Assessing incomplete sampling of disease transmission networks

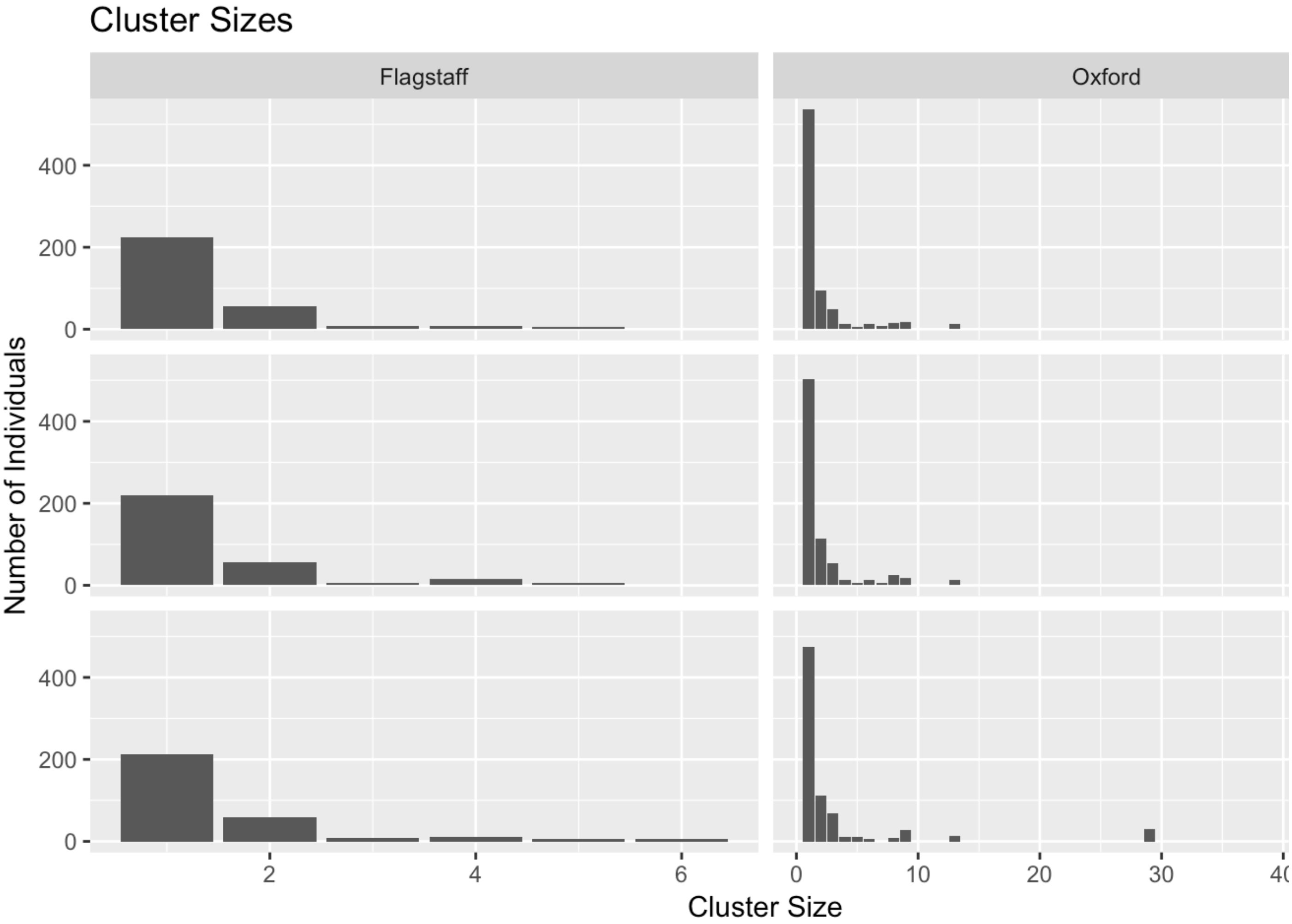
PMI Monthly Meeting

Derek Sonderegger, PhD - Northern Arizona University

May 8, 2019

Collaboration with NAU's Pathogen and Microbiome Institute

Cluster Size Distributions



3/20

Defining γ = HAI rate from full data

- For each cluster, the first time a strain is observed it is considered environmentally acquired.
- The second (or third, or fourth, ..) time a strain is observed, it is healthcare acquired.

$$\gamma = \frac{N - ||\mathcal{I}||}{N} = 1 - \frac{||\mathcal{I}||}{N}$$

$$N = \text{ Number of Patients}$$

$$\mathcal{I} = \text{ Set of strain identifiers}$$

$$||\mathcal{I}|| = \text{ Actual Number of Clusters/Strains}$$

- Knowing $||\mathcal{I}||$ is the key to calculating HAI rate!

Observed Number of Clusters/Strains under Simple Random Sampling

- Define the following

$$\alpha = \text{ proportion of the population sampled}$$

$$n_i = \text{ actual size of the } i\text{th cluster}$$

$$m_i = \text{ observed size of the } i\text{th cluster}$$

Notice that

$$1 \leq m_i \leq n_i$$

and

$$\sum n_i = N$$

$$\sum m_i = \alpha N$$

Conditional Distribution

$$m_i|n_i \sim \text{ZTHyperGeometric}(n_i,\ N - n_i,\ \alpha N) \text{ for } i$$

- Zero Truncated HyperGeometric
- Assume approximate independence between observed cluster sizes
- Distribution requires working with hypergeometric terms

$$f(0|n_i) = \frac{\binom{n_i}{0}\binom{N-n_i}{\alpha N}}{\binom{N}{\alpha N}}$$

Notice that α and $f(0|n_i)$ are inversely related and we could crudely approximate

$$f(0|n_i) \approx 1 - \alpha$$

Critical Expectation

$$E[m_i] = E[E(m_i|n_i)] = E[(1 - f(0|n_i))^{-1} \ \alpha \ n_i]$$

Utilizing this equation, can derive two different estimators.

1. The plug-in estimator that ignores the expectation, and approximates $[1 - f(0)]^{-1} \approx \alpha^{-1}$. This results in $\widehat{n}_i = m_i$.
2. Ignoring the expectations, we could utilize the actual hypergeometric function for $f(0|n_i)$ and solve the following equation for $\widehat{n}_i$. This solution needs to be solved via numerical methods because the “chooses” in $f(0|n_i)$.

Biased Estimator

- Denoting

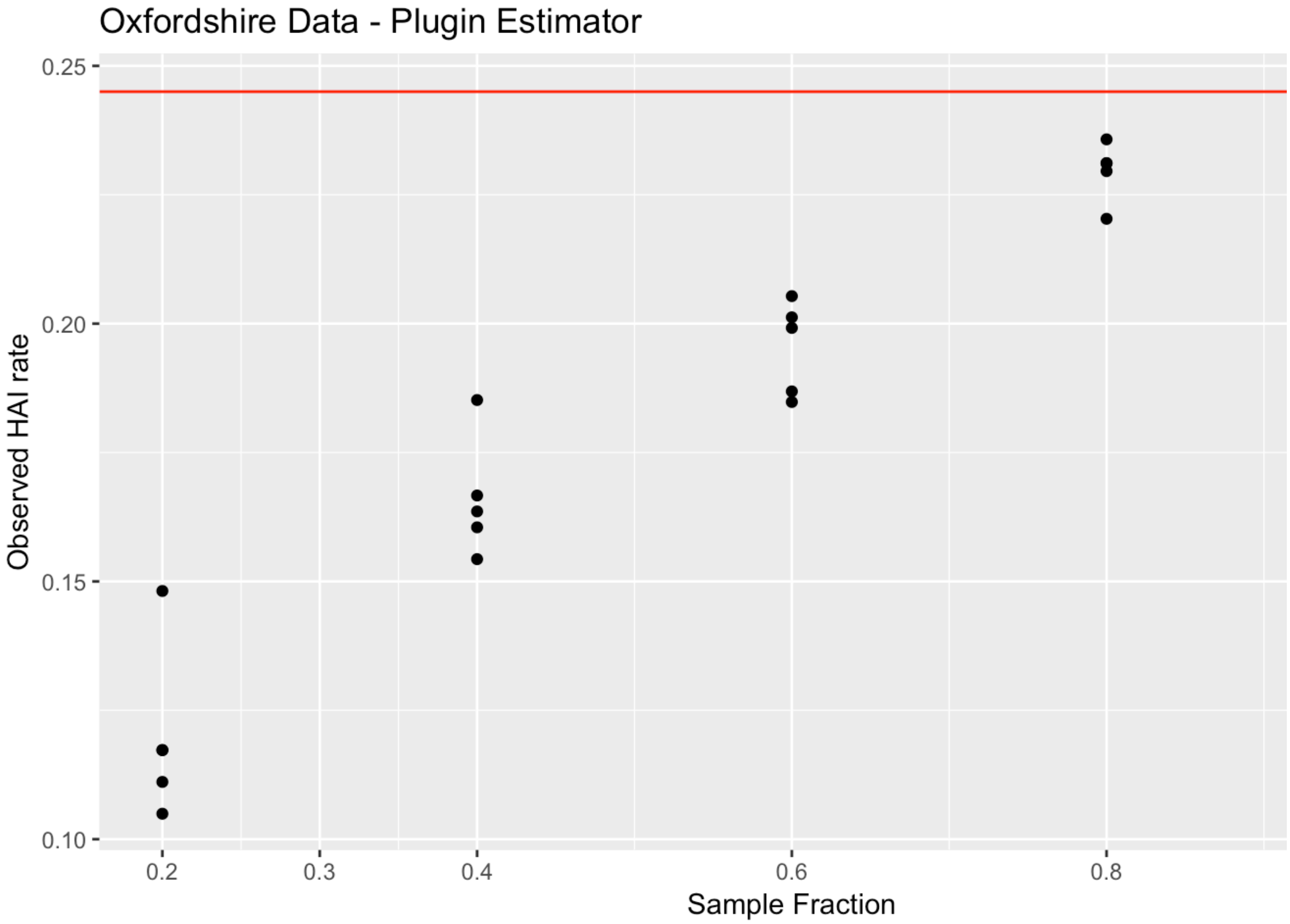
$$\widehat{n} = \sum \widehat{n}_i$$

$$I = \text{ Set of observed strains}$$

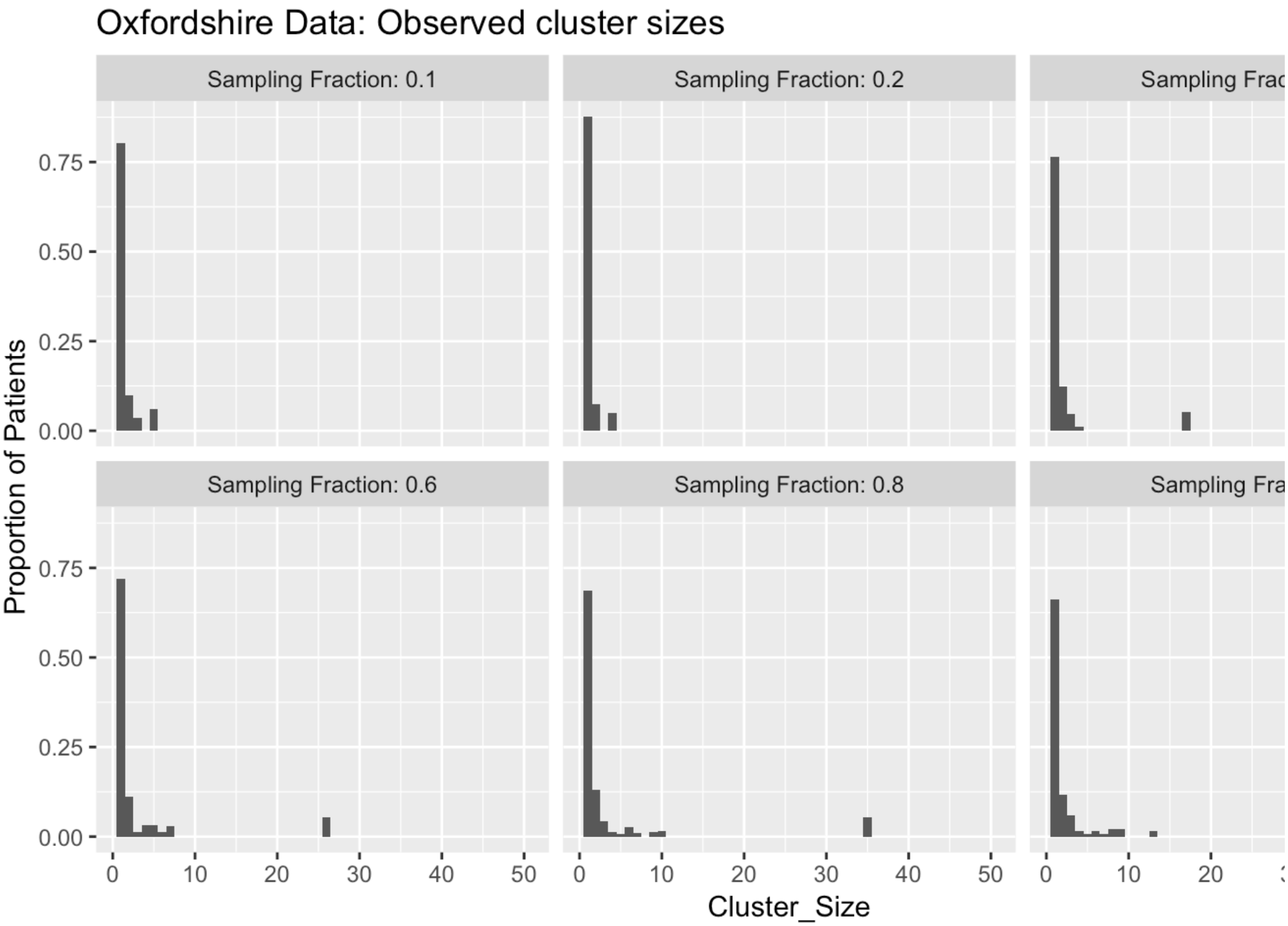
$$||I|| = \text{ Observed Number of Clusters/Strains}$$

$$\widehat{\gamma}^* = \frac{1}{\widehat{n}} \sum_{i \in I} (\widehat{n}_i - 1) = \frac{\widehat{n} - ||I||}{\widehat{n}} = 1 - \frac{||I||}{\widehat{n}}$$

Does the plug-in Estimator Work?



Why doesn’t this work?



10/20

Bias Correction Procedure

1. Calculate the sample HAI rate.
2. Repeatedly subsample the sample at the designated α fraction.
3. For each subsample, calculate the subsample's HAI rate
4. Look at the average discrepancy and use that to adjust the sample HAI rate estimate.
5. The adjustments are made on the logit scale to force the resulting rate to remain in the $[0, 1]$ interval.

Bias Correction Procedure - Math!

By repeatedly sub-sampling at α rate J times and calculating $\widehat{\gamma}_j^*$ for the jth sub-sample,

$$\bar{\delta} = \frac{1}{J} \sum \left[\text{logit}(\widehat{\gamma}^*) - \text{logit}(\widehat{\gamma}_j^*)\right]$$

$$\widehat{\gamma} = \text{ilogit}\left(\text{logit}(\widehat{\gamma}^*) + \bar{\delta}\right)$$

We performed the bias correction step on the logit scale to ensure the resulting estimator is in $[0, 1]$.

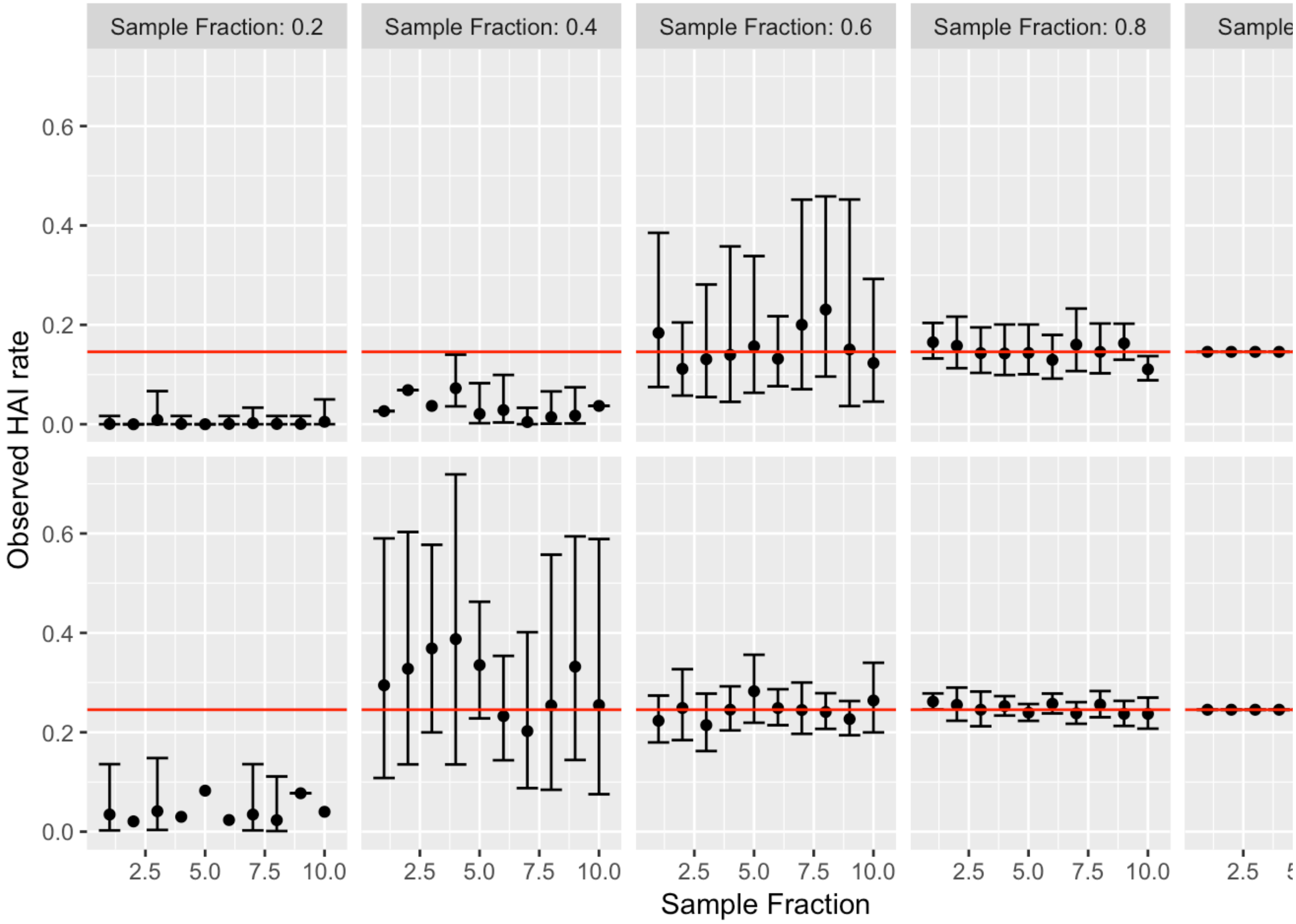
Get approximate Confidence Intervals too!

- Standard deviation of the $\text{logit}(\widehat{\gamma}_j^*)$ values gives a estimated standard error of $\text{logit}(\widehat{\gamma})$ value.
- An approximate 95 confidence interval for γ we use is to add/subtract

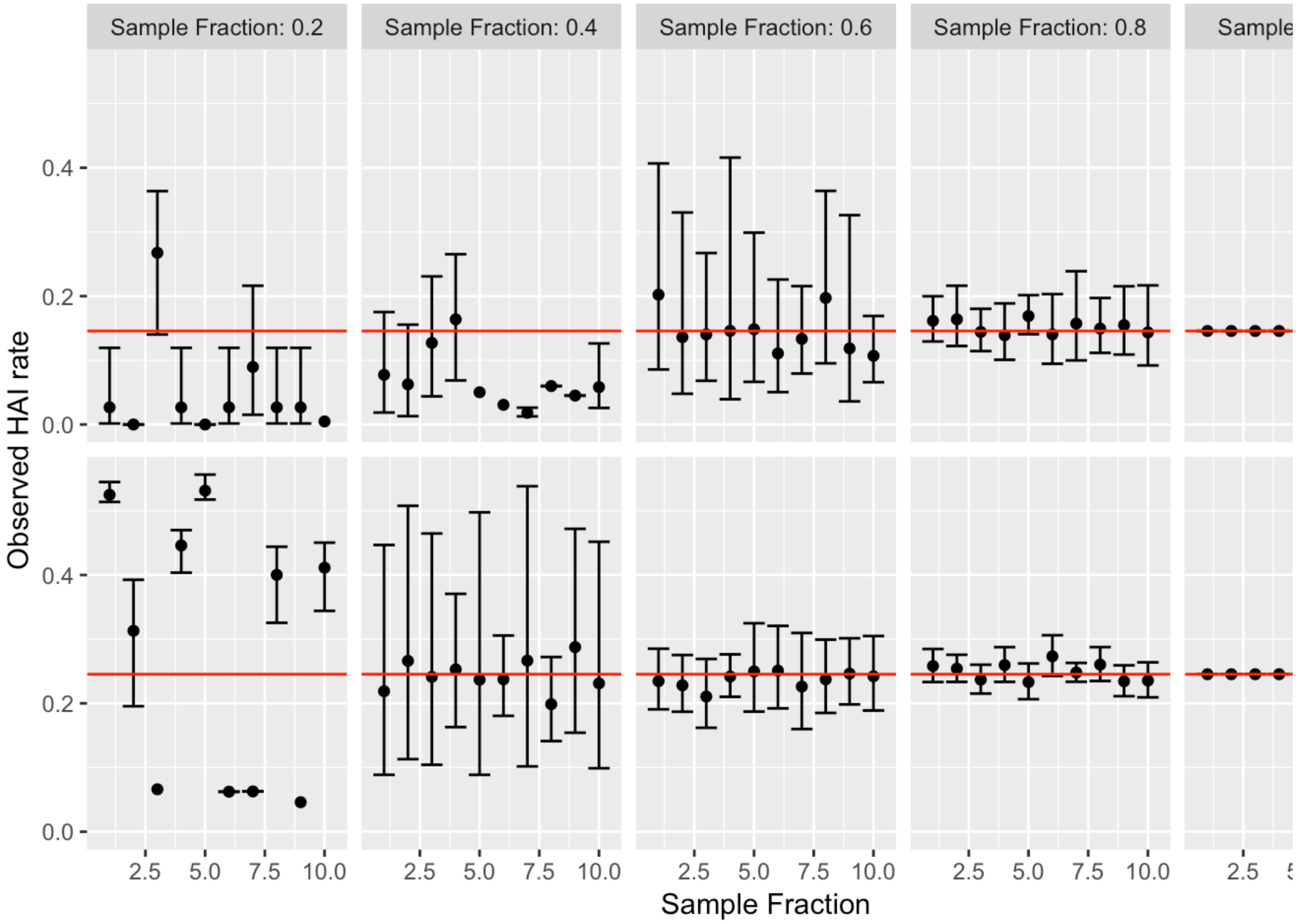
$$\text{ilogit}\left[\text{logit}(\widehat{\gamma}) \pm Z_{0.975} * SE(\text{logit}(\widehat{\gamma}))\right]$$

Results

Plugin Results - Clinical Data



Hypergeometric Results - Clinical Data



16/20

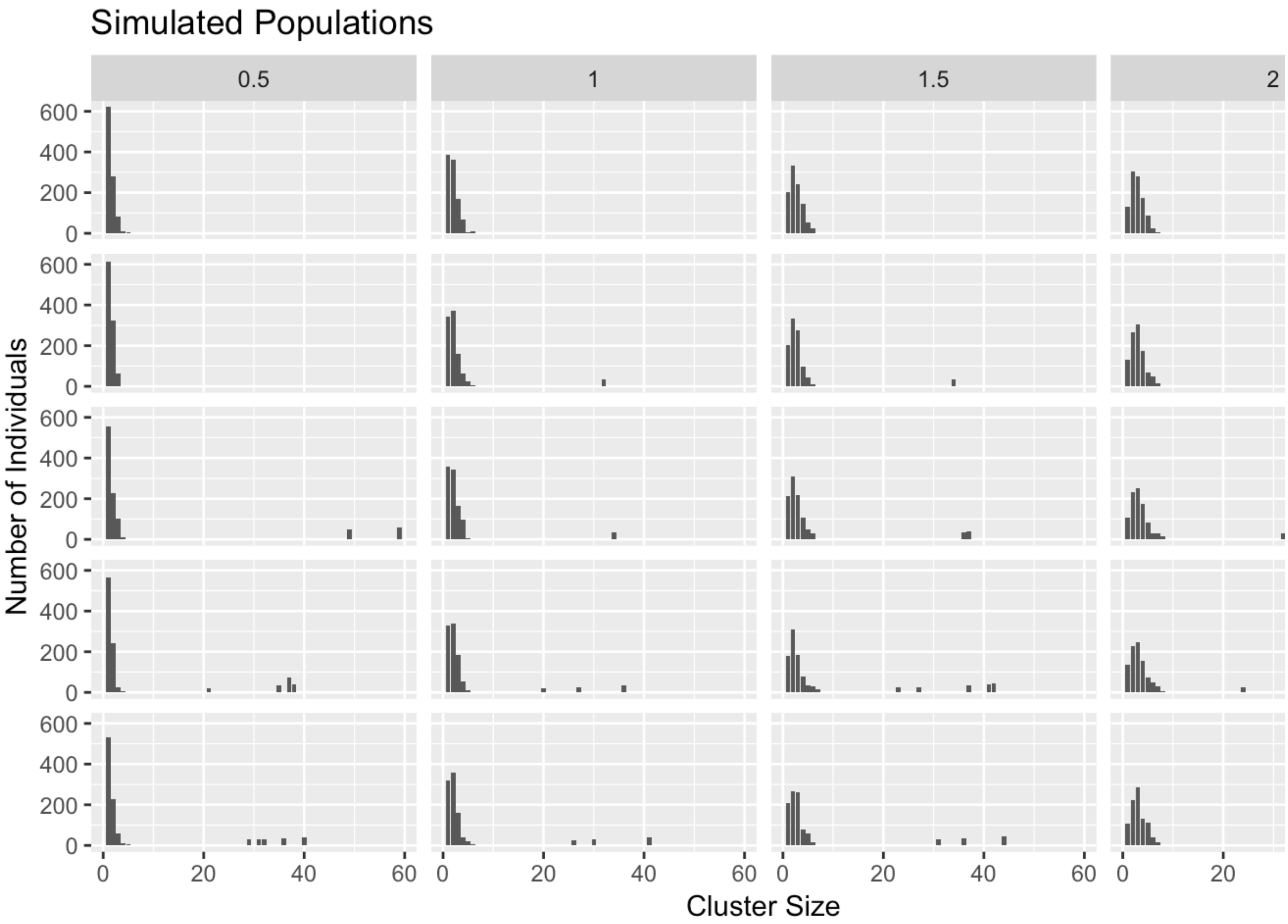
Results - Simulated Populations

The Oxfordshire data could be reasonably modeled using a mixture of two distributions to separate the small clusters sizes from the large. We chose to model the small clusters sizes using a truncated Poisson distribution with the zero truncated out. The large cluster sizes were modeled from a logNormal distribution.

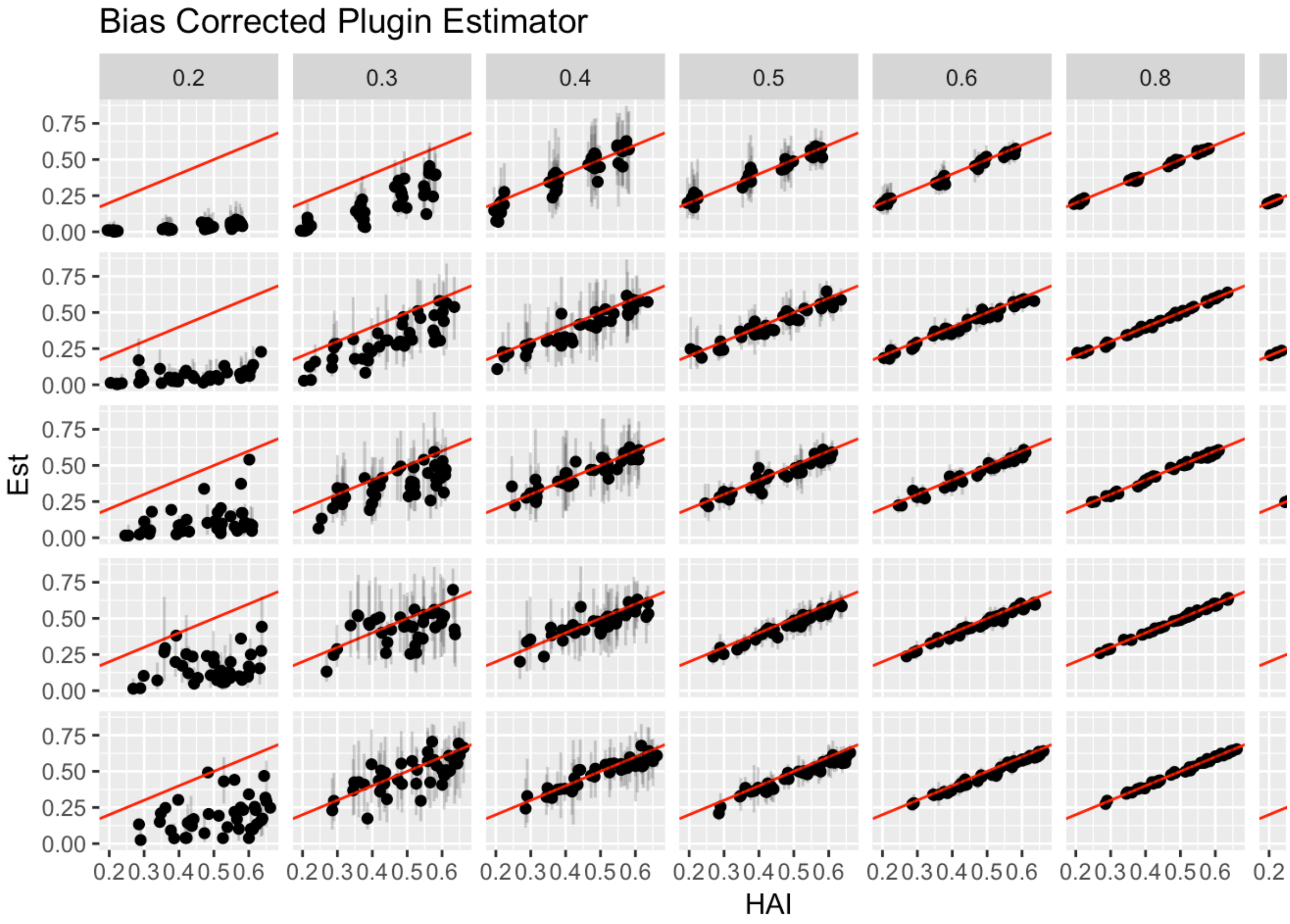
$$n_i \sim \begin{cases} \text{TPoisson}(\lambda) & \text{with probability } 1-\rho \\ \text{logNormal}(\mu, \sigma) & \text{with probability } \rho \end{cases}$$

for i in $\mathcal{I}$.

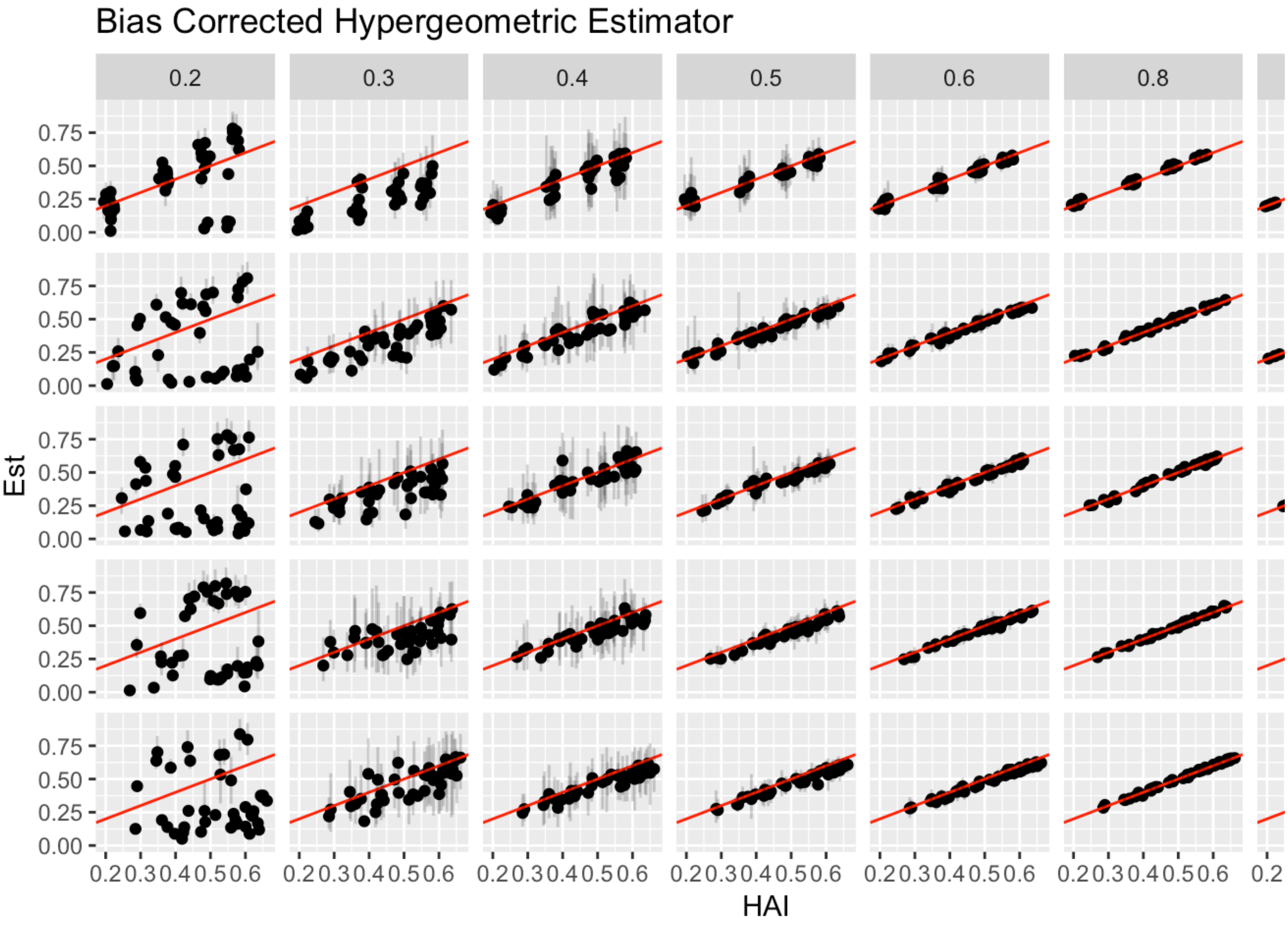
Simulated Data Populations



Simulated Data Populations: Results



Simulated Data Populations: Results


Bias Corrected Hypergeometric Estimator
Est
HAI