Assessing incomplete sampling of disease transmission networks

Dept. Mathematical Sciences, Montana State University

Derek Sonderegger, PhD - Northern Arizona University

March 13, 2019

Colaborators

- Work that I have done with the Pathogen and Microbiome Institute at NAU and we are just a couple months into the project.

PATHOGEN & MICROBIOME INSTITUTE
Northern Arizona University

- Dr Paul Keim
- Dr Jason Sahl

Background Information

Two worrisome Healthcare Aquired Infections (HAIs)

- MRSA
  - Methicillin-resistant Staphylococcus aureus
  - Resistant to many common antibiotics
- *C. Diff*
  - *Clostridioides difficile*
  - Our disease of interest

Clostridioides difficile

- A spore-forming bacteria
  - Spores can survive for months in the environment
  - Bacteria die when exposed to oxygen.
  - Very difficult to work with in the lab.
- *C. diff* is widely distributed
  - Spores are widely found in the environment
  - People and animals can be asymptomatic carriers
- Resistant to many commonly used antibiotics

Human Infection

- Causes diarrhea, fever, nausea, and abdominal pain
- Spread through fecal contamination
- Additional $4.8 billion each year in health care costs
    - 290,000 Americans sickened by the bacteria in a hospital or other health care facility each year.
    - 27,000 people in the U.S. die while infected with *C. diff* annually.

Common infection cycle

- In a healthy gut biome, *C. diff* can't strongly establish due to bacterial competition.
- In patients under a common antibiotic treatment, *C. diff* can flourish.
- Prescribed antibiotics for some other reason (e.g. pneumonia)
    - *C. diff* might already be present in the patient.
    - Come into contact with *C. diff* via live bacteria or spores from another patient.

Medicare Implications

- Won't reimburse costs for treating infections acquired at a healthcare facility
- If the rate of Healthcare Acquired Infections (HAIs) is too high, Medicare will deduct one percent from their OVERALL reimbursements to the facility.
- Medicare defines any diagnosis of *C. diff* that occurs 3 days after admission as "healthcare acquired".

Goal: Estimate HAI rate

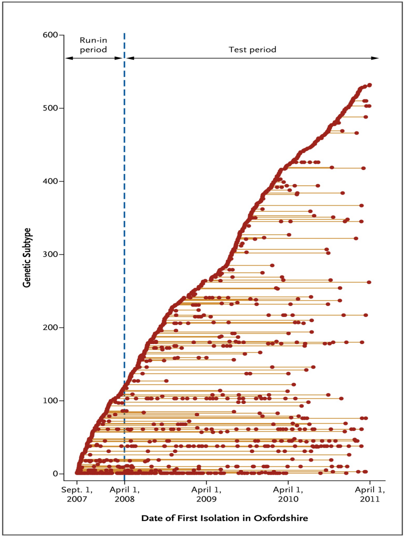
- Individual patients have the genome of their strain of *C. diff* sequenced.
- Group strains into clusters if they differ by at most 2 SNPs.
    - Use Single-Linkage clustering method: represents evolution along a chain of infections
    - Within patient variability suggests that maybe this needs to be evaluated.

Data!

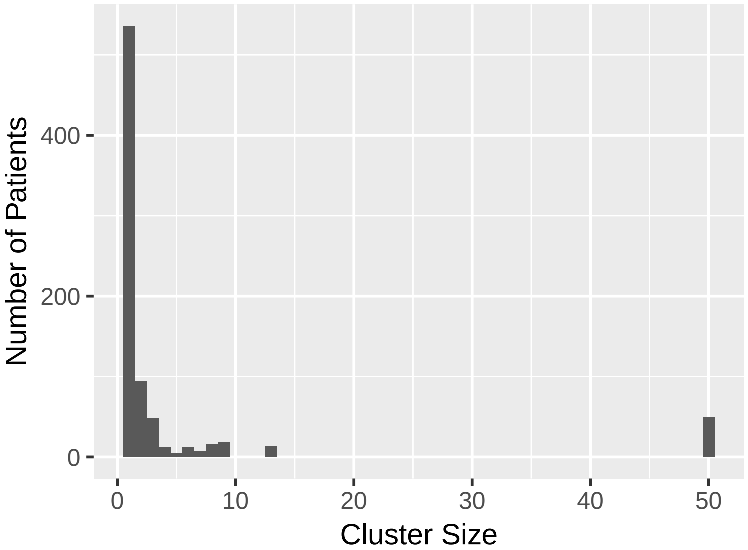
Oxfordshire Data

- Eyre *et al* 2013 describes a study which genotyped nearly all cases of *C. diff* in over three years in Oxfordshire, UK.
- Of the 1250 cases that were evaluated, $N = 1223$ were successfully genotyped.

Oxfordshire Time/Clusters



Oxfordshire Cluster Size Distribution



Defining 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.

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

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

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

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

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

Observed Number of Clusters/Strains under Simple Random Sampling

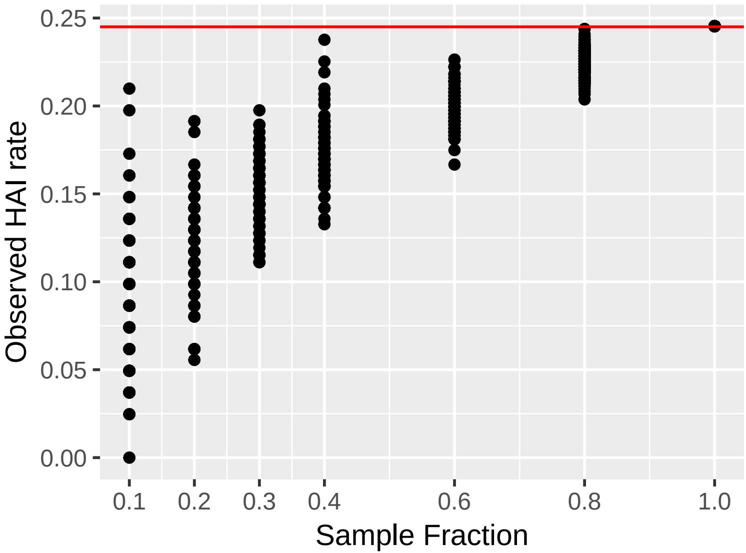
$$\widehat{HAI}_{naive} = 1 - \frac{||I||}{n}$$

$$n = \text{ sample size}$$

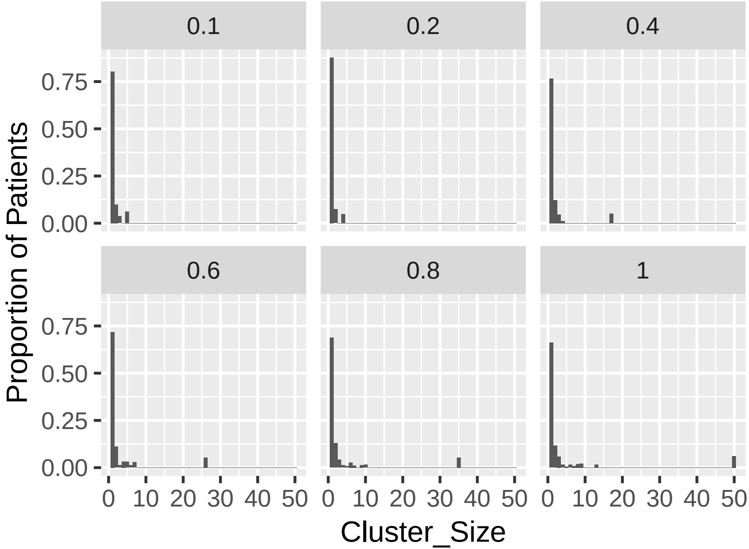
$$I = \text{ Set of observed strains}$$

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

Does the Naive Estimator Work?



Why doesn't this work?



Better Estimators?

HyperGeometric?

- Let
  - n_i be the number of patients with strain i. (This is unknown!)
  - m_i be the observed number of patients with strain i.
  - α be the sampling percentage.

$$n_i \overset{iid}{\sim} \mathrm{ZTPoisson}(\lambda) \text{ for } i \in \mathcal{I}$$

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

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

where I is a subset of $\mathcal{I}$ and the ZT represents the zero truncated distributions.

HyperGeometric?

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

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

Can we just ignore the expectation?

One estimator is to ignore the expectation and solve the following equation for $\widehat{n}_i$.

$$m_i = (1 - f(0|\widehat{n}_i))^{-1} \, \alpha \, \widehat{n}_i$$

which needs to be solved via numerical methods because the "chooses" in $f(0|\widehat{n}_i)$.

$$\widehat{HAI}_{hyper} = 1 - \frac{||I||}{\widehat{n}}$$

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

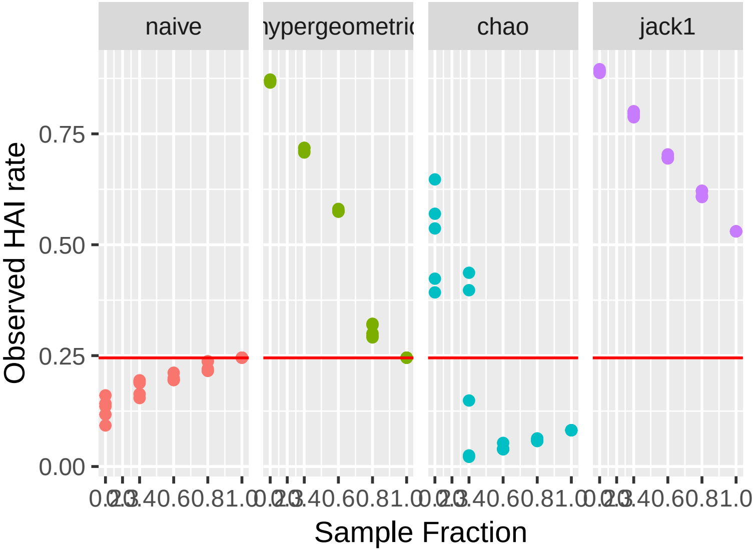
$$I = \text{ Set of observed strains}$$

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

Species Abundance Methods

- A well studied problem is estimating the total number of species based on repeated surveys.
- Each patient represents a survey, which might produce a new strain, or one that has already been seen.
- Several estimators for this problem
  - Chao, Jacknife1, Jacknife2, Bootstrap
  - I'll show Chao and Jacknife
- Used `vegan::specpool`

Estimators of Oxford Data



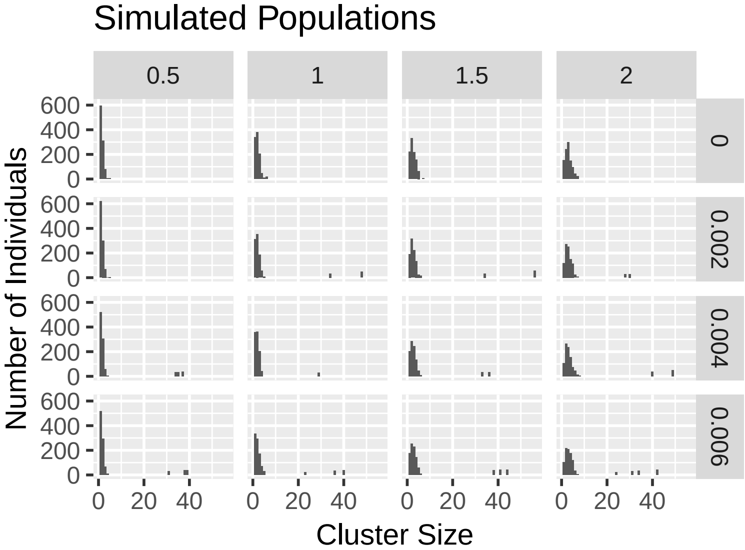
Simulated Populations

- 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.

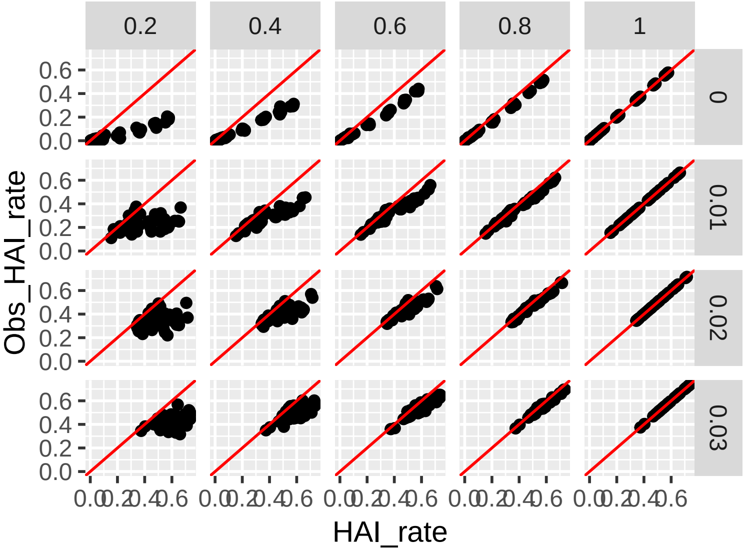
$$n_i \sim \begin{cases} \text{TPoisson}(\lambda) & \text{with probability } 1-\alpha \\ \text{logNormal}(\mu, \sigma) & \text{with probability } \alpha \end{cases}$$

for i in $\mathcal{I}$.

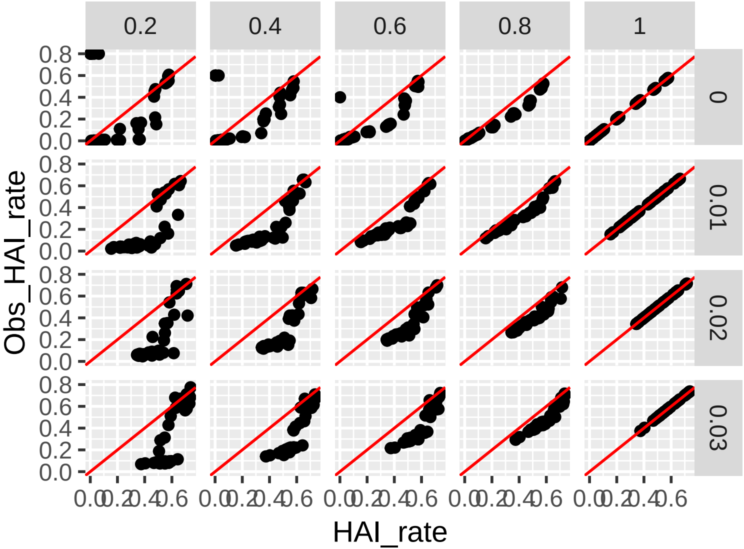
Simulated Populations



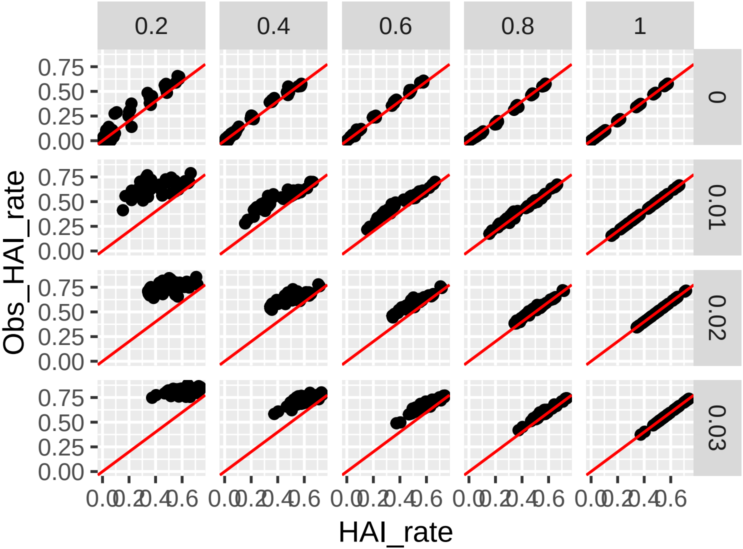
Simulated Data: Naive method



Simulated Data: Chao



Simulated Data: Hypergeometric



Next Steps

- Improve Large Cluster Estimation
  - Evaluate
$$E\left[\binom{N-n_i}{\alpha N}\alpha\ n_i\right]$$
  - Stirling's Approximation?
  - Needs some assumptions about distribution of n_i.
- Confidence Interval for HAI
  - Bootstrap clusters or patients?