Transmission trees contain valuable details about who infected whom in infectious disease outbreaks. Here, we created a database with 81 published, standard transmission trees consisting of 12 directly-transmitted pathogens (mostly viruses). We also demonstrated how the database could be used to help answer research questions in infectious disease epidemiology. Next, we analyzed overall and pathogen-specific patterns between tree parameters (R<sub>e</sub> and variation in secondary infections). We found that outbreak size is nonlinearly associated with R<sub>e</sub> and the dispersion parameter, but additional research is needed to test this hypothesis. Pathogen-specific patterns and intervention efforts may alter theoretical relationships between these variables. Second, we examined how superspreader contribution to onward transmission, either directly or through their tree descendants, varies across pathogens. Superspreaders were responsible for most cases via their descendants and the number of superspreaders varied across pathogens. Additional database exploration matched theory about how the proportion of superspreaders increases at intermediate levels of dispersion, an idea that should be further explored. We hope that our database will assist both theoretical and applied infectious disease epidemiology research.

**INTRODUCTION**

When an infectious disease outbreak occurs, epidemiologists seek time and money-intensive investigations to determine how the outbreak started and the patterns of disease transmission. They often store this information in a transmission tree, where individuals are represented by nodes, and disease transmission by branches. From these trees, one can calculate key statistics like R<sub>e</sub>, the dispersion parameter (variation in R<sub>e</sub>), pathogen mutation rate, and intervention efficacy, though greater standardization in tree format would make the trees and statistics more comparable. Additionally, a better understanding of predictors of outbreak size is important and the importance of superspreaders to onward transmission across different pathogens would help researchers develop transmission tree reconstruction methods and inform public health intervention efforts. This project is our first attempt at standardizing and compiling transmission trees into an open-access database that can be a resource for future researches.

**DATABASE CONSTRUCTION**

- Average individual reproductive number (R<sub>e</sub>): average number of secondary infections across all individuals in the tree
- Initial R<sub>e</sub>: average number of secondary infections in the first two generations of the outbreak, the second generation tree may have zero individuals
- Secondary infections were assumed to follow a negative binomial distribution
- Cases caused directly by superspreaders were defined as the individuals personally infected by the superspreader. Overall cases were defined as all individuals for whom their common infection “ancestor” was the superspreader. They were infected as a result of the superspreader infecting others, even if several generations later.

**TREE ANALYSIS**

- Transmission trees contain valuable information about specific pathogen outbreaks, which is costly to collect. Our database standardizes tree format, allowing for greater comparative analysis.
- Understanding factors associated with increased outbreak size may help predict the extent of outbreaks in the future and lead to more effective preventive measures.
- The impact of superspreading was quantified using a new statistic which we call overall efficacy, which suggested that superspreaders are perhaps more important than previously realized.
- This database provides the information to test theoretical hypotheses about disease transmission and inspire new ideas, such as:
  - How sensitive is outbreak size and length to superspreader introduction timing?
  - Does knowing the transmission tree of a disease allow us to predict the mode of transmission or type of pathogen (bacterial or viral)?

**CONCLUSIONS**

**REFERENCES**


**ACKNOWLEDGEMENTS**

1. Bowdoin College, Brunswick, ME; 2. Odum School of Ecology, University of Georgia, Athens, GA

Juliana Taube, Paige B. Miller, and John M. Drake.

**EXAMPLE QUESTIONS**

1. What is the relationship between individual variation in number of secondary infections, R<sub>e</sub> and total outbreak size?

2. What is the quantitative contribution of superspreading to outbreak size?

3. What determines the frequency of superspreading events?