Introduction

Much of the world is complex, and in seeking to understand it we must make aggregations and assumptions to create groups ***[REF]***. When quantifying a student's academic achievements, or measuring distances, for example, decisions are made to create discrete categories from the underlying continuous data. Categorization e.g., as grades, is essential to the synthesis and interpretation of information, particularly for action; as it would be impractical to simultaneously evaluate every score a student achieved during school ***[REF]***. Infectious diseases are no different. Rather than tracking measles viral loads in a population, for example, individuals are broadly categorized as susceptible, infected, or removed ***[REF]***. Cases are counted, providing estimates of the number of new and cumulative infected individuals, respectively, at any one time; without these groupings, it would be hard to answer questions about disease spread and burden, and allocate resources like preventative vaccines appropriately ***[REF]***. This approach, however, leaves us with some questions; namely, how many groups are appropriate, how should the breakpoints between groups be defined, and are there meaningful differences between the groups that allow for inferences about the system in question? At every scale in an infectious disease system, from variability of infectivity within an individual's infection cycle, to defining outbreaks in a population from the accumulation of infections, these questions must be addressed. In this dissertation, I explore how variability in continuous measures can be discretized, and the interactions that arise from compounding these categorization decisions.

In the first half of my dissertation (Chapters 2 & 3), I explore how differences in infection rates between geographically co-incident groups can be evaluated in the context of the categorization process. In the spring of 2020, the COVID-19 pandemic resulted in many university campuses across the US to shut down, requiring their students to return to their respective homes ***[REF]***. When students were re-introduced to the Pennsylvania State University campus during the start of the Fall 2020 semester, two spatially entwined, but demographically and behaviorally disparate groups were defined: returning students and the surrounding community members. Through this grouping, it is now possible to characterize the burden of SARS-CoV-2 infection (the underlying virus that causes the disease COVID-19). Without discrete categories, there is no denominator for in use calculations of seroprevalence (the proportion of a population that have sufficiently high levels of antibodies, indicating past exposure to a pathogen). In Chapter 2, I show that substantial, unexpected, differences in infection rates can be observed between the student and community populations, highlighting that opportunities exist for infection mitigation efforts to minimize spread between spatially-linked subgroups of a population. To examine differences in COVID-19 infections that may exist in the student body, it was, once again, imperative to define groups to compare. However, with no clear differences in traditional demographic measures that could be used to categorize individuals, such as age, I use Latent Class Analysis (LCA) to define these group from behavioral survey data. The process of discovering categories with unsupervised clustering methods provides a mechanism to quantify the variation in risk perception and behavior, that cannot be directly measured. In Chapter 3, I map the association between these emergent risk groups with infection rates from serological data to parameterize a mechanistic model of infection ***[REF]***, and demonstrate the limits of non-pharmaceutical interventions alone to reduce infections within the student population.

In the second half of my dissertation (Chapters 4 & 5), I examine the necessity and implications of categorizations for action in regions with persistent and emerging infection dynamics. Infectious disease surveillance has 3 primary objectives: to observe and quantify the burden of disease, monitor trends in prevalence, and detect and inform response to outbreaks [1,2]. In pursuit of these goals, numerous continuous values must be discretized. Firstly, cases must be counted, which requires a set of criteria to convert the underlying infection dynamics within an individual into a binary status:

infected or not. This criteria often comes in the form of a diagnostic test, like an enzyme linked immunosorbent assay (ELISA). ELISAs measure the presence and quantity of antibodies in a biological sample that are produced by a person's immune system in response to pathogen exposure, and attempts to discriminate between two hypothetical infection/exposure states ***[REF]***. In practice, no threshold will be able to perfectly discriminate between these groups of individuals, leading to classification errors ***[REF]***. The sensitivity of a test refers to its ability to correctly detect the presence of infection when an infectious individual is tested, also called the true positive rate ***[REF]***. The specificity is the opposite: the ability to correct detect the *lack* of infection in an uninfected individual, also called the true negative rate ***[REF]***. An important third characteristic of diagnostic tests that arises from the discretization of a continuous measure is the positive predictive value (PPV) of a test. The PPV is the probability that a positive test result actually reflects a positive individual ***[REF]***. Unlike the sensitivity, it is not preconditioned on the assumption that the individual tested is truly positive. The complement to the PPV is the negative predictive value (NPV); the probability that a negative test result accurately reflects reality. When counting for infectious disease surveillance, decisions are made on the basis of these imperfect categorizations. In my 4th chapter I explore how fallible diagnostic tests interact with non-target background infections (that change the PPV of test results), producing different time series that are used to detect outbreaks. Additionally, the very notion of an outbreak is itself a categorization of a continuous phenomenon, and attempts to separate a time series of test positive cases by suspected outbreak status will face similar issues of sensitivity/specificity/PPV/NPV. My work demonstrates how uncertainty that arises at each step of the outbreak detection process must be accounted for, highlighting contexts where different combinations of diagnostic tests and outbreak classification criteria can produce equivalent outbreak detection accuracies. In the final chapter, I address how these discontinuity errors affect efforts to build *proactive* rather than *reactive* outbreak alert systems. In contrast to traditional outbreak detection systems that require the observation of test positive cases to trigger an alert i.e., respond to the detection of an ongoing outbreak, proactive alert systems have been developed to predict the risk and potential of future outbreaks. Instead of categorizing incidence to define a prediction target, proactive alert systems calculate summary statistics of test positive time series to predict the approach to the *tipping point* of infectious diseases, $R_{\text{effective}} = 1$. $R_{\text{effective}}$ is the average number of secondary infections each infectious individual is expected to generate before they recover (given the current population size and susceptibility), where values greater than or equal to 1 indicate transmission would be self-sustaining if a population is seeded with infection(s). Predicting the approach to this tipping point would provide advance warning of potential outbreaks, allowing proactive decisions to be made. I show that when imperfect diagnostic tests are utilized to create the underlying summary statistics, much like reactive outbreak detection systems, the alert performance is heavily influenced by the shape and magnitude of the non-target background infections. Addressing the context explicitly when designing a reactive or proactive outbreak surveillance system allows policy-makers to account for the compounding layers of uncertainty, finding zones of equivalence where particular objectives can be given greater prioritization e.g., speed of response vs. the number of false alerts.

When evaluated in its entirety, my dissertation provides a clear and principled approach to evaluating the effects of categorizing continuous infectious disease data. I demonstrate that through acknowledging the imperfect nature of discretization, it is possible to identify meaningfully different clusters of individuals and outcomes that can inform our understanding of the populations most at risk of infection, and how outbreak surveillance systems can be designed to best address context-specific priorities.

Bibliography

[1] Murray J, Cohen A L. Infectious Disease Surveillance. International Encyclopedia of Public Health 2017:222–9. https://doi.org/10.1016/B978-0-12-803678-5.00517-8

[2] World Health Organization. Surveillance in Emergencies 2024