jj}}{\partial B^2} + \frac{k - j}{B^2} \right\}.
    \]
* Rough estimates for starting point

\[
\frac{a_{0k}}{n_k} = \hat{m}_{0k} = B_k^k \Rightarrow \hat{B}_k = \left( \frac{a_{0k}}{n_k} \right)^{1/k} \Rightarrow \hat{B} = \frac{1}{n} \sum_k n_k \hat{B}_k
\]

\[
\frac{a_{1k}}{n_k} = \hat{m}_{1k} = k(1 - \hat{B})\hat{B}^{k-1}Q^{k-1}?
\]

\[
\hat{Q} \approx \left( \frac{1 - \hat{\theta}}{\hat{B}} \right)^{1/\hat{\phi}}
\]

where \( \hat{\phi} = \frac{\sum_{k,j} ja_{jk}}{n} \) and \( \hat{\theta} = \frac{\sum_{k,j} (ja_{jk}/k)}{n} \).

- Inter-group mixing (Addy, Longini and Haber, Biometrics, 1991).
Frailty Hazard Model

  - $\alpha_v$: proportion of full immunity in group $v$ ($1 = \text{vaccine}, 0 = \text{control}$).
    * If $\alpha_1 > \alpha_0$, “all-or-none” effect.
  - $\theta$: reduction rate in susceptibility for the $1 - \alpha_1$ of vaccinated population, “leaky” effect.
  - $VE = 1 - \frac{(1-\alpha_1)\theta\pi}{(1-\alpha_0)\pi} = 1 - \frac{(1-\alpha_1)}{(1-\alpha_0)}\theta$.
  - Frailty (random) hazard
    * $\Pr(Z_v = 0) = \alpha_v$
    * $Z_v | Z_v > 0 \sim f_v(\text{mean} = 1, \text{variance} = \delta_v)$
    * Hazard function: $\lambda_v(t) = Z_v \theta^v c\pi p(t)$. 
* Survival function:

\[ S_v(t) = E_{Z_v} \left[ \exp\{-Z_v \int_0^t \lambda_v(\tau) d\tau\} \right] \]

\[ = \alpha_v + (1 - \alpha_v) \left\{ \frac{1}{1 + \delta_v \int_0^t \lambda_v(\tau) d\tau} \right\}^{1/\delta_v}, \quad \text{if } Z_v \sim \Gamma \]

- For grouped survival data with \( k \) intervals, \( p(t) = \prod_{i=1}^k p_i^{I(t_{i-1} \leq t < t_i)} \).
- \( r_{iv} \): number of subjects at risk at the beginning of interval \([t_{i-1}, t_i)\).
- \( m_{iv} \): number of subjects infected in interval \([t_{i-1}, t_i)\).
- Likelihood function

\[ L = \prod_{i=1}^k \prod_{v=0}^1 \left\{ \frac{S_v(t_i)}{S_v(t_{i-1})} \right\}^{r_{iv} - m_{iv}} \left\{ 1 - \frac{S_v(t_i)}{S_v(t_{i-1})} \right\}^{m_{iv}} \]
Transmission Patterns and Parameters of Interest

$p$: within-household pairwise daily transmission probability without treatment.

$b$: daily probability of infection by the community without treatment (CPI).

$\text{AVE}_{s} = 1 - \theta$ : Efficacy of the antiviral agent in reducing susceptibility.

$\text{AVE}_{t} = 1 - \phi$ : Efficacy of the antiviral agent in reducing infectiousness.
Natural Disease History of Influenza

Latent period (Incubation period)

\[ g(t | t) \]

Time of Infection

Infectious period

\[ 1 - f(t | \tilde{t}) \]

Onset time of symptoms and infectiousness

\[ g(\tilde{t} | t) : \text{The probability of symptom onset on day } \tilde{t} \text{ given infection on day } t. \]

\[ f(t | \tilde{t}) : \text{Probability that the host is infective on day } t \text{ given symptom onset on day } \tilde{t}. \]
Likelihood for complete data

- Likelihood for a person-day

  Probability of pairwise transmission per daily contact:

  \[ p_{ji}(t) = \begin{cases} 
  \theta r_i(t) \phi r_j(t) p f(t|\tilde{t}_j), & j \in H_i \\
  \theta r_i(t) b, & j = c. 
\end{cases} \]

  Define \( D_i = H_i \cup c \). Probability of escaping infection on day \( t \):

  \[ e_i(t) = \prod_{j \in D_i} (1 - p_{ji}(t)) \]

  Probability of escaping infection up to day \( t \):

  \[ Q_i(t) = \prod_{\tau=1}^{t} e_i(\tau) \]
• Likelihood contributed by a single individual

If subject $i$ is known to be infected on day $t$, the probability is

$$U_i(t) = [1 - e_i(t)] Q_i(t - 1),$$

Generally only symptom onset is observable

$$L_i = \begin{cases} 
Q_i(T), & \text{if individual } i \text{ is not infected} \\
\sum_{t=t_i}^{\infty} g(\tilde{t}_i | t) U_i(t), & \text{otherwise}
\end{cases}$$

where $t_i = \tilde{t}_i - l_{max}$, $\overline{t}_i = \tilde{t}_i - l_{min}$ and $T$ is the last observation day for the epidemic.
• Likelihood for the whole population
  – Is \( L = \prod_{i=1}^{N} L_i \) valid? A formal question is: *Given a stochastic infectious period, are exposed people independent of each other?*
  – A simple example: three people, \( \alpha, \beta \) and \( \gamma \) in a household, infected on day 1, 3 and 4 respectively. Assume person \( \alpha \) is infectious only for 1 day, and occurs on day 3 with prob \( s \) and on day 4 with prob \( 1-s \).
    * Likelihood for person \( \beta \): 
      \[
      L^m_\beta = (1 - b)^2 \{ s(1 - (1 - b)(1 - p)) + (1 - s)b \}
      \]
    * Likelihood for person \( \gamma \):
      \[
      L^m_\gamma = (1 - b)^2 \{ s(1 - b)(1 - p)b + (1 - s)(1 - b)(1 - (1 - b)(1 - p)) \}
      \]
* Likelihood for $\beta$ and $\gamma$:

\[
L_{\beta,\gamma}^m = (1 - b)^4 \{ s(1 - (1 - b)(1 - p))(1 - b)(1 - p)b \\
+ (1 - s)b(1 - b)(1 - (1 - b)(1 - p)) \}
\]

* $L_{\beta,\gamma}^m \neq L_{\beta}^m L_{\gamma}^m!$

* However, inference based on the product of individual marginal likelihoods is asymptotically correct.
More on infectious period and dependency

- person $\alpha$ is infected on day 1 and person $\beta$ on day 2. There is a third person $\gamma$ infective on day 1 with probability $s$ and on day 2 with $1-s$.
- Define $z = 1$ if person $\gamma$ is infective on day 1 and $z = 0$ if on day 2.
- Joint likelihood for $\alpha$ and $\beta$

$$L(b, p|z) = [1 - (1 - b)(1 - pz)]$$
$$\times (1 - b)(1 - pz) \{1 - (1 - b)[1 - p(1 - z)]\}, \quad (1)$$

or equivalently,

$$L(b, p|z) = b^{1-z}[1 - (1 - b)(1 - p)]^z$$
$$\times (1 - b)(1 - p)^z b^z [1 - (1 - b)(1 - p)]^{1-z}. \quad (2)$$
• Inference should be based on $E_z L(b, p|z)$, but could be numerically difficult if there are many infectives and infectious period is long.

• Is $\log E_z L(b, p|z) \approx E_z \log L(b, p|z)$? Probably not, but at MLE $(\hat{b}, \hat{p})$,

$$\frac{\partial}{\partial p} \log E_z L(\hat{b}, \hat{p}|z) = 0 \iff \frac{1}{E_z L(\hat{b}, \hat{p}|z)} E_z \frac{\partial}{\partial p} L(\hat{b}, \hat{p}|z) = 0$$

$$\iff E_z \frac{\partial \log L(\hat{b}, \hat{p}|z)}{\partial p} L(\hat{b}, \hat{p}|z) = 0.$$ 

Note that $0 \leq L(\hat{b}, \hat{p}|z) \leq 1$ and thus $VAR_z(L[\hat{b}, \hat{p}|z]) \leq 0.25$, but $VAR_z[\frac{\partial \log L(\hat{b}, \hat{p}|z)}{\partial p}] \rightarrow \infty$. Hence $L(\hat{b}, \hat{p}|z)$ is asymptotically constant relative to $\frac{\partial \log L(\hat{b}, \hat{p}|z)}{\partial p}$, and from (2)

$$\frac{\partial \log L(\hat{b}, \hat{p}|E_z(z))}{\partial p} = E_z \frac{\partial \log L(\hat{b}, \hat{p}|z)}{\partial p} \approx 0$$
• If $\alpha$ infected on day 1 and $\beta$ infected on day 2,
\[
\frac{\partial \log L(\hat{b}, \hat{p} | z)}{\partial p} = \frac{1-b}{1-(1-b)(1-p)} - \frac{z}{1-p}.
\]

• If $\alpha$ not infected and $\beta$ infected on day 2,
\[
\frac{\partial \log L(\hat{b}, \hat{p} | z)}{\partial p} = -\frac{z}{1-p} + \frac{(1-z)(1-b)}{1-(1-b)(1-p)}.
\]

• When such approximation is in doubt
  – E-M algorithm, treating $z$ as missing data. Iteratively replace $z$ with $E_{z|b,p,y}(z)$ in (2), and maximize $\log L(b, p | z)$. What we did in our old papers is replacing $z$ in (1) with $s$, which implies using only empirical information to calculate $E_{z|b,p,y}(z)$. A new paper!
  – Bayesian method.
Likelihood for truncated data: case-ascertained design

- Problems
  - *Selection bias:* Only households with infected members are followed.
  - *Left truncation:* Individuals are observed after the ascertainment of index cases.

- Solution: conditioning on something.

- Why conditioning? A simple example:
  - Completely observed $k$ geometric sequences, $\text{Pr}(T_i) = q^{T_i-1}p$, $i = 1, \ldots, k$, $p = 1 - q$.
  - Likelihood is $q^{\sum_i(T_i-1)}(1-q)^k$, and the MLE is $\hat{q} = \frac{\{\frac{1}{k} \sum_i T_i\}^{-1}}{\frac{1}{k} \sum_i T_i}$.
    
    Quick check for consistency, $E[T_i] = \frac{1}{1-q}$.
  - We start observation at $T_{(1)} = \min_i T_i$. How do we estimate $q$?
- Answer: conditioning on observation up to $T_{(1)}$.
  * Let $M = \{ i : T_i = T_{(1)} \}$ and $m$ be the size of $M$.
  * Conditional likelihood is $q \sum_{i \notin M} (T_i - T_{(1)} - 1) (1 - q)^{k-m}$, and conditional MLE is $\hat{q}^c = \frac{\sum_{i \notin M} (T_i - T_{(1)} - 1)}{k-m + \sum_{i \notin M} (T_i - T_{(1)} - 1)}$.
* quick check for consistency:
  \[
  \cdot \hat{q}^c = \frac{\{ \sum_{i} (T_i - 1) \} - kT_{(1)} + m}{\{ \sum_{i} (T_i - 1) \} - kT_{(1)} + k} = 1 - \frac{1 - \frac{m}{k}}{\{ \frac{1}{k} \sum_{i} (T_i) - T_{(1)} \}} \\
  \cdot E[T_i] = \frac{1}{1-q}.
  \]
  \[
  \cdot E[T_{(1)}] = \sum_{j=0}^{\infty} Pr\{T_1 > j, \ldots, T_k > j\} = \frac{1}{1-q^k}.
  \]
  \[
  \cdot E[m] = \sum_{m=1}^{k} \left\{ m \binom{k}{m} \sum_{j=1}^{\infty} q^{(j-1)m} (1 - q)^m q^j (k-m) \right\} = \frac{k(1-q)}{1-q^k}
  \]
  \[
  \cdot \hat{q}^c \xrightarrow{p} q.
  \]
• Possible solution: conditioning on the disease history (infection and symptom) up to the symptom onset day of the index case \( \tilde{t}_{d_i} \).

1. \( \tilde{t}_i \leq \tilde{t}_{d_i} \): \( L_i^m = L_i \), e.g., index case.
2. \( \tilde{t}_{d_i} < t_i \): \( L_i^m = Q_i(\tilde{t}_{d_i}) \)
3. \( t_i \leq \tilde{t}_i \leq \tilde{t}_{d_i} < \tilde{t}_i \):

\[
L_i^m = Q_i(t_i - 1) \sum_{t=\tilde{t}_i}^{\tilde{t}_i} \left\{ (1 - e_i(t)) \prod_{\tau=\tilde{t}_i}^{t-1} e_i(t) \right\} \times \Pr(\tilde{t}_i > \tilde{t}_{d_i} | t)
\]

4. \( \tilde{t}_i \leq \tilde{t}_{d_i} < \tilde{t}_i < \tilde{t}_i \):

\[
L_i^m = Q_i(t_i - 1) \sum_{t=\tilde{t}_i}^{\tilde{t}_{d_i}} \left\{ (1 - e_i(t)) \prod_{\tau=\tilde{t}_i}^{t-1} e_i(t) \right\} \times \Pr(\tilde{t}_i > \tilde{t}_{d_i} | t)
\]

\[
+ Q_i(t_i - 1) \prod_{\tau=\tilde{t}_i}^{\tilde{t}_{d_i}} e_i(t)
\]
- Use the conditional likelihood $L^c_i = L_i / L^m_i$ for inference.
- More accurate conditioning: $L^c_H = L_H / L^m_H$. If infectious period starts before symptom onset, calculation of $L^m_H$ could be complicated. Bayesian approach may be the solution.
Likelihood for right-censored data: real-time analysis

- No symptoms observed could mean either escape from infection or incubation period.
- Calculate the marginal probability of observing no symptom onset up to day $T$:

$$L^m_i = Q_i(T - l_{min}) + \sum_{t=T-l_{max}+1}^{T-l_{min}} \left\{ (1 - e_i(t))Q_i(t - 1) \right\} \times \Pr(\tilde{t}_i > T|t)$$
Assessing goodness of fit


- The probability of symptom onset on day $t$ for subject $i$ is

$$
\pi_i(t) = \sum_{\tau=t-l_{\text{min}}}^{t-l_{\text{min}}} \left\{ (1 - e_i(\tau)) \prod_{s=t-l_{\text{max}}}^{\tau-1} e_i(s) \right\} g(t|\tau).
$$

- Choose $0 = c_0 < c_1 < \ldots < c_m = 1$, then $\hat{n}_k = \sum_{c_{k-1} < \pi_i < c_k} \pi_i$ is the fitted count in level $k$. Let $N_k$ be the total person-days and $\tilde{n}_k$ be the observed count in level $k$, then

$$
\sum_{k=1}^{m} \frac{N_k(\tilde{n}_k - \hat{n}_k)^2}{\hat{n}_k(N_k - \hat{n}_k)} \sim \chi^2_{m-2}.
$$

- If $\hat{n}_k \ll N_k$ for all $k$, it is simplified to $\sum_{k=1}^{m} \frac{(\tilde{n}_k - \hat{n}_k)^2}{\hat{n}_k}$. 
Simulation study

Population: a community composed of households of size two or larger with 1000 people is generated based on the age distribution and household sizes from the US Census 2000.

Table 1: Empirical distributions of the latent period and the infectious period (Elveback et al., 1976)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>4</td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>5</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 2: Comparison of MLEs by randomization schemes and household follow-up schemes

<table>
<thead>
<tr>
<th>Parameter‡</th>
<th>Estimate</th>
<th>MonteCarlo standarderrors</th>
<th>95%CI coverage (%)§§</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_$</td>
<td>$H_$</td>
<td>$T_$</td>
</tr>
<tr>
<td>$\theta$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prospective</td>
<td>0.70</td>
<td>0.71</td>
<td>0.083</td>
</tr>
<tr>
<td>Case-ascertained</td>
<td>0.70</td>
<td>0.71</td>
<td>0.083</td>
</tr>
<tr>
<td>$\phi$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prospective</td>
<td>0.20</td>
<td>0.24</td>
<td>0.045</td>
</tr>
<tr>
<td>Case-ascertained</td>
<td>0.20</td>
<td>0.24</td>
<td>0.044</td>
</tr>
</tbody>
</table>

‡ True efficacy-related parameters are set to $\theta = 0.70$ and $\phi = 0.20$.
§ I, individual-level randomization; H, household-level randomization.
§§ The 95% CI is obtained as $\exp\{\log(\hat{\lambda}) \pm 1.96 \times \text{se}\{\log(\hat{\lambda})\}\}; \lambda = \theta, \phi.$
Table 3: Two randomized multi-center trials of Oseltamivir, an influenza antiviral agent.

<table>
<thead>
<tr>
<th></th>
<th>Trial I</th>
<th>Trial II</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Welliver et al. 2001)</td>
<td>(Hayden et al. 2004)</td>
<td></td>
</tr>
<tr>
<td>Households</td>
<td>372</td>
<td>277</td>
</tr>
<tr>
<td>Population</td>
<td>1329</td>
<td>1110</td>
</tr>
<tr>
<td>Treatment for illness</td>
<td>None</td>
<td>Oseltamivir</td>
</tr>
<tr>
<td>Duration of medication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illness treatment</td>
<td>N/A</td>
<td>5 days</td>
</tr>
<tr>
<td>Prophylaxis</td>
<td>7 days</td>
<td>10 days</td>
</tr>
<tr>
<td>Follow up (symptom diary)</td>
<td>14 days</td>
<td>30 days</td>
</tr>
<tr>
<td>Infected/Exposed(index)</td>
<td>165/372</td>
<td>179/298</td>
</tr>
<tr>
<td>Infected/Exposed(susceptible)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control†</td>
<td>38/464</td>
<td>45/392</td>
</tr>
<tr>
<td>Oseltamivir</td>
<td>4/493</td>
<td>14/420</td>
</tr>
</tbody>
</table>
Table 4: Maximum likelihood estimates by age (1-17 vs 18+) for pooled oseltamivir trials conducted in 1998-1999 and 2000-2001, North America and Europe.

<table>
<thead>
<tr>
<th>With Assumption</th>
<th>Parameter</th>
<th>MLE</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>$b_c^\dagger$</td>
<td>0.0023</td>
<td>(0.0015, 0.0035)</td>
</tr>
<tr>
<td></td>
<td>$b_a$</td>
<td>0.00055</td>
<td>(0.0003, 0.001)</td>
</tr>
<tr>
<td></td>
<td>$p_{cc}$</td>
<td>0.038</td>
<td>(0.023, 0.063)</td>
</tr>
<tr>
<td></td>
<td>$p_{ca}$</td>
<td>0.012</td>
<td>(0.007, 0.021)</td>
</tr>
<tr>
<td></td>
<td>$p_{ac}$</td>
<td>0.018</td>
<td>(0.008, 0.040)</td>
</tr>
<tr>
<td></td>
<td>$p_{aa}$</td>
<td>0.022</td>
<td>(0.014, 0.034)</td>
</tr>
<tr>
<td></td>
<td>$\text{AVE}_S$</td>
<td>0.85</td>
<td>(0.52, 0.95)</td>
</tr>
<tr>
<td></td>
<td>$\text{AVE}_I$</td>
<td>0.66</td>
<td>(-0.10, 0.89)</td>
</tr>
<tr>
<td></td>
<td>$\text{AVE}_T$</td>
<td>0.95</td>
<td>(0.77, 0.99)</td>
</tr>
<tr>
<td>No</td>
<td>$\text{AVE}_S$</td>
<td>0.93</td>
<td>(0.50, 0.99)</td>
</tr>
<tr>
<td></td>
<td>$\text{AVE}_I$</td>
<td>0.78</td>
<td>(-0.27, 0.96)</td>
</tr>
<tr>
<td></td>
<td>$\text{AVE}_T$</td>
<td>0.87</td>
<td>(0.41, 0.97)</td>
</tr>
<tr>
<td></td>
<td>$\text{SAR}_{cc}^\ddagger$</td>
<td>0.15</td>
<td>(0.074, 0.21)</td>
</tr>
<tr>
<td></td>
<td>$\text{SAR}_{ca}$</td>
<td>0.049</td>
<td>(0.021, 0.075)</td>
</tr>
<tr>
<td></td>
<td>$\text{SAR}_{ac}$</td>
<td>0.071</td>
<td>(0.014, 0.13)</td>
</tr>
<tr>
<td></td>
<td>$\text{SAR}_{aa}$</td>
<td>0.086</td>
<td>(0.047, 0.12)</td>
</tr>
</tbody>
</table>

$\dagger$, $\ddagger$ Subscription $c$ denotes child (1-17), $a$ denotes adult (18+), and $ca$ denotes child-to-adult transmission.

$\ddagger$ $\text{SAR}_{vu}$ is based on the average 4.1 days of infectious period, i.e., $\text{SAR}_{vu} = 1 - (1 - p_{vu})^{4.1}$. 

<table>
<thead>
<tr>
<th>Risk Level</th>
<th>Total Person-days</th>
<th>Observed # of illness onsets</th>
<th>Predicted # of illness onsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2084</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1321</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>15878</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>1434</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>8165</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>933</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>935</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>1241</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>1084</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>894</td>
<td>25</td>
<td>27</td>
</tr>
</tbody>
</table>

† With assumption of $\psi = \theta \phi$. 
Maximum Likelihood Estimates Based on Augmented Data

- **Unobserved data**
  - Pairwise transmission outcomes $Y_{ji}(t)$ (1: transmission, 0: escape) were not observed when $j$ is infective and $t_i \leq t \leq \bar{t}_i$.
  - $Y_{ji}(t)$ is observable only if $Y_{ji}({\tau}) = 0$ for all $\tau < t$.
  - $Y_{ji}(t)$ is independent of $Y_{ki}(t)$ for the same day $t$.
  - More convenient to work with $Z_{ji}(t) = Y_{ji}(t) \prod_{k \in D_i, \tau < t} (1 - Y_{ki}(\tau))$ and $\tilde{Z}_{ji}(t) = (1 - Y_{ji}(t)) \prod_{k \in D_i, \tau < t} (1 - Y_{ki}(\tau))$. 
Potential transmissions not observable

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Outcome</th>
<th>Expected Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 2</td>
<td>0</td>
<td>( \Pr[ Z_{21}(t_i - 1) = 1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>( \Pr[ Z_{21}(t_i - 1) = 1</td>
</tr>
<tr>
<td>1 c</td>
<td>0</td>
<td>( \Pr[ Z_{11}(t_i - 1) = 1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>( \Pr[ Z_{11}(t_i - 1) = 1</td>
</tr>
</tbody>
</table>
• The likelihood of the augmented data

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

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

where \( \max_{j \in D_i} Z_{ji}(t) \) indicates if \( Z_{ji}(t) = 1 \) for any \( j \) on day \( t \). The log-likelihood is

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

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

$$\Pr[I_{ji}(t)] = \hat{Q}_i(t-1)\hat{p}_{ji}(t)$$
$$\Pr[I_i(t)] = \hat{Q}_i(t-1)\{1 - \hat{e}_i(t)\},$$
$$\Pr[S_i(\tilde{t}_i)] = \sum_{\tau = \tilde{t}_i}^{\bar{t}_i} g(\tilde{t}_i | \tau) \times \Pr[I_i(\tau)],$$
The conditional distributions of $Z_{ji}(t)$ and $\tilde{Z}_{ji}(t)$ are

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

and

\[
\Pr(\tilde{Z}_{ji}(t) = 1|b, p, \theta, \phi, \tilde{t}_i) = \begin{cases} 
g(\tilde{t}_i|t) \times \left\{ \frac{\Pr[I_i(t)] - \Pr[I_{ji}(t)]}{\Pr[S_i(\tilde{t}_i)]} \right\} + \sum_{\tau=t+1}^{\tilde{t}_i} g(\tilde{t}_i|\tau) \times \Pr[I_i(\tau)] \times \Pr[S_i(\tilde{t}_i)], & t_i \leq t < \tilde{t}_i \\
1, & t < t_i \\
0, & \text{otherwise}
\end{cases}
\]
• Variance estimation

Let \( Z = \{Z_{ji}(t), \tilde{Z}_{ji}(t)\} \), \( \tilde{t} = \{\tilde{t}_i\} \), and \( \lambda = \{b, p, \theta, \phi\} \). Louis’ method states that

\[
\frac{\partial^2 \log(L(\lambda|\tilde{t}))}{\partial \lambda^2} = \mathbb{E}_{Z|\tilde{t},\lambda}\left\{ -\frac{\partial^2 \log(L(\lambda|\tilde{t}, Z))}{\partial \lambda^2} \right\} \\
+ \text{VAR}_{Z|\tilde{t},\lambda}\left\{ -\frac{\partial \log(L(\lambda|\tilde{t}, Z))}{\partial \lambda} \right\}
\]
Linear Model Based on Augmented Data

- Log-linear model

Modelling pairwise transmission:

\[ p_{ji}(t) = \begin{cases} 
\theta r_i(t) \phi r_j(t) p f(t|\tilde{t}_j), & j \in D_i \\
\theta r_i(t) b, & j = c
\end{cases} \]

Define \( \psi_j \) (1: household member, 0: community), then

\[ \log(p_{ji}(t)) = \beta_0 + \beta_1 \psi_j + \beta_2 r_i(t) + \beta_3 r_j(t) + \log(f(t|\tilde{t}_j)) \]

where \( b = e^{\beta_0}, \ p = e^{\beta_0 + \beta_1}, \ \theta = e^{\beta_2}, \ \phi = e^{\beta_3}. \)
• Log-linear model in final-size model (Haber, Longini and Cotsonis, 1988)

$$\log m_{jk} - \log m_{jj} = \beta_1 (k - j) + \beta_2 j (k - j) + \log \binom{k}{j}$$

where $B = e^{\beta_1}$, $Q = e^{\beta_2}$. WLS can be used.
• **E-step** Assign weight \( \Pr(Z_{ji}(t) = 1|b, p, \theta, \phi, \tilde{t}_i) \) to response 1 and \( \Pr(\bar{Z}_{ji}(t) = 1|b, p, \theta, \phi, \tilde{t}_i) \) to response 0.

• **M-step** (Iteratively Re-weighted Least Square)
  - Summarize binary responses into proportions \( P_h, h = 1, \ldots, H \), where \( H \) is the number of covariate patterns.
  - Minimize objective function \( \sum_{h=1}^{H} w_h \{ \log(\tilde{P}_h) - \log(P_h) \}^2 \), where \( \tilde{P}_h \) is observed proportion.
  - If \( \tilde{P}_h = 0 \), replace it with \( \hat{P}_h \), the fitted proportion.
  - Note that \( \text{VAR}(\log(\tilde{P}_h)) = \frac{1-P_h}{n_h \times P_h} \), set \( w_h = \sqrt{\frac{n_h \times \tilde{P}_h}{1-P_h} \times \frac{n_h \times \tilde{P}_h}{1-P_h}} \).

• **Variance estimation**
  \[ \text{VAR}\{\hat{\lambda}\} = \mathbb{E}\{\text{VAR}[\hat{\lambda}|Z]\} + \text{VAR}\{\mathbb{E}[\hat{\lambda}|Z]\} \]
• Adjustment for left truncation and selection bias for case-ascertained follow-up scheme

  – Problem: adjustment with straightforward statistical meaning is difficult.
  
  – Solution: borrow the adjusting term in the likelihood method.

   Adjusted objective function:

   \[
   \sum_{h=1}^{H} w_h \left\{ \log(\tilde{P}_h) - \log(P_h) \right\}^2 + \sum_i \log(A_i(\beta)).
   \]

   Then, at the \( l^{th} \) iteration,

   \[
   \hat{\beta}_l = (X'WX)^{-1}\left\{ X'WY - \frac{1}{2} \sum_i \frac{d \log(A_i(\hat{\beta}_{l-1}))}{d \hat{\beta}_{l-1}} \right\}
   \]
Simulation study

Table 6: Comparison between MLEs and IRLS estimates. Results are based on 1000 simulations.†‡

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean of Point Estimates</th>
<th>Monte Carlo SD</th>
<th>Mean of SD Estimates</th>
<th>Coverage of 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE</td>
<td>IRLS</td>
<td>MLE</td>
<td>IRLS</td>
</tr>
<tr>
<td>$b$</td>
<td>0.0051</td>
<td>0.0051</td>
<td>0.00028</td>
<td>0.00028</td>
</tr>
<tr>
<td>$p$</td>
<td>0.10</td>
<td>0.10</td>
<td>0.011</td>
<td>0.011</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.40</td>
<td>0.40</td>
<td>0.067</td>
<td>0.067</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.71</td>
<td>0.71</td>
<td>0.13</td>
<td>0.13</td>
</tr>
</tbody>
</table>

† True parameters are set to $b=0.005$, $p=0.1$, $\theta=0.40$, $\phi=0.70$.
‡ MLEs are the same for observed and augmented likelihoods.
Table 7: Comparing sensitivity to initial estimates between observed likelihood, augmented likelihood and IRLS methods for sparse data. Results are based on 1000 simulations.†

<table>
<thead>
<tr>
<th>Initial Values $(b_0, p_0)$‡</th>
<th>Method</th>
<th>Conv. Rate (/1000)</th>
<th>$b$</th>
<th>$p$</th>
<th>$\theta$</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(0.002, 0.01)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs. ML</td>
<td>903</td>
<td>0.0020 (0.00016)</td>
<td>0.010 (0.0049)</td>
<td>0.42 (0.25)</td>
<td>0.98 (0.95)</td>
<td></td>
</tr>
<tr>
<td>Aug. ML</td>
<td>889</td>
<td>0.0020 (0.00016)</td>
<td>0.010 (0.0048)</td>
<td>0.42 (0.24)</td>
<td>1.01 (0.96)</td>
<td></td>
</tr>
<tr>
<td>IRLS</td>
<td>937</td>
<td>0.0020 (0.00016)</td>
<td>0.011 (0.0047)</td>
<td>0.48 (0.24)</td>
<td>1.07 (0.93)</td>
<td></td>
</tr>
<tr>
<td>$(0.02, 0.1)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs. ML</td>
<td>524</td>
<td>0.0020 (0.00016)</td>
<td>0.010 (0.0048)</td>
<td>0.41 (0.24)</td>
<td>1.19 (1.11)</td>
<td></td>
</tr>
<tr>
<td>Aug. ML</td>
<td>878</td>
<td>0.0020 (0.00016)</td>
<td>0.010 (0.0049)</td>
<td>0.42 (0.24)</td>
<td>0.99 (1.00)</td>
<td></td>
</tr>
<tr>
<td>IRLS</td>
<td>920</td>
<td>0.0020 (0.00016)</td>
<td>0.011 (0.0048)</td>
<td>0.48 (0.24)</td>
<td>1.07 (1.00)</td>
<td></td>
</tr>
<tr>
<td>$(0.0002, 0.001)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs. ML</td>
<td>92</td>
<td>0.0020 (0.00016)</td>
<td>0.010 (0.0054)</td>
<td>0.38 (0.23)</td>
<td>1.04 (0.79)</td>
<td></td>
</tr>
<tr>
<td>Aug. ML</td>
<td>864</td>
<td>0.0020 (0.00015)</td>
<td>0.010 (0.0047)</td>
<td>0.44 (0.26)</td>
<td>1.03 (1.08)</td>
<td></td>
</tr>
<tr>
<td>IRLS</td>
<td>928</td>
<td>0.0020 (0.00015)</td>
<td>0.011 (0.0047)</td>
<td>0.49 (0.24)</td>
<td>1.08 (0.90)</td>
<td></td>
</tr>
</tbody>
</table>

† True parameters are set to $b=0.002$, $p=0.01$, $\theta=0.40$, $\phi=0.70$.
‡ Initial values for $\theta$ and $\phi$ are set to the true values.
§ Values in the parentheses are Monte Carlo standard deviations.
# Data Analysis on Trials of Zanamivir

Table 8: Two randomized multi-center trials of zanamivir, an influenza antiviral agent

<table>
<thead>
<tr>
<th></th>
<th>Hayden et al., 2000</th>
<th>Monto et al., 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Households</strong></td>
<td>336</td>
<td>484</td>
</tr>
<tr>
<td><strong>Population</strong></td>
<td>1186</td>
<td>1770</td>
</tr>
<tr>
<td><strong>Index case randomization</strong></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Duration of medication</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Index case</td>
<td>5 days</td>
<td>N/A</td>
</tr>
<tr>
<td>Contact</td>
<td>10 days</td>
<td>10 days</td>
</tr>
<tr>
<td>Follow up (symptom diary)</td>
<td>14 days</td>
<td>14 days</td>
</tr>
<tr>
<td>Infected(^\d)/Symptomatic(index)</td>
<td>164/336</td>
<td>281/484</td>
</tr>
<tr>
<td>Infected(^\d)/Exposed(contacts)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>52/435</td>
<td>76/626</td>
</tr>
<tr>
<td>Zanamivir</td>
<td>17/415</td>
<td>27/660</td>
</tr>
</tbody>
</table>

Numbers may slightly differ from references due to different criteria of data inclusion for analysis.
\(^\d\) Laboratory-confirmed infections with clinical symptoms
Table 9: Estimates of efficacies and transmission probabilities by age (1-17 vs. 18+) for pooled zanamivir trials conducted in 1998-1999 and 2000-2001.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>IRLS</th>
<th>MLE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point Estimate</td>
<td>SD</td>
<td>Point Estimate</td>
</tr>
<tr>
<td>$b_c$</td>
<td>0.0024</td>
<td>0.00052</td>
<td>0.0028</td>
</tr>
<tr>
<td>$b_a$</td>
<td>0.00086</td>
<td>0.00030</td>
<td>0.0010</td>
</tr>
<tr>
<td>$p_{cc}$</td>
<td>0.040</td>
<td>0.0074</td>
<td>0.040</td>
</tr>
<tr>
<td>$p_{ca}$</td>
<td>0.028</td>
<td>0.0045</td>
<td>0.029</td>
</tr>
<tr>
<td>$p_{ac}$</td>
<td>0.023</td>
<td>0.0071</td>
<td>0.020</td>
</tr>
<tr>
<td>$p_{aa}$</td>
<td>0.040</td>
<td>0.011</td>
<td>0.032</td>
</tr>
<tr>
<td>AVE$_S$</td>
<td>0.68</td>
<td>0.086</td>
<td>0.75</td>
</tr>
<tr>
<td>AVE$_I$</td>
<td>0.24</td>
<td>0.38</td>
<td>0.23</td>
</tr>
<tr>
<td>AVE$_T$</td>
<td>0.81</td>
<td>0.094</td>
<td></td>
</tr>
</tbody>
</table>

† Subscript $c$ denotes child (1-17), $a$ denotes adult (18+), and $ca$ denotes child-to-adult transmission.
Summary

- Amount of information in data determines which model to use.
- Study design is important for statistical inference.
- Left truncation and selection bias can be adjusted for by conditioning on disease status on the symptom onset day of index cases.
- Methods based on data-augmentation are less sensitive to initial estimates.
- Linear model can be used for initial estimates.
- Invaded households might be more prone to infection. Random effect? (Halloran et al., 2003).
- Depending on definition of clinical symptoms, there may be 30% or more asymptomatic (silent) cases. Bayesian method? (O’Neill et al.,
2000).

- How to deal with index cases with negative laboratory test results.