Escherichia coli O157:H7 Phylogenetics: Implications for Virulence and Disease Distribution

Gillian Tarr

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Amanda Phipps, Chair

Jonathan Mayer

Peter Rabinowitz

Program Authorized to Offer Degree:

Epidemiology

University of Washington

Abstract

Escherichia coli O157:H7 Phylogenetics: Implications for Virulence and Disease Distribution

Gillian Tarr

Chair of the Supervisory Committee:

Amanda Phipps, Assistant Professor

Epidemiology

This dissertation uses phylogenetic classifications to investigate heterogeneity in the epidemiology of Shiga toxin-producing *Escherichia coli* (STEC) O157:H7, one of the top causes of foodborne illness and hospitalization. Isolates of a given phylogenetic lineage may express similar traits. Understanding differences between lineages may increase our understanding of STEC O157:H7 incidence and its progression to hemolytic uremic syndrome (HUS). The goals of this study were to: 1) validate HUS case status and compare common definitions of HUS; 2) test the association between phylogenetic lineage and HUS, determine how age affects this association, and quantify the portion of the association due to Shiga toxin genes (*stx*); and 3) determine whether STEC O157:H7 isolates of the same lineage cluster together or are equally distributed with other lineages.

I conducted a retrospective cohort study of all culture-confirmed STEC O157:H7 cases reported to the Washington State Department of Health from 2005 through 2014. Isolates were typed using an established single nucleotide polymorphism (SNP) assay and grouped into phylogenetic lineages. Lineage Ib was the most common, followed by IIa and IIb. The remaining lineages were grouped in a “rare” category. I abstracted medical records of all hospitalized cases

to validate HUS status using stringent clinical criteria. Compared to other common definitions of HUS, these stringent criteria best identified cases with severe disease as evidenced by dialysis. The definition used for public health reporting performed poorly when judged against the stringent criteria, overestimating HUS burden by twofold.

Using the validated HUS outcome, I assessed the association between phylogenetic lineage and HUS using generalized estimating equations to account for lack of independence among STEC O157:H7 isolates with the same pulsed field gel electrophoresis profile. In unadjusted analysis, lineage IIb was associated with a significantly higher odds of HUS than lineage Ib [odds ratio (OR) =1.65; 95% confidence interval (CI) 1.05, 2.60]. However, when assessing effect modification of the OR by age, both lineage IIa and IIb were associated with higher odds of HUS among adults 20-59 but not among children <10 years-old. Associations between lineages IIa or IIb and HUS appeared to be mediated by the *stx2a* genotype.

The concept of spatial segregation was used to assess the distribution of STEC O157:H7 cases by phylogenetic lineage. Using a kernel estimation method, statistically significant spatial segregation was detected (p=0.001), with foci of segregation among lineage IIb in the southwest region of the state and among lineage IIa in the south-central region. A generalized additive model adjusted for age and sex identified increased risk of lineage IIb infections in the southwest, consistent with the spatial segregation detected. Two additional methods confirmed the results. In exploratory analysis, I identified multiple risk factors potentially associated with infection by particular STEC O157:H7 lineages.

Distinguishing lineages of the STEC O157:H7 serotype can provide insight into its maintenance, transmission, and virulence. The heterogeneity between lineages provides an opportunity to target interventions, both clinical and public health, to the specific form of the pathogen dominant in an area. This dissertation furthers our understanding of STEC O157:H7, as well as uncovering multiple questions that should be addressed to effectively limit the incidence and impact of this disease.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This dissertation would not have been possible without the collaborative efforts of the Washington State Department of Health and Washington State University. Hanna Oltean and the team in the Office of Communicable Disease Epidemiology at DOH generously provided space and resources to conduct the HUS review, as well as provided case data and clinical isolates. Tom Besser and Smriti Shringi at WSU completed the SNP and SBI typing to provide phylogenetic lineage and Shiga toxin gene classifications. This project would not exist without these immense contributions.

I owe a large debt of gratitude to my dissertation committee. Peter Rabinowitz gave me an academic home in the Center for One Health Research, providing support, mentoring, and resources. It was on a tour of DOH with Peter that I learned about this project and decided to make it my dissertation. In addition to isolate typing, Tom Besser provided the initial idea for Chapter 3, wanting to investigate the association between phylogenetic lineage and HUS, and in doing so spurred the entire project. Jonathan Mayer supported my efforts from the start, providing valuable feedback in the early stages of the project that helped me make it what it is. Jon Wakefield was instrumental in developing the sampling scheme for isolate testing and identifying appropriate spatial analysis methods. Phil Tarr gave selflessly of his time to this project, reviewing drafts thoroughly and quickly. He has helped shape my thinking on STEC, providing immeasurable insight into the field and directing me toward important areas of research, including the verification of HUS status in Chapter 2. I am also incredibly thankful for the networking he has assisted me in doing. Finally, I owe a huge thank you to my chair Amanda Phipps. She has been a constant source of support through this process, providing feedback and guidance on all aims. I greatly appreciate her willingness to step outside her subject area to share her methodological expertise with my project.

I would also like to thank all the sources of funding for this project. Tom Besser's grants from the U.S. Department of Agriculture National Institute of Food and Agriculture, #2009-

04248 and #2010-04487, made the isolate typing possible. I am immensely grateful to Lianne Sheppard for funding me on the Biostatistics, Epidemiology, and Bioinformatics Training in Environmental Health training grant, funded by the National Institutes of Health National Institute of Environmental Health Sciences (T32ES015459). Lianne also guided me through the process of applying for an F31, which funded the final quarters of my program (NIH National Institute of Allergy and Infectious Disease F31AI126834).

I am also grateful to all of the hospitals that responded to my request for records during the HUS review. I would like to thank the local health jurisdiction staff who initially investigated the cases, completing case report forms that became the basis of my study data.

Finally, I would like to thank the family and friends who supported and encouraged me through this process. I would particularly like to thank my mom, dad, and aunt for providing childcare, often with little notice, so that I could attend a meeting or work on a paper. Thank you to my husband for reminding me frequently that this is all worth it, and to my daughters for giving me two more reasons to be a good role model and follow my dreams.

For Lois and Ada, may you never hesitate to live the life you want.

Chapter 1. Introduction

Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 is a leading foodborne pathogen, estimated to cause more than 90,000 illnesses, 3,000 hospitalizations, and costs exceeding $270 million each year in the United States (1, 2). STEC was first recognized for its link to hemolytic uremic syndrome (HUS) in 1983 (3). The Shiga toxin (Stx) released by STEC enters cells and interrupts protein synthesis, potentially triggering apoptosis (4). If Stx reaches the blood stream, it can lead to HUS, defined by hemolytic anemia, thrombocytopenia, and renal failure (4). STEC O157:H7 is the most common cause of HUS. HUS occurs in ~15% of children under the age of 10 infected with STEC O157:H7 and has 3-5% case fatality (5, 6).

Following a high-profile outbreak associated with a fast food hamburger chain in 1993 (7), STEC O157:H7 became reportable to the Centers for Disease Control and Prevention (CDC) in 1995. Incidence in 1997 was 2.1 per 100,000, and Healthy People 2010 set a goal of halving incidence, which was accomplished by 2009 (8). Worldwide, declines in incidence have been attributed to improvements in food safety, pre-market food inspection, and outbreak investigation (9-11). The Healthy People 2020 goal is to halve STEC O157:H7 incidence from 1.2 per 100,000 in 2006-2008 to 0.60 per 100,000 by 2020 (12). National reporting was extended to include other STEC serotypes in 2001. The most recent national estimate for all STEC serotypes combined was 1.94 per 100,000 in 2014 (13). Estimates from the Foodborne Diseases Active Surveillance Network (FoodNet) for 2016 put STEC incidence at 2.84 per 100,000 (14). Increased recognition and testing are suspected to have caused an increase in the reported incidence of STEC in recent years, while the incidence of serotypes other than O157 has likely declined similarly O157 incidence (15, 16). STEC O157:H7 remains the most common serotype, accounting for 36% or 1.0 per 100,000 of the STEC cases ascertained by FoodNet. However, its incidence has remained stable or shown non-significant declines in recent years (14, 17).

With the current trajectory, the United States is not on target to meet the Healthy People 2020 goal for STEC O157:H7, and new approaches are needed to decrease incidence and prevent

severe outcomes such as HUS. The STEC O157:H7 serotype shows considerable diversity. More than 5000 distinct pulsed field gel electrophoresis (PFGE) STEC O157:H7 profiles have been detected, with the 100 most common strains accounting for over 50% of reported cases (CDC, unpublished data). Individual STEC O157:H7 strains can vary substantially in their virulence, with some resulting in no cases of HUS and others leading to HUS in >20% of cases (18, 19). Better understanding of the genetic characteristics responsible for variation in frequency and severity by strain may provide opportunities for intervening in disease transmission or disease progression and ultimately meeting public health goals.

The STEC O157:H7 phylogenetic tree summarizes evolutionary relationships among observed strains of the bacteria. Strains on a given branch of the tree share important genetic markers and may share traits, such as high virulence. In 2008, Manning and colleagues (20) demonstrated increased odds of HUS associated with one particular phylogenetic grouping, termed clade 8. Other studies have now shown similar findings (21, 22). At the time of the 2008 study by Manning et al., few STEC O157:H7 strains had been fully genotyped and there was not yet a substantial list of single nucleotide polymorphisms (SNPs) that could be used as evolutionary markers for phylogenetic analyses. Further definition of the tree was necessary (23, 24). Since then, many more STEC O157:H7 strains have been sequenced (25, 26), and an alternative phylogenetic tree has been developed for STEC O157:H7 (23, 27). The updated tree (Figure 1.1) uses SNPs systematically selected from conserved regions of the STEC O157:H7 genome and adds strains isolated from cattle, STEC O157:H7's primary reservoir.

Studies that combine STEC phylogenetics with epidemiology are a natural and critical step in understanding pathogen heterogeneity, both in terms of virulence and distribution. Phylogenetic signals associated with differences in disease manifestation may highlight virulence genes beyond those already recognized and may open exploration into novel treatments. Heterogeneity can also be used to elucidate mechanisms behind STEC O157:H7 epidemiology.

In this dissertation, I leverage differences in the distribution of phylogenetic lineages across person, space, and time to improve our understanding of STEC O157:H7 maintenance, transmission, and virulence (Figure 1.2). I conducted a population-based retrospective cohort study of all culture-confirmed STEC O157:H7 cases reported to the Washington State Department of Health (DOH) from 2005 through 2014.

As described in Chapter 2, I validated HUS status for hospitalized members of the cohort. Washington State does not report HUS to the CDC or otherwise systematically track HUS incidence. To standardize HUS status classification and to inform DOH about HUS reporting in the state, I reviewed medical records for hospitalized HUS cases and classified them using objective laboratory criteria. I compared multiple definitions of HUS and demonstrated the implications of misclassification on disease surveillance and clinical research.

In Chapter 3, I used the validated HUS status to investigate the association between phylogenetic lineage and HUS (Figure 1.2). Contrary to expectation, we found that this association was nullified after adjustment for age. Age is the strongest predictor of progression to HUS, but its relationship with lineage is unclear. I found that lineages differ in their age distribution, prompting questions about differences in exposure or early disease course. Age also modified the association between lineage and HUS, indicating that more work is warranted to understand how mechanisms between lineage and HUS differ by age. Finally, I tested mediation of the lineage-HUS association by Stx gene (*stx*) composition. Although the correlation between lineages and *stx* genotypes is very close, I detected mediation that may account for much of the signal associated with lineage.

Chapter 4 moves backward in the natural history of disease to assess spatial patterns of STEC O157:H7 by phylogenetic lineage (Figure 2.1). I identified spatial segregation of lineages, which argues against widespread dispersal of strains, such as one would expect to occur through national food distribution networks. The segregation suggests that local reservoirs of STEC O157:H7 may exist. Alternatively, lineages may vary in their modes of transmission and some

modes of transmission may be more common in some regions than others. The potential for preferential exposures is intriguing and may play a role in persistent reservoirs or repeated introduction. I identified STEC O157:H7 risk factors more strongly associated with certain lineages than others in an exploratory analysis.

Bacterial phylogenetics provide a method of dissecting the epidemiology of STEC O157:H7 to understand how and why certain associations exist. In this study, I have identified an HUS definition that accurately captures the most severe disease without overinflating disease burden, elucidated roles for age and *stx* genes in the association between phylogenetic lineage and HUS, and discovered spatial segregation of STEC O157:H7 lineages, which provides clues as to how maintenance and transmission may be occurring in local communities.

Figures

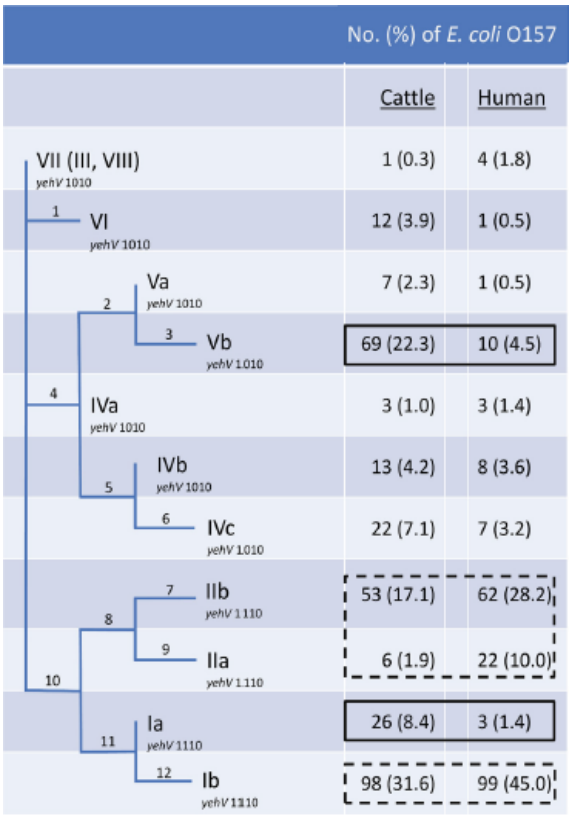

Figure 1.1. Phylogenetic tree of Escherichia coli O157:H7. To the right are the number of strains in each lineage from cattle and humans. From Jung et al. (27).

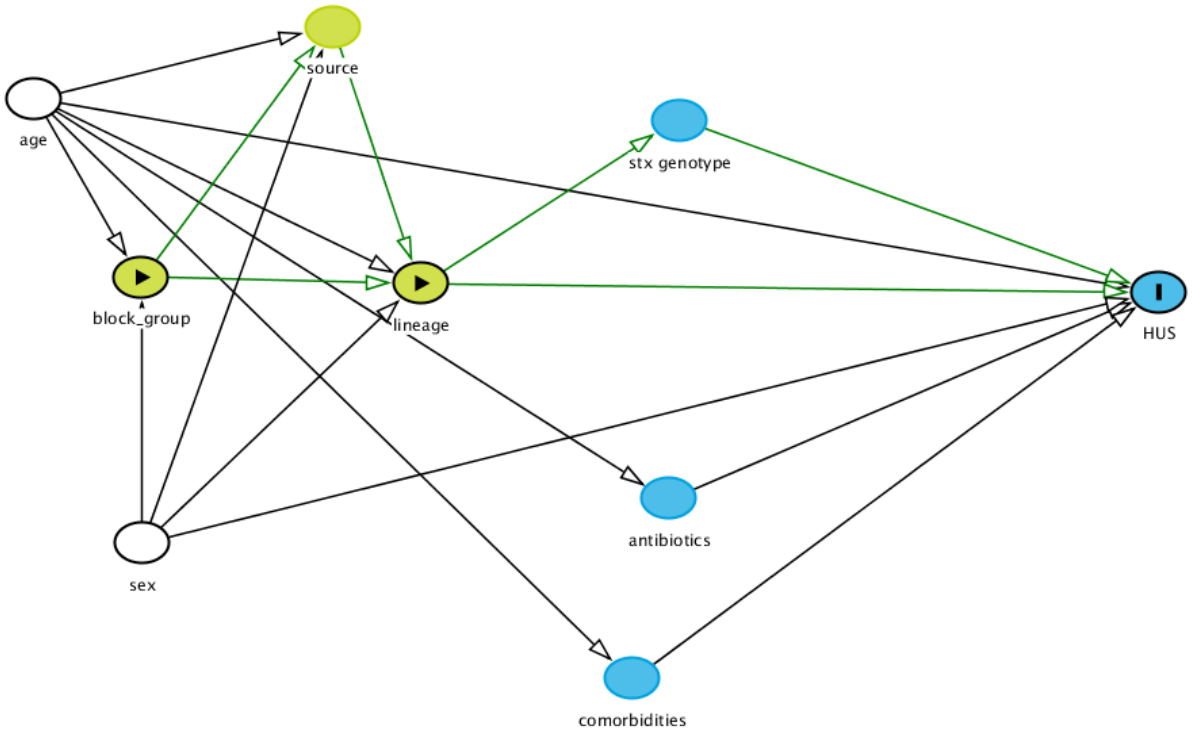

Figure 1.2. Directed acyclic graph (DAG) of associations under study.

Chapter 2. Defining Hemolytic Uremic Syndrome Following *Escherichia coli* O157:H7 Infection: Implications for Public Health Planning and Clinical Epidemiology

Gillian A. M. Tarr[1,2], Hanna N. Oltean[3], Peter Rabinowitz[2,4], Amanda I. Phipps[1], Phillip I. Tarr[5]

[1]Department of Epidemiology, University of Washington

[2]Center for One Health Research, University of Washington

[3]Washington State Department of Health

[4]Department of Environmental and Occupational Health Sciences, University of Washington

[5]Department of Pediatrics, Washington University in St. Louis School of Medicine

Introduction

Hemolytic uremic syndrome (HUS) is characterized by hemolytic anemia, thrombocytopenia, and renal injury, which often necessitates dialysis (28). HUS has a case fatality of 3-5% (6, 29, 30). HUS incidence is highest among children <5 years old (13, 14, 31, 32). Shiga toxin-producing *Escherichia coli* (STEC) O157:H7 is the predominant cause of postdiarrheal HUS worldwide. In 2014, 250 HUS cases, or 0.08 cases per 100,000, were reported to the Centers for Disease Control and Prevention (CDC) in all age groups (13). Active surveillance of HUS in children <18 years old through the Foodborne Diseases Active Surveillance Network (FoodNet) identified 0.53 cases per 100,000 in this age group in 2014 (17).

Public health planning and clinical epidemiologic studies have employed diverse HUS case definitions. All generally include criteria of anemia (diminished hemoglobin concentrations or low hematocrit, and/or evidence of hemolysis), thrombocytopenia, and an index of impaired renal function (e.g., elevated serum creatinine concentrations or oliguria). In the context of initial public health investigations, sensitivity of the working case definition for HUS is often prioritized over specificity in order to avoid missing cases. However, to plan public health responses, a highly accurate case definition (i.e. high sensitivity and high specificity) is desirable.

Specific components of HUS definitions have raised concerns in the literature, such as overdiagnosis from using urinalysis results to establish renal injury (33-35) and true cases being excluded because hemoconcentration prevents them from meeting anemia criteria (36, 37). We therefore evaluated HUS definitions used for public health reporting, epidemiologic research, and clinical practice, with the goal of informing public health regarding data collection improvements, current barriers to accurate reporting, and education needs among clinicians diagnosing HUS cases. We demonstrate the importance of HUS definition to tracking disease through time. We also show the impact of misclassifying HUS in the context of a significant

public health and clinical question: the effect of using antibiotics to treat STEC O157:H7 infections.

Methods

Case Ascertainment & Record Abstraction

We conducted a retrospective review of all hospitalized, culture-confirmed *E. coli* O157:H7 cases reported to the Washington State Department of Health (DOH) between 2005 and 2014 through passive surveillance. Hospitalization status and hospitals where treated were noted for each case. Cases were considered hospitalized if admitted at least overnight to an inpatient medical facility. A standardized DOH case report form (CRF) was used by local health jurisdictions.

We obtained records from each hospital listed on the CRFs (Appendix). We reviewed these records and abstracted the following information: demographics; dates of onset, hospitalization, and stool specimen collection; signs and symptoms; laboratory values for platelet count, hematocrit, and serum creatinine; evidence of microangiopathic changes; urine output rate; receipt of dialysis; antibiotic use (drug, date initiated, and duration); and diagnoses. If urine output rate per hour was not noted explicitly in the chart, we calculated it as urine output documented over a 24 hour period per hour per kilogram of admission weight.

All case data were de-identified before analysis. This review was deemed exempt by the Washington State Institutional Review Board.

HUS Definitions

Six primary definitions of HUS were considered in this review (Table 2.1). The first, hereafter referred to as the stringent clinical definition (SCD), has been used in publications of HUS case series from the Pacific Northwest (38-41), elsewhere in North America (42), and in

Europe (43, 44). The Council of State and Territorial Epidemiologists (CSTE) criteria for postdiarrheal HUS, established in 1996 and reaffirmed in 2009, are used as the standard for reporting cases to CDC (45). Combined CSTE confirmed and probable cases are used in HUS surveillance, and so this combination was used in this review. An alternative definition from the literature is similar to the CSTE definition, but with the addition of thrombocytopenia and without allowing hematuria or proteinuria as sufficient evidence of renal injury. This definition, hereafter referred to as the hematology-focused confirmed definition (HCD), has been used in multiple studies of FoodNet data (30, 46, 47). We also considered HUS designation on the CRF and diagnosis indicated in the discharge note or charge codes of the hospital chart. Finally, to mirror surveillance systems such as FoodNet (17), which ascertain all cases typically reported (i.e. using the CSTE confirmed and probable definitions) and all cases clinically diagnosed as HUS, we combined the CSTE and hospital diagnosis case groups in a sixth definition.

To align with previous work, we used age-specific normal values for serum creatinine concentrations from Meites (48) for the SCD and hematocrit normal values for age from the Harriet Lane Handbook (49) to define anemia for the CSTE and hematology-focused confirmed definitions (HCD).

Most Favorable Definition for Identifying Severe Disease

HUS definitions were evaluated on their ability to accurately ascertain cases requiring dialysis, one of the most common severe outcomes of HUS. Sensitivity to dialysis was calculated as the number of dialysis cases identified as HUS by the definition over the number of cases receiving dialysis as a consequence of HUS. The ideal definition is expected to ascertain all cases needing dialysis as a consequence of their acute episode of HUS. However, because not all HUS cases require dialysis, specificity for dialysis was not appropriate for assessing definition performance. Instead, the proportional incidence of dialysis was calculated for each definition and compared to that reported in the literature. The definition with the best combination of

sensitivity to dialysis and correspondence to published incidence of dialysis was used as the comparator for evaluation of sensitivity and specificity in the other definitions.

HUS Definition Validity

Using the definition that best identified cases needing dialysis as the standard, we calculated sensitivity and specificity of the different case definitions for all cases and for cases stratified by age (<10 years-old) and antibiotic use. We conducted sensitivity analyses to determine how changes in various definition components affected ascertainment. For the SCD, we tested the impact of age-specific normal serum creatinine concentrations from the Harriet Lane Handbook (49) and of relaxing the requirement that all laboratory criteria be met on the same day. For the CSTE definition, we tested the impact of inclusion of thrombocytopenia criteria. For the HCD, we tested the performance of the definition without criteria for evidence of microangiopathic changes.

Impact of HUS Definition on Disease Trends and Association with Antibiotic Use

To understand how HUS definition affects the ability to capture trends in disease over time, the annual incidence rate of HUS across all age groups was calculated (50). Trends in incidence as determined by each definition were graphed.

To understand how HUS definition affects data needed for public health recommendations and clinical practice, we tested the impact of using different HUS case definitions on the estimated association between antibiotic use and HUS. Antibiotic use was determined from notes, medication administration records, or charges in the hospital record, and could include antibiotics received prior to the admission (e.g. from an outpatient care provider) or during the hospital stay. For those cases deemed HUS by the SCD, antibiotics must

have been initiated prior to meeting SCD criteria for HUS. To account for differing time at risk of receiving antibiotics by HUS status, the nadir of illness (defined as the day on which anemia, thrombocytopenia, and azotemia were at their most severe) was used as a reference date for those who did not meet the SCD criteria. If this was not a single day (e.g. the patient had low hematocrit and platelets but creatinine was not elevated until later in the illness), the day with the most severe combination of criteria (e.g. the day with the lowest hematocrit and platelet count) was chosen as the reference date.

We identified appropriate confounders for adjustment from previous studies of the association between antibiotic use and HUS. We adjusted analyses for age (continuous) (38, 51), sex (52), vomiting (38, 52), bloody stool (52), and days from diarrhea onset to stool specimen collection (38). White blood cell count was not used, as we did not consistently have these values early in illness. We used logistic regression to test the association between antibiotic use and HUS as determined by each definition, adjusted for the above covariates. R (53) was used for all analyses.

Results

Of 1160 culture-confirmed STEC O157:H7 cases in Washington State, 471 (41%) were hospitalized (Figure 2.1). Hospital records were obtained and abstracted for 433 cases: no hospital was listed on the CRF for 18 cases, and records for 20 additional cases could not be located at the hospital listed (Appendix).

Of the 433 cases whose records we reviewed, 429 had sufficient data to determine HUS status using laboratory results, and 164 fulfilled one or more case definitions of HUS. Individual definitions classified 58 to 160 cases as having HUS (Table 2.2). Using the SCD, the overall average annual incidence of HUS was 0.11 per 100,000, with a range of 0.09 to 0.24 per 100,000 for the other definitions. Among children <18 years of age, average annual incidence of

HUS was 0.41 per 100,000 based on the SCD, with a range of 0.31 to 0.63 per 100,000 using the other definitions.

The frequencies of bloody diarrhea and vomiting were similar across definitions (Table 2.2). Indicators of disease severity were length of hospitalization, urine output, and dialysis. Length of stay for all reviewed STEC O157:H7 cases ranged from 1 to 68 days and varied by case definition. Median (IQR) stays among HUS cases ranged from 7 (4, 14) days for CSTE-defined cases to 13.5 (10, 21) days for cases classified according to the HCD. Of the 164 cases classified as having HUS by at least one definition, 30 (18%) were anuric (Table 2.2).

Five cases died during the study period (Appendix, Table S1). Case fatality was 1.2% among hospitalized STEC O157:H7 cases and 3.9% among SCD-defined HUS cases. One fatal case was not defined as HUS by any definition, and two were HUS cases by all definitions.

Determining Comparator Definition

Of the 396 reviewed cases with sufficient information to determine if they received dialysis, 41 (10%) received dialysis during their STEC O157:H7 hospitalization, indicating severe HUS (Table 2.2). Two dialyzed cases had pre-existing chronic or end-stage renal disease and were considered to have received dialysis for a cause other than their infection. Four definitions identified 100% of the 39 cases receiving dialysis as a consequence of their STEC O157:H7 infection: SCD, CSTE, hospital diagnosis, and combined.

Approximately 27-28% of cases defined according to the CSTE or combined definitions required dialysis (Table 2.2), which is considerably lower than average estimates of 50-60% provided by prior studies (10, 30, 54). Fifty-two percent of SCD-defined HUS cases received dialysis, which was the closest proportional incidence of dialysis to prior studies. Therefore, we used the SCD to assess the sensitivity and specificity of the other HUS definitions.

Definition Validity

Relative to the SCD, the combined definition had the highest sensitivity at 99% (95% CI 93%, 100%) (Table 2.3). The HCD had the lowest sensitivity, identifying 74% of HUS cases (95% CI 62%, 83%). Specificity relative to the SCD ranged from 76% (95% CI 71%, 80%) for the combined definition to 99% (95% CI 98%, 100%) for the HCD.

Factors affecting SCD disagreement with the CRF and hospital diagnosis HUS designations could not be evaluated, because these criteria are subjective. Where the SCD and the CSTE or HCD disagreed, we examined component criteria to determine source(s) of the discordance (Appendix). Low specificity of the CSTE definition appeared to be driven by lack of criteria for thrombocytopenia and inclusion of hematuria or proteinuria as evidence for kidney injury (Appendix, Table S2). Low sensitivity of the HCD appeared to be driven by the requirement for evidence of microangiopathic changes and crude criteria for serum creatinine concentration.

For all definitions but the CSTE and combined, sensitivity was higher and specificity was lower among children <10 years of age than among cases 10 and older (Table 2.3). This reflects greater ascertainment of HUS in children, both true positives and false positives. For the CSTE and combined definitions, both sensitivity and specificity were higher in children <10, suggesting that the CSTE criteria are more accurate in young children than in older children and adults.

HUS case identification also varied by antibiotic use. All five definitions had poorer sensitivity among patients who received antibiotics compared to those who did not (Table 2.3). All definitions were better able to correctly exclude non-HUS cases among antibiotic users than non-users. The greatest difference in specificity by antibiotic use was observed in hospital-

diagnosed cases and suggests that hospital providers are more likely to over-diagnose HUS in patients who have not received antibiotics.

Sensitivity Analysis

Using creatinine normal values from the Harriet Lane Handbook (49) instead of Meites (48) changed the classification of three cases, two from HUS to non-HUS and one from non-HUS to HUS (Appendix). Relaxing the requirement that all criteria be met on the same day resulted in the inclusion of two additional cases. Neither of these modifications to the definition appreciably altered the HUS case group or ascertainment of dialysis cases (Appendix, Table S3).

When excluding cases without thrombocytopenia, the modified CSTE definition classified 102 STEC O157:H7 cases as HUS (Appendix, Table S3). Without the requirement for microangiopathic changes, the modified HCD classified 69 cases as HUS. Thus, both modifications moved their respective definitions closer to the SCD case count. The modified CSTE definition identified all 39 cases dialyzed as a result of their STEC O157:H7 infection, and the modified HCD identified 38. Reflecting the nature of the changes, the modified CSTE definition exhibited improved specificity (92%; 95% CI 88%, 94%) and the modified HCD exhibited improved sensitivity (87%; 95% CI 77%, 94%) (Appendix, Table S4).

Temporal Trends of HUS

Incidence trends were similar across most definitions (Figure 2.2). The CSTE and, by extension, combined definitions estimated substantially higher incidence in all years. Most trends, such as the surge of cases in 2013, were detected by all six definitions. However, the definitions were not always consistent. In 2009, the CSTE and hospital diagnosis classifications indicate an increase in incidence while the SCD and CRF indicate a decrease and the HCD detected no change.

Antibiotic Use and HUS

Of those STEC O157:H7 cases who had complete data for adjusted analysis of the association between antibiotic use and HUS, antibiotics were administered to 144 patients prior to their onset of HUS or their reference date (Figure 2.1). Twenty (14%) developed HUS according to the SCD. Among the 212 who did not receive antibiotics, 45 (21%) developed HUS. Despite a crude OR of 0.60 based on the SCD, adjustment for *a priori* confounders revealed an elevated risk of HUS among those given antibiotics (Table 2.4): when the SCD of HUS was applied, the adjusted odds of HUS were 2.48 times higher among STEC O157:H7 cases receiving antibiotics than those who did not receive antibiotics (95% CI 1.15, 5.41). Estimates based on all other definitions were attenuated, ranging from OR=1.08 for the combined definition to OR=2.18 for hospital diagnosis (Table 2.4). Published knowledge about the negative effects of antibiotics on STEC O157:H7 disease progression increased during the study period, but the proportion of cases receiving antibiotics remained at approximately 40% (Figure 2.3). On average, 11% of children <10 years-old hospitalized with reported STEC O157:H7 infections received antibiotics each year, compared to 60% of older children and adults.

Discussion

We found that the SCD best identified STEC O157:H7 cases with severe HUS manifestations, as evidenced by the need for dialysis. Of the six primary definitions examined, the SCD identified all dialyzed patients without preexisting chronic or end-stage renal disease, without inflating the case count with patients lacking anemia, thrombocytopenia, and/or azotemia. While not a gold standard, the SCD's correspondence with expectations for dialysis prompted us to use it as the comparator for evaluating the other five definitions.

HUS incidence according to the SCD (0.11 cases per 100,000) was slightly above the average annual incidence of postdiarrheal HUS of all etiologies reported to CDC during the study period (0.09 cases per 100,000). Although the incidence in our study might be expected to be lower than that observed nationally because it does not include HUS due to non-STEC O157:H7 pathogens, the higher incidence is not surprising given that Washington's reported STEC incidence is twice the national average (13). The incidence among children <18 years-old in our study was 0.41 per 100,000, a substantial decline from studies in urban and suburban Seattle, Washington, in the 1970s and 1980s (55, 56).

HUS status reported on the CRF and hospital diagnosis were both subjective assignments of disease. With no current standard for HUS reporting in Washington State, we cannot hypothesize about differences between the SCD and CRF designations. However, our results demonstrate the danger of relying on an unvalidated and non-standardized classification. The rationale for a hospital diagnosis is not always apparent when examining objective criteria in medical records. Compared to the SCD, using hospital diagnosis to define HUS status yielded good sensitivity, missing only three cases. However, the positive predictive value (PPV) of hospital diagnosis was only 79%, meaning that 1 in 5 individuals diagnosed with HUS likely does not have it according to objective criteria.

The CSTE and combined definitions had reasonable sensitivity but poor specificity. The CSTE definition PPV was only 47%, meaning that over half of the cases classified as HUS by the CSTE definition do not have HUS per the SCD. Correspondingly, we see a two-fold increase in incidence with the CSTE and combined definitions relative to the SCD definition. The HCD faces the converse limitation. It is so restrictive that 20 cases, including six receiving dialysis, were not considered HUS cases. With a PPV of 97%, however, almost all HUS cases identified by the HCD were true cases per the SCD.

The four domains in which the SCD, CSTE, and HCD diverged (i.e., thrombocytopenia, microangiopathic changes, serum creatinine concentrations, and use of urinalysis criteria) each

warrant comment. The CSTE statement on postdiarrheal HUS (45) notes, "If a platelet count obtained within 7 days after onset of the acute gastrointestinal illness is not less than 150,000/mm3, other diagnoses should be considered." However, thrombocytopenia is not included in the matrix of criteria for HUS. As evidenced by a recent review that catalogued HUS definitions as part of its methods (57), thrombocytopenia criteria are common. Our results show that they are indeed critical to the differentiation of HUS and non-HUS cases. When thrombocytopenia was added to the CSTE definition in sensitivity analysis, 52 cases were removed from the HUS count, and specificity of the modified definition reached 92%, comparable to hospital diagnosis.

Criteria requiring microangiopathic changes on smear examination may be overly restrictive, as almost half of the STEC O157:H7 patients in our review did not have peripheral blood smear data in their hospital chart. Smears may also appear normal if done only early in illness before evidence of injury to erythrocytes appears. In our study, 12 of the HUS cases missed by the HCD lacked evidence of microangiopathic changes; in two, smears were interpreted as normal and in ten, there were no documented smear results. Even among patients receiving dialysis and diagnosed with HUS, a peripheral blood smear was not documented for four patients, and was reported as normal when performed in a fifth. Evidence of intravascular erythrocyte destruction hemolysis is ideal, but, in reality, case management may not require this information if other clinical and laboratory elements are consistent with STEC-associated HUS. When the HCD was modified in sensitivity analysis to not require microangiopathic changes, sensitivity increased without compromising specificity.

The use of rigid creatinine criteria in the HCD, grouping cases into only two age groups, resulted in suboptimal sensitivity for even the modified HCD when compared to the age-specific criteria of the SCD, which groups cases into five age groups based on Meites (48). The nine SCD-classified HUS cases who were excluded from the modified HCD based on these criteria included one child who received dialysis and eight with evidence of hemolysis. While our

findings support the use of age-specific serum creatinine concentrations as an important component of the HUS case definition, we acknowledge problems with this determination. First, the degree of renal injury in STEC O157:H7 infections is likely on a spectrum, and employment of a rigid cut-point for categorical definition purposes is somewhat arbitrary. For example, there is evidence of renal tubular injury in infected children who do not develop HUS (58). Also, as a consequence of illness, many infected children have had poor protein intake for several days, and creatinine values might be misleadingly low. Under such a scenario, a normal value might actually reflect some degree of renal dysfunction.

The CSTE definition uses the same rigid creatinine serum concentrations as the HCD but accepts hematuria or proteinuria in lieu of elevated creatinine. This resolves much of the sensitivity problem but diminishes specificity. Urines may be contaminated in patients with diarrhea and urinalysis therefore might not accurately reflect kidney injury (33-35). Moreover, hematuria, based on dip-criteria, could reflect filtered serum free hemoglobin secondary to intravascular erythrocyte destruction, and not red cells of kidney origin, thus not necessarily a marker of kidney injury. Because of these factors, classifying patients as having HUS based on hematuria or proteinuria when their serum creatinine concentrations are normal inflates the HUS burden.

Choice of HUS definition has implications for patients, public health practitioners, and clinicians relying on HUS research. One implication of low specificity is the potential for adverse future consequences for patients who are classified as having had HUS even though they had no azotemia. Specifically, a diagnosis of HUS infers lifelong risk for chronic renal disease, most particularly if the HUS was accompanied by anuria (28). Non-stringent HUS definitions might lead to unwarranted concerns and subsequent life and health insurance challenges for patients and their families. Such issues arise in genetic testing, where inappropriate predictive value is inferred from finding a risk locus (59-61). Given that HUS definitions based on urinalysis-dependent criteria offer no value in acute illness management in the absence of azotemia, lack of

specificity introduced by these criteria, and potential adverse consequences on individuals of an inappropriate diagnosis, we believe that age-specific serum creatinine criteria provide a sensitive and specific means of resolving these limitations.

There is considerable value in accurate HUS diagnosis for public health practitioners. HUS surveillance complements specific pathogen surveillance. First, HUS is a syndrome that is highly unlikely to be overlooked. In contrast, microbiologic-dependent diagnosis relies on appropriate testing and, to a variable extent, pathogen isolation. Given that ~15% of STEC O157:H7 cases in children <10 years-old progress to HUS (5), there is a ratio of approximately 1 case of HUS for every 7 cases of culture-confirmed STEC O157:H7 infections in children. Hence, by extrapolation, reasonably credible estimates of annual STEC O157:H7 incidence, and of outbreak magnitude, can be made. Second, if a pathogen causes a diffuse outbreak, as in STEC O157:H7 infections that are transmitted via widely disseminated food vehicles, the geographically dispersed sites where patients present for care can hinder provider perception of a cluster. However, HUS cases are generally clustered in highly specialized pediatric facilities, offering another opportunity for outbreak detection. Third, even though STEC O157:H7 retains its role as the predominant cause of HUS worldwide, non-O157 STEC can cause HUS, and it is critical to remain vigilant to the potential emergence of these agents. While the preponderance of non-O157:H7 STEC infections do not culminate in HUS, it is important to ascertain the relative proportions of HUS that can be attributed to these agents, so as to learn more about their epidemiology and pathogenesis. Surveillance of individual HUS cases can prompt heightened attempts to recover pathogens for much-needed analyses, and epidemiologic and environmental investigations.

Given the value of tracking HUS, inaccurate estimates of the HUS burden could have implications for public health efforts. We found that some definitions consistently over- or underestimated HUS burden, and the difference in incidence between definitions was more than three-fold in some years. Choice of HUS definition also affects the ability to detect fluctuations

in HUS incidence, such as the variation in trends in 2009. Between the peaks of 2010 and 2013, incidence increased only 14% by the HCD but more than 50% by the SCD, CSTE definition, and hospital diagnosis. These results show that the case definition is important to not only accurately estimating and planning for the burden of HUS but also in detecting secular changes.

Our data also have implications for interpreting biomedical studies for clinical decisions. Notably, the SCD generated an OR of 2.48 for developing HUS following antibiotic use, similar to that calculated in a recent meta-analysis, which estimated an OR of 2.24 (57). However, this measure was attenuated to as low as 1.08 when other case definitions were applied, a diminished risk also noted in the meta-analysis. Consistent with the direction of the bias, sensitivity was lower and specificity higher among cases who had received antibiotics. Our results demonstrate the impact that misclassification of HUS status, which in this case was differential, may have on assessing clinical interventions.

With data from only one state, one limitation of this analysis is the small sample size. This is particularly relevant to the comparison of temporal trends, which reflect case counts as small as four in some years based on the HCD. However, we observed a similar magnitude of difference between incidence estimates for the HCD and hospital diagnosis as observed in a study by Ong et al. that compared similar definitions applied to FoodNet cases (47).

Data quality also varied substantially across the more than 70 hospitals involved in the review. This resulted in missing information for multiple cases but also reflects a real-world assessment of definition performance. To be accurate across multiple settings, a definition should not rely on clinical practices that vary considerably across facilities, such as that we observed in the performance of peripheral blood smears. The association between antibiotic use and HUS was the analysis most likely biased by missing data. We used a complete case approach, which limited us to 356 cases. However, only 16 of the 73 cases excluded for missing data were missing antibiotic use, so we were able to compare the crude association from the 356 cases analyzed to that observed in 416 cases. In this larger sample, 13% of antibiotic users and

21% of non-users progressed to HUS, compared to 14% and 21%, respectively, in the smaller sample.

Concerns have been raised about reliance on anemia as a criterion with which to define HUS, because patients can present with hemoconcentration while other criteria are met (36, 37). In our study, many cases presented with apparent hemoconcentration, with high hematocrits that dropped precipitously during the first few days of hospitalization. Those who met the SCD had been hospitalized for an average two and a half days when HUS criteria were satisfied. Without criteria for anemia, two additional cases, both with azotemia, would have been defined as HUS. One, admitted to the hospital on the tenth day of diarrhea, had a smear containing schistocytes and was diagnosed by hospital providers as having HUS. The other had no HUS diagnosis. Neither required dialysis. The initial presentation of hemoconcentration is an important consideration for clinical practitioners but does not appear to warrant modification of public health reporting criteria.

There are often tradeoffs in implementing more exact case definitions. Our review indicates that the CSTE definition, the standard for public health reporting, may overestimate the HUS burden by more than two-fold. At the same time, the HCD, common in the literature, underestimates the burden by a sizeable portion and excludes some of the most severe cases. Modifying criteria for thrombocytopenia and microangiopathic changes improves these definitions, and removing hematuria/proteinuria criteria and replacing crude serum creatinine cutoffs with age-specific values would improve them further. For states, such as Washington, considering the adoption of a definition for standardized HUS tracking, the SCD provides the best ascertainment of severe cases while minimizing inflation of HUS burden. For agencies in which age-specific serum creatinine concentrations are not available on historical cases, the modified HCD provides a reasonable alternative, though likely an underestimate of the true HUS burden.

In summary, tracking postdiarrheal HUS enables public health departments to monitor the severity of STEC infections in their regions and to dedicate necessary resources. Differences in HUS definitions may incorrectly estimate the burden of disease and mask fluctuations in incidence, compromising public health efforts. Additionally, clinical research involving HUS, such as assessments of the effect of antibiotics on disease progression, may be subject to bias due to misclassification if reported cases are used without validation.

Figures & Tables

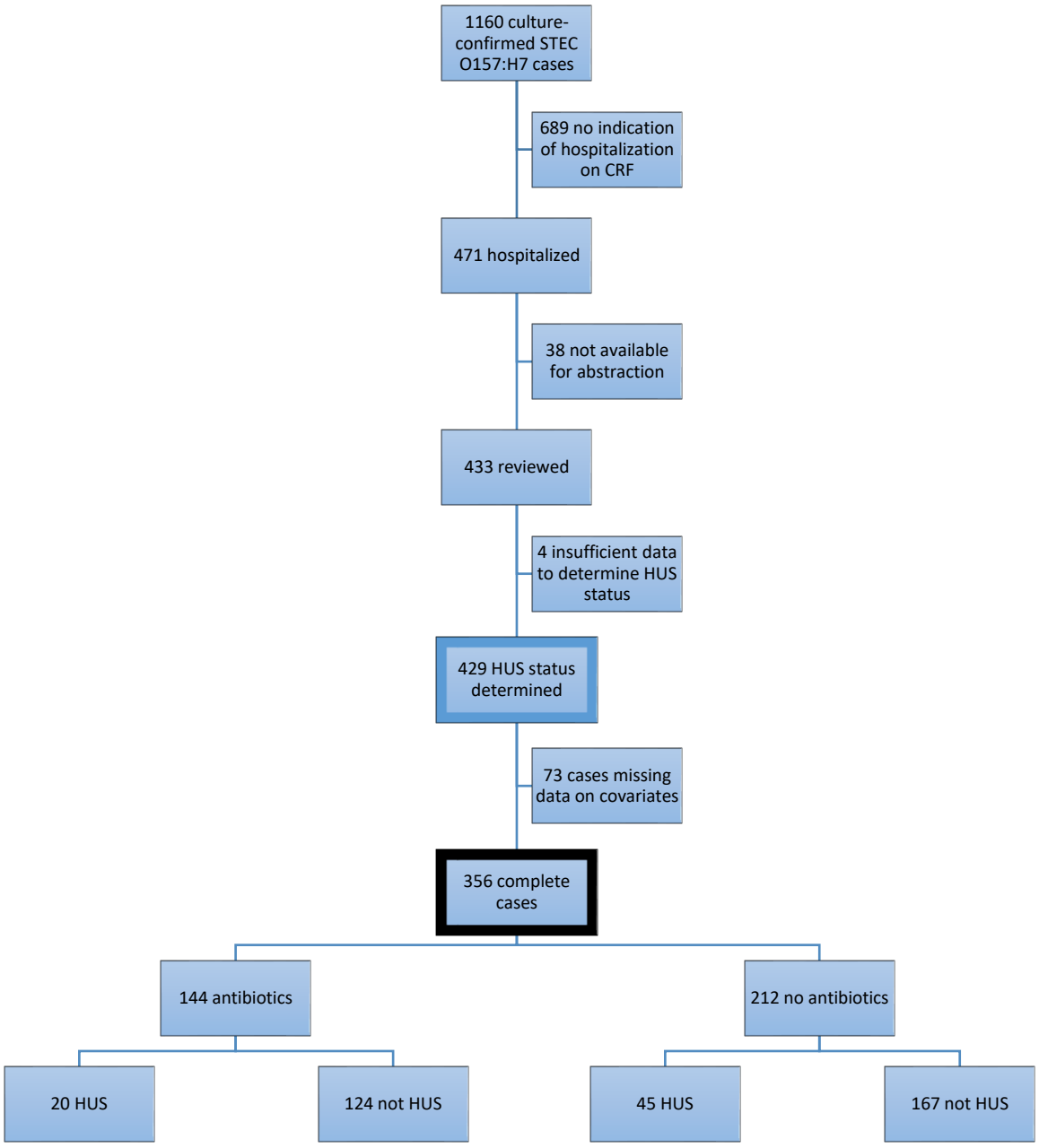

Figure 2.1. Study population. The population for definition comparison (heavy blue outline) and association of antibiotic use and HUS (heavy black outline). HUS determinations are according to the Stringent Clinical Definition.

Abbreviations: CRF, case report form; HUS, hemolytic uremic syndrome; STEC, Shiga toxin-producing *Escherichia coli*

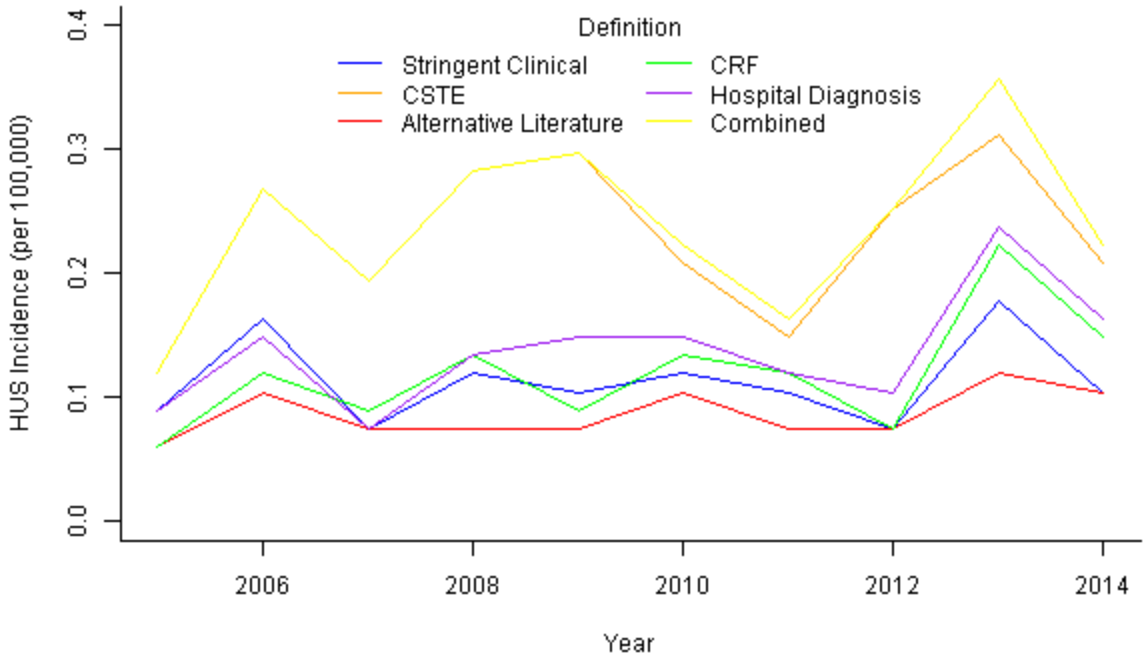

Figure 2.2. Temporal patterns of HUS by definition. The CSTE definition and combined definition, which assigns HUS status to any case considered to have HUS according to the CSTE definition or hospital diagnosis, estimate substantially higher incidence of HUS than other definitions. The HCD obscures some of the variation in incidence over time.

Abbreviations: CRF, case report form; CSTE, Council for State and Territorial Epidemiologists; HUS, hemolytic uremic syndrome

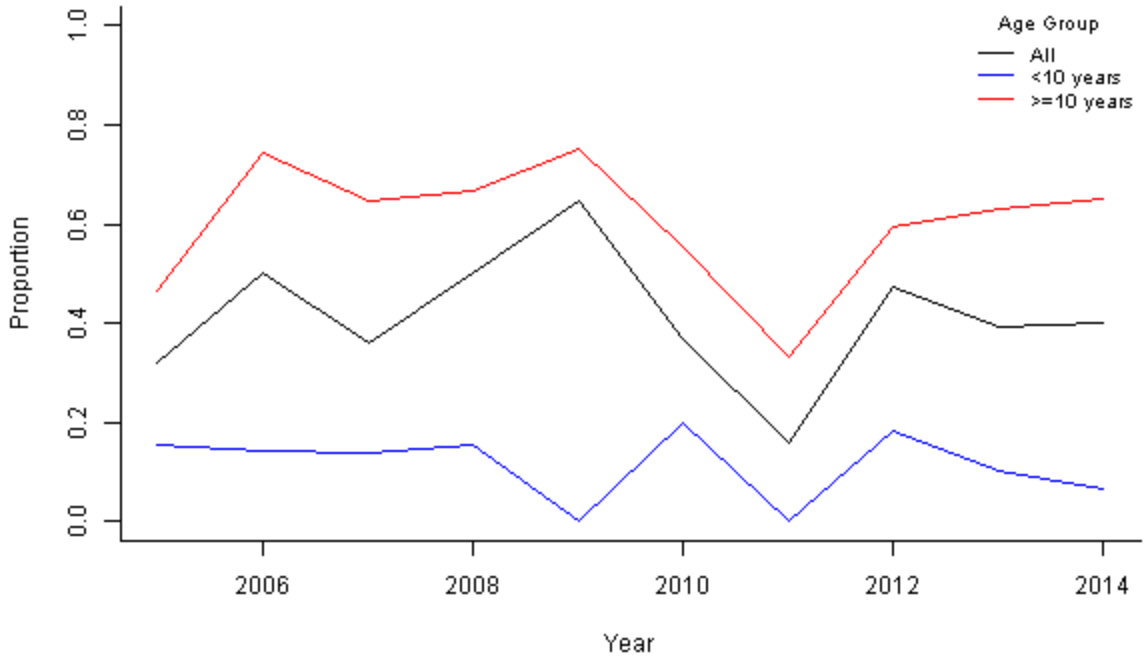

Figure 2.3. Temporal pattern of proportion of cases receiving antibiotics among all STEC O157:H7 cases. Antibiotic use was generally low among children <10 years-old, with little fluctuation. Over half of children and adults ≥10 years-old hospitalized with STEC O157:H7 infections received antibiotics in most years.

Abbreviation: STEC, Shiga toxin-producing *Escherichia coli*

Table 2.1. Hemolytic Uremic Syndrome Definitions

Criteria	Anemia	Thrombocytopenia	Renal injury	Timing	Comments
Stringent clinical definition (SCD) (38-44)	Hematocrit <30%	Platelet count >150,000 mm^{-3}	Serum creatinine > upper limit of normal for age (48)	All must be present on a single day.	Microangiopathic changes on peripheral blood smear are also expected but are not necessary to meet the definition if a smear was not performed or only performed early in the illness.
Council of State and Territorial Epidemiologists (CSTE) (45)	*Confirmed*: Anemia with microangiopathic changes *Probable*: Anemia		Hematuria; proteinuria; creatinine ≥1.0 mg/dL for children <13 years-old and ≥1.5 mg/dL for ≥13-year-olds; 50% increase in creatinine from baseline	*Confirmed*: HUS onset within 21 days of diarrhea onset *Probable*: HUS follows diarrhea	Thrombocytopenia is expected but not required. Baseline creatinine values were not known for most patients in this review. Probable cases may be missing only one confirmed criterion.
Hematology-focused confirmed definition (HCD) (30, 46, 47)	*Confirmed*: Anemia with microangiopathic changes *Probable*: Anemia	Platelet count <150,000/mm^3	Creatinine ≥1.0 mg/dL for children <13 years-old and ≥1.5 mg/dL for ≥13-year-olds	HUS onset within 21 days of diarrhea onset	
Case report form (CRF)	HUS indicated on the DOH case report form based on investigation by the local health jurisdiction.				

Hospital diagnosis	HUS indicated in the hospital discharge note or diagnostic codes.
Combination definition	HUS as indicated by the CSTE definition or hospital diagnosis.

Table 2.2. Clinical Outcomes by HUS Definition

Variable	Stringent Clinical Definition	CSTE Definition*	Hematology-focused Confirmed Definition	Case Report Form	Hospital Diagnosis	Combination Definition	Full Cohort
Number of cases	76	154	58	80	92	160	433
Incidence per 100,000							
<18 years-old	0.41	0.59	0.31	0.44	0.51	0.63	-
All ages	0.11	0.23	0.09	0.12	0.14	0.24	-
Bloody diarrhea (%)	74 (99%)	149 (98%)	57 (100%)	78 (99%)	91 (99%)	155 (98%)	407 (96%)
Missing	1	2	1	1	0	2	11
Vomiting (%)	65 (89%)	114 (77%)	49 (88%)	66 (86%)	75 (85%)	118 (77%)	259 (65%)
Missing	3	5	2	3	4	6	37
Days hospitalized, median (IQR)	13 (10, 19)	7 (4, 14)	13.5 (10, 21)	12 (7, 16.5)	12 (7, 17)	7 (4, 13)	3 (2, 6)
Missing	0	1	0	1	0	1	7
Urine output							
Anuria (%)	29 (44%)	30 (28%)	26 (49%)	25 (36%)	28 (36%)	30 (27%)	34 (17%)
<1.0 ml/kg/hr (%)	55 (83%)	75 (69%)	43 (81%)	50 (72%)	57 (73%)	76 (68%)	117 (59%)
Missing	10	46	5	11	14	48	233
Underwent dialysis (%)	39 (52%)	41 (28%)	33 (58%)	36 (47%)	39 (42%)	41 (27%)	41 (10%)
Missing	1	8	1	3	0	8	37

Each column describes data for only those patients considered to have HUS according to the stated definition. The full cohort column refers to the 433 cases whose hospital records were abstracted.

*Includes confirmed and probable definitions.

Abbreviation: CSTE, Council for State and Territorial Epidemiologists; HUS, hemolytic uremic syndrome; IQR, interquartile range; SD, standard deviation

Table 2.3. Sensitivity and Specificity of HUS Definitions, Using the Stringent Clinical Definition as Comparator

	CSTE Definition	Hematology-focused Confirmed Definition	Case Report Form	Hospital Diagnosis	Combined Definition
All cases	429	429	384	410	429
Sensitivity (95% CI)	96% (89%, 99%)	74% (62%, 83%)	86% (76%, 93%)	97% (91%, 100%)	99% (93%, 100%)
Specificity (95% CI)	77% (72%, 81%)	99% (98%, 100%)	94% (91%, 96%)	94% (91%, 97%)	76% (71%, 80%)
Cases <10 years-old	174	174	152	169	174
Sensitivity (95% CI)	97% (88%, 100%)	75% (62%, 85%)	89% (78%, 96%)	100% (91%, 100%)	100% (94%, 100%)
Specificity (95% CI)	81% (72%, 88%)	99% (95%, 100%)	88% (79%, 93%)	89% (82%, 94%)	77% (69%, 85%)
Cases ≥10 years-old	255	255	232	241	255
Sensitivity (95% CI)	94% (71%, 100%)	71% (44%, 90%)	75% (48%, 93%)	88% (62%, 98%)	94% (71%, 100%)
Specificity (95% CI)	75% (69%, 81%)	100% (98%, 100%)	97% (93%, 99%)	97% (94%, 99%)	75% (69%, 81%)
Received antibiotics	178	178	163	170	178
Sensitivity (95% CI)	92% (73%, 99%)	67% (45%, 84%)	77% (55%, 92%)	91% (72%, 99%)	96% (79%, 100%)
Specificity (95% CI)	77% (69%, 83%)	100% (96%, 100%)	97% (93%, 99%)	98% (94%, 100%)	77% (69%, 83%)
Did not receive antibiotics	238	238	209	236	238
Sensitivity (95% CI)	98% (90%, 100%)	78% (65%, 89%)	90% (77%, 97%)	100% (90%, 100%)	100% (93%, 100%)

Specificity (95% CI)	77% (70%, 83%)	99% (96%, 100%)	93% (87%, 96%)	92% (87%, 95%)	75% (68%, 81%)

Sensitivity and specificity were calculated for those cases with complete data for the stringent clinical definition and the comparative definition or source. For subgroup analyses, age and antibiotic use, respectively, must also be non-missing.

Abbreviation: CI, confidence interval; HUS, hemolytic uremic syndrome

Table 2.4. Association of Antibiotic Use and HUS Development by HUS Definition

Definition	Odds Ratio	95% Confidence Interval	*P*
SCD	2.48	1.15, 5.41	0.020
CSTE	1.13	0.66, 1.92	0.657
HCD	1.97	0.86, 4.48	0.105
Case Report Form	2.08	0.95, 4.61	0.068
Hospital Diagnosis	2.18	1.05, 4.60	0.038
Combined	1.08	0.64, 1.84	0.773

Logistic regression of HUS status on antibiotic use. Antibiotic use is defined as administration of any antibiotic prior to HUS diagnosis (cases) or illness nadir (non-cases). All analyses adjusted for age, sex, vomiting, bloody stool, and day of specimen collection.

Abbreviations: CSTE, Council for State and Territorial Epidemiologists; HCD, hematology-focused confirmed definition; HUS, hemolytic uremic syndrome; SCD, stringent clinical definition

Chapter 3. Identifying the Role of Age and Shiga Toxin in the Association between *Escherichia coli* O157:H7 Phylogenetics and Hemolytic Uremic Syndrome

Gillian A. M. Tarr[1,2], Smriti Shringi[3], Hanna N. Oltean[4], Jonathan Mayer[1], Peter Rabinowitz[2,5], Jon Wakefield[6], Phillip I. Tarr[7], Thomas E. Besser[3], Amanda I. Phipps[1]

[1]Department of Epidemiology, University of Washington

[2]Center for One Health Research, University of Washington

[2]Veterinary Microbiology and Pathology, Washington State University

[4]Washington State Department of Health

[5]Department of Environmental and Occupational Health Sciences, University of Washington

[6]Department of Biostatics, University of Washington

[7]Department of Pediatrics, Washington University in St. Louis School of Medicine

Introduction

Although significant progress has been made in reducing the incidence and impact of Shiga toxin-producing *Escherichia coli* (STEC) O157:H7, it remains the largest cause of postdiarrheal hemolytic uremic syndrome (HUS) (62). Postdiarrheal HUS incidence varies by age, with the greatest burden among children <5 years old (13, 14, 31, 32). In 2014, children 1-4 years old experienced 0.73 HUS cases per 100,000, followed by 5-14 year olds at 0.21 per 100,000 (13). In comparison, 2014 incidence among adults ages 35-64 was <0.02 per 100,000.

Beyond age, pathogen characteristics are an important factor in determining progression to HUS. Shiga toxin (Stx), STEC O157:H7's cardinal virulence factor, can be encoded by multiple genes (most commonly *stx1*, *stx2a*, and *stx2c*), with some genotypes more highly associated with HUS than others (63-66). Expanding on the potential role of pathogen characteristics in HUS etiology, a seminal study in 2008 identified a subtype of STEC O157:H7, termed clade 8, associated with increased risk of HUS (20). This study found that clade 8 strains were most commonly isolated from children 0-18 years of age and carried *stx2a* either alone or in combination with *stx2c* (20). Although numerous studies have investigated virulence factor expression that may be responsible for this association (67-70), population-based replications to confirm the association have been limited, varying in size, methodology, and result (21, 22, 71).

The phylogenetic definition of the STEC O157:H7 serotype has advanced since the 2008 discovery of hypervirulent clade 8. Bono et al. (23) reported a tree of phylogenetic lineages based on isolates from humans, animals, and the environment. As opposed to the earlier tree, the newly characterized phylogenetic lineages drew on a large pool of systematically chosen single nucleotide polymorphisms (SNPs) and incorporated isolates from a diverse set of sources. From limited supplemental data published on this newer tree (27) and analysis of the correspondence of Manning's clades and Jung's lineages in our own data, lineages IIa and IIb appear to overlap with clade 8. We tested the association between these lineages and HUS. This assessment serves two purposes: 1) to increase the specificity of the association and 2) to

determine whether further studies are warranted to explore areas of overlap and difference between clades and lineages. A more refined understanding of the association between STEC O157:H7 phylogenetics and HUS will enable improved targeting of prevention and intervention efforts.

Given the higher incidence of HUS among young children and the preponderance of clade 8 strains isolated from children by Manning et al. (20), we investigated the role of age in the association between phylogenetic lineage and HUS, evaluating it as a potential confounder and effect modifier. To elucidate the mechanism between lineage and HUS, we evaluated *stx* subtypes as mediators and determined the portion of the total association due to specific *stx* profiles.

Methods

Study Setting and Design

We conducted a population-based retrospective cohort study of all culture-confirmed STEC O157:H7 cases reported to the Washington State Department of Health (DOH) between 2005 and 2014. STEC case reporting, mandated by Washington Administrative Code, occurs primarily through diagnostic laboratories and healthcare providers. Local health jurisdiction personnel use a standardized DOH case report form to abstract medical records and interview cases to obtain demographic information, potential exposures, and details of the course of illness.

HUS was defined as <30% hematocrit, <150,000 platelets/mm^3, and above the normal serum creatinine concentration for age. All criteria needed to be met on the same day. HUS status was confirmed during a DOH review of all reported, hospitalized, culture-confirmed STEC O157:H7 cases from the study period (Chapter 2). Cases were considered hospitalized if they had stayed at least overnight at an inpatient medical facility. Non-hospitalized cases were assumed to not have HUS due to the severity of the disease.

This study was deemed exempt by the Washington State Institutional Review Board.

Isolate Typing

All STEC O157:H7 isolates were sent to DOH for microbiologic confirmation and pulsed field gel electrophoresis (PFGE) analysis. We obtained these isolates from DOH and determined their phylogenetic lineage by single nucleotide polymorphism-typing a subset of isolates using the 48-plex SNP assay developed by Jung et al. (27). Concordance among isolates with identical PFGE profiles was confirmed by stratifying each PFGE profile by lineage (Appendix). Isolates that did not undergo SNP-typing were then assigned the lineage of a SNP-typed isolate with the same PFGE profile. Phylogenetic lineage was classified as Ib, IIa, or IIb. Remaining lineages were grouped into a single "rare" category.

Because *stx* genes were expected to vary within isolates sharing a PFGE pattern, a case-cohort sampling strategy was used to identify isolates for Stx-encoding bacteriophage insertion (SBI) typing. A subcohort was randomly sampled from the full cohort, and all remaining HUS cases were added to it. The SBI typing methods have been described previously (27, 72). Briefly, PCR was used to detect 12 targets, including *stx1*, *stx2a*, and *stx2c*.

Statistical Analysis

Case data were merged with isolate results using a unique identifier, and the dataset was de-identified prior to analysis. Age (73-76) and sex (77-79) were considered *a priori* confounders based on prior literature. A directed acyclic graph was constructed to summarize the potential associations of other variables. Distributions of potential confounders and mediators were summarized in contingency tables. Death and whether a case was outbreak-related were generally reported on the case report form only if positive. For these variables, the number and percent of cases with the status indicated as positive were determined, implicitly

assuming that missing values were negative. Aside from age, none of the examined variables were significantly associated with both lineage and HUS.

R (53) was used for all analyses.

Logistic Regression with Generalized Estimating Equations

Logistic regression with generalized estimating equations (GEE) was used to estimate the association between lineage and HUS. In this context, lineage is considered a group-level variable, with groups defined by PFGE types. Lineage is the exposure of interest, and the correlation among isolates of the same PFGE type is considered a nuisance. An exchangeable working correlation matrix was used for all analyses. Robust standard errors calculated using the sandwich estimator accounted for any potential misspecification of the correlation structure. The effect of lineage was estimated using lineage Ib, the most common lineage, as the comparator. Unadjusted and adjusted models provided estimates of the odds ratio (OR) for risk of HUS in the other lineages (e.g. IIa) relative to Ib.

To determine the importance of accounting for lack of independence among isolates of the same PFGE type, we conducted two additional analyses. First, we regressed HUS on lineage, age, and sex without using GEE. This provided an estimate of the OR as though all isolates were independent. Second, we conducted a simulation study in which one isolate per PFGE type was randomly drawn and used in a single-level logistic regression, which emulated the approach used in previous studies. After 10,000 repetitions of the simulation, we calculated the proportion of lineage OR estimates that were above 1.0 and the proportion that were statistically significant.

Effect Modification by Age

To elucidate the role of age in the association between lineage and HUS, we examined effect modification of the OR estimate by age. The primary GEE analysis was stratified by age group and the lineage OR estimates were compared across strata.

Mediation by Shiga Toxin Genotype

Mediation of the lineage-HUS association by *stx* genes was tested using the potential outcomes framework for formal mediation analysis (80). This causal mediation approach accounts for limitations in traditional mediation analyses (81-84), such as bias from mediator-outcome confounders. Potential outcomes are those a subject would have experienced had their exposure been different from their actual observed exposure. Potential outcomes mediation analysis allows calculation of the average causal mediated effect (ACME) and the average direct effect (ADE) (Appendix).

We tested *stx* genes in two ways. First, the three main subtypes, *stx1*, *stx2a*, and *stx2c*, were treated as independent. Presence or absence of each gene was coded for each isolate, and mediation was tested for each gene individually. Second, because of literature showing that the effect of Stx subtypes may vary depending on their exact gene combinations, mediation was also tested according to overall *stx* genotype (e.g. having both *stx1* and *stx2a*). The dominant genotypes were identified for lineages IIa and IIb, and the remaining genotypes were grouped together as the comparator.

We used the parametric estimation algorithm outlined by Imai et al. (85). First, a binomial model with a probit link was fit for the mediator (*stx* gene or genotype) with lineage, age, and sex as predictors. A second binomial with probit link model was then fit for HUS status with the mediator, lineage, age, and sex as predictors. From these models, the *mediation* package in R was used to estimate the ACME and ADE (86). Use of the probit link enabled interpretation of results as the increase in probability of HUS due to the mediated or direct path.

Each mediation analysis was conducted with 10,000 simulations and used robust standard errors to estimate 95% confidence intervals (CIs).

In sensitivity analyses, we tested the assumption of no interaction between exposure and mediator by allowing the ACME and ADE to vary by lineage for mediators that were statistically significant in their initial mediation analysis. In a *post hoc* analysis, we also explored modification of mediation effects by age group (Appendix).

Results

During the years 2005-2014, 1196 STEC O157:H7 cases were reported to DOH. Of these, 1160 were culture-confirmed and eligible to be included in the study. Using a standardized definition, HUS status was validated for 1118 cases. HUS occurred in 76 cases. HUS status differed by age, with children <5 years-old constituting over half of HUS cases but less than a quarter of non-HUS cases (Table 3.1). Case fatality was 3.9% among HUS cases and was 0.4% among non-HUS cases.

Lineage was assigned to 1121 isolates, including 1082 of those with validated HUS status. Of the 39 excluded isolates, six were biochemically atypical STEC O157:H7 and 33 were not available for typing (Appendix). We SNP-typed 793 isolates, and matched, by extension, another 328 to a known lineage using PFGE.

Phylogenetic Association with HUS

In the unadjusted GEE model, lineage IIb was associated with increased risk of HUS relative to lineage Ib (OR=1.65; 95% CI 1.05, 2.60) (Table 3.2). There was no elevation in HUS risk among lineage IIa cases, compared to lineage Ib cases. No HUS cases occurred in the group of rare lineages; effect estimates for this group are not presented because of statistical instability. After adjustment for age and sex, the association between IIb and HUS was attenuated and no longer distinguishable from the null (OR=1.43; 95% CI 0.90, 2.25).

We demonstrated the impact of accounting for non-independence among isolates of the same PFGE type using adjusted logistic regression without use of GEE. As expected, identical point estimates as the GEE analysis were obtained for lineage IIa (OR=0.97; 95% CI 0.53, 1.73) and lineage IIb (OR=1.43; 95% CI 0.79, 2.53), with a larger confidence interval for lineage IIb.

We demonstrated the impact of using all isolates in 10,000 simulations of randomly selecting one isolate per PFGE type. The association between lineage IIb and HUS was >1 for 97% of the draws, and 25% were statistically significant (Appendix, Figure S1). In other words, approximately one quarter of risk estimates of lineage IIb association with HUS will be statistically significantly >1 when analyzing a single representative isolate due purely to the isolates randomly selected.

Effect Modification by Age

The ratio of Ib isolates to IIb isolates increased with age, while the ratio of Ib to IIa isolates remained stable across age groups until the ≥60-year group (Table 3.2). In 0-4 year-olds, the ORs for both lineage IIa and IIb, relative to lineage Ib, were <1 but not statistically significant. The OR for both lineages increased with age through the 20-59 year-old age group. Among those aged 20-59 years, lineage IIa had an OR of 12.7 (95% CI 1.57, 103) and lineage IIb had an OR of 8.50 (95% CI 1.13, 63.7). There were no lineage IIa or IIb HUS cases in the ≥60-year group.

Mediation by Shiga Toxin Genotypes

The case-cohort sample in which SBI typing was conducted consisted of 469 cases, 453 of which also had a validated HUS status. This included 393 non-HUS and 27 HUS members of the randomly selected subcohort, as well as 33 additional HUS cases. Isolates from 16 HUS cases were not SBI-typed and thus could not be included in the case-cohort sample (Appendix). The subcohort was moderately younger, contained a greater proportion of cases with lineage IIa and

IIb isolates, and had somewhat more person-to-person transmission than the full cohort (Table 3.1). HUS cases within the subcohort also had more comorbidities and lower occurrence of dialysis than HUS cases in the full cohort but were otherwise similar in their exposures and clinical course.

Distribution of *stx* genotypes by lineage showed that 92% of isolates contained *stx2a*, whether alone or in combination with another *stx* gene (Table 3.3). Lineage Ib isolates were dominated by the *stx1-stx2a* genotype (90%). Lineage IIa isolates were predominantly the *stx2a-stx2c* genotype (84%). Most lineage IIb isolates (94%) had only the *stx2a* gene. Six isolates had none of the three probed *stx* genes at the time of typing.

Single *stx* genes did not display any measurable mediation of the association between lineage and HUS (Table 3.4).

After the dominant genotypes for lineages IIa (*stx2a-stx2c*) and IIb (*stx2a*) were identified, the remaining genotypes (Table 3.3) were grouped as the comparator for genotype mediation analyses. However, the *stx2a-stx2c* genotype could not be tested for mediation due to the small sample size in the Ib comparator lineage. The *stx2a*-only genotype displayed statistically significant mediation of the lineage-HUS association (Table 3.4). Similar patterns were observed for both lineages IIa and IIb. The total effects were close to zero, reflecting direct and mediated effects of opposite sign. The ADE were negative, although not statistically significant, indicating a tendency toward decreasing probability of HUS associated with non-*stx* gene characteristics of the lineages. The ACME for lineage IIa was 0.13 (95% CI 0.012, 0.28) and for IIb was 0.17 (95% CI 0.018, 0.31). These indicate an increase in the probability of HUS due to higher frequency of *stx2a* genotypes among lineage IIa and IIb isolates. For example, 0.17 is the average of two quantities: 1) the difference in HUS probability between lineage IIb with the *stx2a* genotype frequency regularly associated with lineage IIb (i.e. observed) and the potential outcome of lineage IIb if it had the *stx2a* genotype frequency associated with lineage Ib (i.e. counterfactual), and 2) the difference in HUS probability between the potential outcome of

lineage Ib if it had the *stx2a* genotype frequency associated with lineage IIb (i.e. counterfactual) and HUS probability for lineage Ib with the *stx2a* genotype frequency associated with lineage Ib (i.e. observed).

The assumption of no interaction could only be tested for the mediated lineage IIb-HUS association, because there were insufficient isolates with comparator genotypes in lineage IIa. For the lineage IIb-HUS association, relaxing the assumption to allow interaction between *stx2a* and lineage produced a marginally significant ACME for lineage Ib, indicating an increase in the probability of HUS due to *stx2a* of 0.19 (95% CI -0.004, 0.43; p = 0.06). The ACME for lineage IIb was indistinguishable from 0. However, the test for difference between the ACME estimates was not significant (p = 0.50), precluding conclusions that interaction was present. We did not detect any modification of the mediation effect by age (Appendix, Table S5).

Discussion

We sought to refine the association between STEC O157:H7 phylogeny and HUS by using systematically-defined phylogenetic lineages. To better understand drivers of this association at the population level, we elucidated the role of age and Stx subtypes. While lineage IIb was associated with increased risk of HUS in unadjusted analysis, there was no significant impact when adjusting for age. However, stratified analysis indicated that age modifies the effect of both lineages IIa and IIb on HUS. Finally, we found evidence of mediation of the association between lineage and HUS by *stx2a* when it is the only type of *stx* gene present in the bacteria.

Our results deepen the body of literature on virulence heterogeneity of the STEC O157:H7 serotype and explicate roles for age and Stx subtype that explain part of the association between phylogeny and HUS. Other studies have identified an elevated risk of HUS among patients infected with clade 8 strains (20-22). Lineages IIa and IIb in the present study overlap with clade 8 (27) and show a similar elevation in risk of HUS. However, our study shows that this increase is only among adults. In young children, no statistically significant association

between lineage and HUS exists, although in the youngest age group there is a suggestion of protective effect for lineages IIa and IIb as compared to lineage Ib. Yet lineage IIb is more common in children <10 years-old, who are more likely to develop HUS, giving the appearance in unadjusted and unstratified analyses of increased risk associated with lineage IIb. This may be a driver of observed associations in previous studies of clade 8. These studies have either not adjusted for age (22) or adjustment by large age groups left room for residual confounding (20).

Iyoda et al. (21) explored effect modification by age of the odds of clade 8 infection between HUS patients and asymptomatic STEC O157:H7 carriers. They obtained an OR of 6 for 0-9 year-olds and an OR of 3 for children and adults ≥10 years. There are multiple reasons our results may differ from these. Iyoda et al. (21) used a case-control design with HUS cases drawn from reported STEC O157:H7 cases and controls drawn from food handlers and daycare workers, with child controls added from an unidentified source. Although the population from which the controls was drawn is likely a subset of the larger population from which the cases were drawn, using a counterfactual framework we can see that only a minority of cases would have been eligible to be controls had they not developed HUS. Lack of comparability between cases and controls may have influenced their results. Additionally, both Manning et al. (20) and Iyoda et al. (21) used one representative isolate from each outbreak or PFGE-defined strain. We demonstrated through simulation that studies using only one isolate per strain may show an association in 25% of analyses merely by chance of the isolates selected.

The largest factor likely explaining the difference between our results and those of Iyoda et al. (21) is that study's use of asymptomatic controls. Thus, their effect estimates combine the effect of clade 8 on both becoming ill and progressing from illness to HUS. In other words, their study assessed the combination of pathogenicity and virulence, whereas our analysis focus on estimating virulence alone. Notwithstanding the potential biases outlined above, their results could be interpreted together with our results, which focus on the progression from illness to HUS, to better understand the point in STEC O157:H7 natural history when clade 8 or lineages

IIa and IIb have their greatest impact. Specifically, if, as our results show, there is no significant increased risk of developing HUS once illness is established among 0-9 year olds with lineage IIa or IIb infection, as compared to lineage Ib infection, there would need to be a significant effect on the development of illness following exposure for the OR of 6 between asymptomatic controls and HUS cases reported by Iyoda et al. (21) to be true.

Age has long been considered the strongest predictor of progression to HUS among those with STEC O157:H7 infection. The lack of association between lineages and HUS among those aged <10 years suggests that differential infection by high virulence lineages does not explain why young children are more likely to progress from STEC O157:H7 illness to HUS. However, our findings show that lineage IIb strains disproportionately establish disease in young children, driving the observed unadjusted association between lineage IIb and HUS. That lineages IIa and IIb do not confer a higher risk of HUS among 0-9 year-olds, relative to lineage Ib, suggests that there is a difference in either exposure or early disease manifestation that leads to more IIb-infected cases being reported in this age group than cases infected with other lineages. As outlined above, if the results of Iyoda et al. (21) are correct and can be applied to lineages, lineage IIb bacteria may more easily establish disease in children aged <10 years, compared to lineage Ib bacteria. Alternatively, exposure to different lineages may differ across age groups, with more young children exposed to lineage IIb. For example, if lineage IIb is preferentially associated with raw milk (Chapter 4), and young children drink more milk than older children and adults, they may be exposed to more IIb strains. Using phylogenetics to better understand the distribution of infection will enable us to target prevention strategies and explore therapeutic advances.

The elevation of HUS risk among older children and adults with lineage IIa or IIb infections must not be misinterpreted as a higher absolute risk than among young children. The risk is relative to those infected with lineage Ib strains: among 10-59 year-olds with STEC O157:H7 infections, those with lineage IIa or IIb strains are more likely to progress to HUS than

those with lineage Ib strains. This association may suggest that older children and adults are more resistant to severe disease development from lineage Ib infection. We see substantially more reported cases infected by Ib than IIa and IIb lineages in this age group, as though lineage IIa and IIb strains have more difficulty establishing disease in these individuals. However, if IIa or IIb strains are successful in establishing disease, they are more likely to cause HUS than the more easily-established lineage Ib strains. In the oldest group, ≥60 years-old, lineage Ib may confer greater risk of HUS than either lineage IIa or IIb. However, our data in this group were sparse and preclude definitive conclusions. More work is needed to identify the characteristics of STEC O157:H7 lineages that explain differences in virulence across age groups.

Studies of clade 8 isolates have described the potential for high Stx2 production (68, 87, 88). The *stx2a*/*stx2c* and *stx2a*-only genotypes, shown to be independently associated with progression to HUS (63, 64), are the most common genotypes among clade 8 isolates (20, 21, 71). However, it has been unclear what portion of the clade 8-HUS association is due to the common clade 8 *stx* gene profiles as opposed to other clade 8 virulence factors. We show mediation of the lineage-HUS association by the *stx2a*-only genotype, which is consistent with the literature on clade 8.

We observed very close correlation of lineage and *stx* genotype, which is similar to previous studies (27, 89). While this precluded mediation analysis for the *stx2a*/*stx2c* genotype due to lack of sufficient variability, identification of the *stx2a* genotype as a significant mediator of the association between both lineages IIa and IIb and HUS is reassuring. As the only statistically significant pathway of effect, these results suggest that much of the virulence associated with lineage IIb is due to the *stx2a*-only genotype. We did not gain clarification of whether the mediated effect of *stx2a* is modified by age. Modification of the mediation effect would have helped explain the modification of the overall effect seen in the primary analysis, and lack of modification would have pointed to other pathogen characteristics at play in the

modification of the overall effect. This question should be studied with a larger SBI-typed cohort.

Although there is overlap between clade 8 and lineages IIa and IIb, as well as overlap between clade 2 and lineage Ib (27), these classifications are based on separate sets of SNPs, and neither classification is a subset of the other. However, cases in which the clade and lineage classifications diverge could provide useful clues as to the mechanism behind differences in epidemiology. For example, future studies could investigate whether the epidemiology of clade 8/lineage IIa strains differs from clade 8/lineage IIb strains. Such a finding would suggest the differences between lineages, but not necessarily clades, are important drivers of population patterns.

This study was limited to reported cases. STEC O157:H7 cases who did not seek care, were not tested, or were not reported by their providers were not included. Unreported cases were likely milder, and it is unlikely that any HUS cases went unreported. If there is an association between phylogenetic lineage and severity, this could have induced selection bias. However, our simulation study of this potential bias suggests that it would have induced only a small attenuation of the effect if present (Appendix).

We were not able to assign phylogenetic lineage to 39 isolates. These isolates tended to be from earlier in the study period, indicating that they are not missing completely at random. Composition of the bacterial population did shift slightly during the study period (Chapter 4), with lineage Ib more dominant early in the period. The proportion of HUS cases among the 39 untyped isolates was 3%, comparable to the proportion among Ib cases. Given the temporal and HUS patterns of the untyped isolates, their inclusion would likely have accentuated the lower HUS frequency among Ib isolates, increasing the apparent effect of other lineages. However, the small number of untyped isolates relative to the whole sample was unlikely to have meaningfully altered our results.

In the mediation analysis, we were limited to only 453 SBI-typed cases with validated HUS status and lineage designation. Although we used a case-cohort sampling scheme to reduce the bias that would have been introduced by including all isolates SBI-typed for other reasons, this reduced our power to identify mediation effects. Additionally, 16 HUS cases were not SBI-typed, leading to their exclusion from the mediation analysis. Including these cases may have altered the results. Consistent with previous studies of *stx* genotype association with HUS (63, 64), that we observed mediation only with the *stx2a* genotype and not any of the individual *stx* genes is reassuring. A larger cohort with more variation in genotypes is needed to rigorously test mediation by other genotypes and further explore mediation modified by age.

This study benefited from over 1,100 STEC O157:H7 cases, including 76 HUS cases, which is many times the number of HUS cases in the original Manning et al. work (20). We validated HUS outcomes with hospital records using a standardized definition to ensure the comparability of our outcome. Detailed data from DOH case report forms enabled us to sensitively investigate effect modification by age. By employing correlated data methods, we were also able to incorporate data from the entire cohort instead of limiting the study to representative isolates from each PFGE type, which our simulation study showed is an important step in avoiding bias.

This study demonstrates that the association between phylogenetic lineage and HUS is modified by age and due, at least in part, to lineage IIa and IIb containing a high proportion of *stx2a*-only strains. Our findings suggest that future efforts should focus on determining how STEC O157:H7 exposure by lineage differs across age groups. Identifying lineage-correlated pathogen characteristics that affect both disease establishment and progression to HUS will also be important in explaining age-specific patterns. With overlap between the lineages and clades, it is likely that our results can be extended to clades. More work is needed to leverage areas of overlap and difference between clades and lineages to elucidate potential mechanisms.

Tables

Table 3.1. Frequency of Case Characteristics of Patients Reported in Washington State with Confirmed STEC O157:H7 Infection, 2005-2014

Variable	Full Cohort		Subcohort	
	No HUS (n = 1042)	HUS (n = 76)	No HUS (n = 393)	HUS (n = 27)
Sex				
Female	591 (57.0%)	46 (60.5%)	216 (55.0%)	17 (63.0%)
Male	445 (43.0%)	30 (39.5%)	177 (45.0%)	10 (37.0%)
Age group				
<5 years	227 (21.8%)	41 (53.9%)	93 (23.7%)	11 (40.7%)
5-9 years	145 (13.9%)	18 (23.7%)	64 (16.3%)	9 (33.3%)
10-19 years	184 (17.7%)	5 (6.6%)	64 (16.3%)	2 (7.4%)
20-59 years	358 (34.4%)	7 (9.2%)	124 (31.6%)	2 (7.4%)
≥60 years	127 (12.2%)	5 (6.6%)	48 (12.2%)	3 (11.1%)
Ethnicity				
Hispanic or Latino	97 (12.5%)	8 (11.1%)	50 (15.9%)	4 (14.8%)
Not Hispanic or Latino	676 (87.5%)	64 (88.9%)	264 (84.1%)	23 (85.2%)
Race				
American Indian or Alaskan Native	10 (1.3%)	2 (2.8%)	5 (1.6%)	0
Asian	55 (7.1%)	4 (5.6%)	22 (7.2%)	2 (7.7%)
Black	25 (3.2%)	0	5 (1.6%)	0
Multiracial	11 (1.4%)	0	1 (0.3%)	0
Native Hawaiian or Pacific Islander	5 (0.6%)	0	3 (1.0%)	0
Other	20 (2.6%)	4 (5.6%)	11 (3.6%)	1 (3.8%)
White	651 (83.8%)	62 (86.1%)	260 (84.7%)	23 (88.5%)
Phylogenetic lineage				
Ib	531 (52.7%)	37 (49.3%)	186 (47.3%)	13 (48.1%)

IIa	235 (23.3%)	18 (24.0%)	108 (27.5%)	5 (18.5%)
IIb	173 (17.2%)	20 (26.7%)	72 (18.3%)	9 (33.3%)
Rare*	68 (6.8%)	0	27 (6.9%)	0
Outbreak-associated†	105 (10.1%)	7 (9.2%)	45 (11.5%)	3 (11.1%)
Most likely source of infection				
Animal	70 (17.5%)	6 (18.8%)	23 (15.2%)	2 (15.4%)
Environment	43 (10.7%)	0	19 (12.6%)	0
Food	201 (50.1%)	18 (56.2%)	72 (47.7%)	7 (53.8%)
Person	75 (18.7%)	7 (21.9%)	33 (21.9%)	4 (30.8%)
Water	12 (2.9%)	1 (3.1%)	4 (2.6%)	0
Contact with lab-confirmed case				
Yes	127 (13.7%)	14 (20.3%)	55 (15.5%)	7 (28.0%)
No	799 (86.3%)	55 (79.7%)	300 (84.5%)	18 (72.0%)
Direct animal contact				
Yes	490 (55.9%)	39 (60.0%)	174 (54.4%)	15 (62.5%)
No	387 (44.1%)	26 (40.0%)	146 (45.6%)	9 (37.5%)
Underlying condition				
Yes	101 (10.7%)	9 (13.6%)	33 (9.1%)	5 (21.7%)
No	842 (89.3%)	57 (86.4%)	328 (90.9%)	18 (78.3%)
Bloody diarrhea				
Yes	870 (87.1%)	73 (96.1%)	334 (87.0%)	26 (96.3%)
No	129 (12.9%)	3 (3.9%)	50 (13.0%)	1 (3.7%)
Received dialysis				
Yes	2 (0.2%)	40 (52.6%)	1 (0.3%)	12 (44.4%)
No	1036 (99.8%)	36 (47.4%)	391 (99.7%)	15 (55.6%)
Documented death††	4 (0.4%)	3 (3.9%)	2 (0.5%)	0

The full cohort included 1,118 culture-confirmed STEC O157:H7 cases with validated HUS status. The randomly sampled subcohort of 420 cases was used for case-cohort analyses of mediation by Shiga toxin genes.

*"Rare" lineages includes 12 different lineages

†Whether a case was associated with an outbreak was not reported for most cases, so only positive responses are shown.

††Death status was not reported for most cases. There were eight deaths. Only seven are shown in the table. The eighth was hospitalized, but the chart could not be abstracted to determine HUS status.

Abbreviation: HUS, hemolytic uremic syndrome

Table 3.2. Association of Phylogenetic Lineage and HUS

	N HUS/Total	Odds Ratio	95% Confidence Interval	*P*
Crude				
Lineage Ib	37/568	1	-	-
Lineage IIa	18/253	1.11	0.63, 1.96	0.711
Lineage IIb	20/193	1.65	1.05, 2.60	0.031
Adjusted*				
Lineage Ib	37/561	1	-	-
Lineage IIa	18/253	0.98	0.54, 1.78	0.937
Lineage IIb	20/193	1.43	0.90, 2.25	0.126
Age-stratified: 0-4 years old†				
Lineage Ib	22/118	1	-	-
Lineage IIa	8/71	0.61	0.27, 1.39	0.24
Lineage IIb	10/62	0.73	0.39, 1.36	0.31
Age-stratified: 5-9 years old†				
Lineage Ib	8/79	1	-	-
Lineage IIa	4/31	1.39	0.62, 3.12	0.42
Lineage IIb	6/34	2.38	0.79, 7.13	0.12
Age-stratified: 10-19 years old†				
Lineage Ib	1/95	1	-	-
Lineage IIa	2/50	4.92	0.89, 27.1	0.067
Lineage IIb	2/30	4.99	0.94, 26.4	0.059
Age-stratified: 20-59 years old†				
Lineage Ib	1/200	1	-	-
Lineage IIa	4/78	12.7	1.57, 103	0.017
Lineage IIb	2/48	8.50	1.13, 63.7	0.037
Age-stratified: ≥60 years old†				
Lineage Ib	5/75	1	-	-
Lineage IIa	0/23	0	0, 0	<0.001
Lineage IIb	0/19	0	0, 0	<0.001

Logistic regression using generalized estimating equations (GEE) of HUS status on phylogenetic lineage. No HUS occurred in the group of cases infected with rare lineages, so results are not shown for this group.

*Model adjusted for age as a continuous variable and sex.

†Model adjusted for sex.

Abbreviation: HUS, hemolytic uremic syndrome

Table 3.3. Distribution of Shiga Toxin Genotypes by Phylogenetic Lineages

	All Lineages n (%)	Lineage Ib n (%)	Lineage IIa n (%)	Lineage IIb n (%)	Rare Lineages n (%)
No *stx*	6 (1.3)	2 (0.9)	1 (0.8)	3 (3.3)	0
HUS	0	0	0	0	-
stx1	6 (1.3)	4 (1.9)	0	0	2 (7.4)
HUS	1 (1.7)	1 (3.7)	-	-	0
stx1-stx2a	192 (42.4)	192 (90.1)	0	0	0
HUS	22 (36.7)	22 (81.5)	-	-	-
stx1-stx2c	15 (3.3)	0	0	0	15 (55.6)
HUS	0	-	-	-	0
stx2a	117 (25.8)	13 (6.1)	19 (15.4)	85 (94.4)	0
HUS	24 (40.0)	4 (14.8)	2 (13.3)	18 (100)	-
stx2a-stx2c	106 (23.4)	1 (0.5)	103 (83.7)	2 (2.2)	0
HUS	13 (21.7)	0	13 (86.7)	0	-
stx2c	11 (2.4)	1 (0.5)	0	0	10 (37.0)
HUS	0	0	-	-	0

No isolates in the case-cohort sample were observed with the *stx1-stx2a-stx2c* genotype.

Abbreviations: HUS, hemolytic uremic syndrome; *stx*, Shiga toxin gene

Table 3.4. Average Change in the Probability of HUS Due to Mediated and Direct Effects

	Lineage IIa Estimate (95% CI)	Lineage IIb Estimate (95% CI)
	Gene Presence/Absence	
stx1		
Total effect	-0.018 (-0.088, 0.057)	0.054 (-0.034, 0.15)
ACME	0.084 (-0.049, 0.22)	0.098 (-0.059, 0.25)
ADE	-0.10 (-0.26, 0.048)	-0.044 (-0.22, 0.14)
Proportion mediated	-1.27 (-35.0, 29.3)	1.35 (-17.3, 19.2)
stx2a		
Total effect	-0.017 (-0.089, 0.058)	0.056 (-0.030, 0.15)
ACME	0.001 (-0.008, 0.007)	-0.0001 (-0.007, 0.007)
ADE	-0.018 (-0.089, 0.057)	0.056 (-0.030, 0.15)
Proportion mediated	-0.003 (-0.96, 0.79)	0.001 (-0.38, 0.33)
stx2c		
Total effect	-0.015 (-0.085, 0.061)	0.056 (-0.030, 0.15)
ACME	-0.005 (-0.10, 0.10)	0.00001 (-0.004, 0.005)
ADE	-0.010 (-0.13, 0.11)	0.056 (-0.030, 0.15)
Proportion mediated	0.15 (-14.6, 14.8)	-0.0002 (-0.20, 0.20)
	Genotype*	
stx2a		
Total effect	-0.027 (-0.14, 0.16)	0.057 (-0.029, 0.15)
ACME	0.13 (0.012, 0.28)†	0.17 (0.018, 0.31)†
ADE	-0.16 (-0.32, 0.025)	-0.11 (-0.27, 0.061)
Proportion mediated	-1.10 (-22.7, 21.2)	2.36 (-22.2, 27.3)

Each mediation analysis was conducted for lineages IIa and IIb separately vs. lineage Ib. Under the assumption of no interaction, the ACME, ADE, and proportion mediated shown are the average of the estimates for Ib and the lineage being analyzed. Mediation analyses were

conducted using Imai et al.'s parametric estimation algorithm (85) with 10,000 simulations and robust standard errors.

*Genotype is compared against all non-IIa- and -IIb-dominant genotypes: no *stx*, *stx1*, *stx1-stx2a*, *stx1-stx2c*, *stx1-stx2a-stx2c*, and *stx2c*. Insufficient variation existed to test mediation by the *stx2a-stx2c* genotype.

$^{\dagger}p < 0.05$

Abbreviations: ACME, average causal mediated effect; ADE, average direct effect; CI, confidence interval; HUS, hemolytic uremic syndrome; *stx*, Shiga toxin gene

Chapter 4. Geogenomic Segregation and Temporal Trends of Human Pathogenic *Escherichia coli* O157:H7: Evidence for Persistent Local Reservoirs

Gillian A. M. Tarr[1,2], Smriti Shringi[3], Amanda I. Phipps[1], Thomas E. Besser[3], Jonathan Mayer[1], Hanna N. Oltean[4], Jon Wakefield[5], Phillip I. Tarr[6], Peter Rabinowitz[2,7]

[1]Department of Epidemiology, University of Washington

[2]Center for One Health Research, University of Washington

[2]Veterinary Microbiology and Pathology, Washington State University

[4]Washington State Department of Health

[5]Department of Biostatics, University of Washington

[6]Department of Pediatrics, Washington University in St. Louis School of Medicine

[7]Department of Environmental and Occupational Health Sciences, University of Washington

Introduction

Shiga toxin-producing *Escherichia coli* (STEC) cause major public health challenges, with the global burden of disease ranging from 2.4 to 2.8 million cases annually (90, 91). The STEC O157:H7 serotype is recognized for its disproportionate virulence and relative frequency of isolation, causing >2,000 hospitalizations each year in the United States (1) and leading to hemolytic uremic syndrome (HUS) in 6-15% of reported cases (5, 10, 29). HUS has a case fatality between 3 and 5% (6, 29).

Over two-thirds of reported STEC O157:H7 infections are thought to be attributable to foodborne transmission (1), but epidemiologic studies have implicated other risk factors, including rural residence, untreated water consumption, farm animal contact, and local land use (92-94). Most STEC O157:H7 infections occur sporadically, and the source of infection is often difficult to identify with certainty (1, 11). However, STEC O157:H7 can persist in certain locales, posing ongoing risk to humans (95-103). Most notably, multiple studies demonstrate persistence of specific STEC O157:H7 strains within cattle farms, as well as transmission between neighboring farms (95-100). Potential reservoirs enabling this persistence remain unknown, but include, water, soil, and wild birds (101-103). It is, therefore, possible that humans incidentally acquire these infections by virtue of dwelling in specific geographic regions.

While the globalization of the food supply may deliver a great diversity of pathogens to consumers, the importance of local STEC O157:H7 reservoirs needs to be more fully explored to control endemic infections. Here, we sought to test the hypothesis that geographic clusters, most likely of environmental origin, cause STEC O157:H7 human disease over time, using a generalizable population-based cohort, taking into account the genomic relatedness of different isolates (23, 27), and the geographic, temporal, and secular attributes of their corresponding infections.

Methods

Study Setting and Design

We conducted a population-based retrospective cohort study of all culture-confirmed *E. coli* O157:H7 cases reported to the Washington State Department of Health (DOH) between 2005 and 2014. Approximately 16% of Washington State's 7.1 million inhabitants reside in rural areas, amidst intensive agricultural production, including dairy farming (104). STEC case reporting, mandated by Washington Administrative Code, occurs primarily through diagnostic laboratories and healthcare providers. Local health jurisdiction personnel use a standardized DOH case report form to abstract medical records and interview cases to obtain demographic information (including residence address), potential exposures, and details of the course of illness. Case addresses were geocoded and census block group (BG) determined. All case data were de-identified for analysis.

All STEC O157:H7 isolates were sent to DOH for microbiologic confirmation and pulsed field gel electrophoresis (PFGE) analysis. We obtained these isolates from DOH and determined their phylogenetic lineage using single nucleotide polymorphism (SNP)-typing a subset of isolates using the 48-plex SNP assay developed by Jung et al. (27). We typed each PFGE pattern present in the dataset and oversampled HUS cases. Concordance among isolates with identical PFGE profiles was confirmed by stratifying each PFGE profile by lineage (Appendix). Isolates that did not undergo SNP-typing were then assigned the lineage of a SNP-typed isolate with the same PFGE profile. Phylogenetic lineage was classified as Ib, IIa, or IIb. Remaining lineages were grouped as a single "rare" category.

This study was deemed exempt by the Washington State Institutional Review Board.

Spatial Segregation of Phylogenetic Lineages

Spatial segregation is the ecological concept that one species or species type is more likely to be surrounded by like than by non-like individuals (105). We used Diggle's kernel regression method to test spatial segregation of STEC O157:H7 by phylogenetic lineage (106).

This method of estimating spatial segregation provides an overall test of spatial segregation, and identifies statistically significant regions of segregation (and association) by lineage. We used a Gaussian kernel function and selected a model bandwidth using a cross-validated log-likelihood function. We tested bandwidths between 0.02 and 1 at 0.0098 increments, and the bandwidth associated with the highest cross-validated log-likelihood was selected (0.6472). The test statistic for spatial segregation summed the square of the difference between the kernel regression-estimated lineage-specific probability at a given location and the overall probability that a case isolate belongs to that lineage over all lineages and all case locations. To determine statistical significance, we performed 999 Monte Carlo replications with cases randomly distributed across space and found the proportion of replications with test statistics higher than that observed from the data. We plotted lineage-specific probability surfaces on individual maps. The analysis was conducted in R (53) using the *spatialkernel* package (107).

To account for potential confounders and detect geographic trends, we modeled the risk surface using a multinomial generalized additive model (GAM). We estimated the effect of a bivariate thin plate regression spline smooth of latitude and longitude on the odds of infection with a given lineage as compared to the most common lineage. We compared lineages IIa and IIb and the group of rare lineages separately to lineage Ib, the reference lineage. The multinomial analysis entailed logistic-type equations for each of the three lineage comparisons. The model was adjusted for sex and age group (<5, 5-9, 10-19, 20-59, and ≥60 years); isolates from cases of unknown age (n=1) or sex (n=10) were excluded from analysis. Parameters were estimated using restricted maximum likelihood (REML). We used the *mgcv* package in R (108, 109).

We conducted a series of sensitivity analyses to determine the robustness of our results to specific parameters used in each analysis, and we confirmed our results with two independent methods, Dixon's nearest-neighbor test (105) and multinomial spatial scan statistics (110) (Appendix).

Temporal Variation in Spatial Segregation

To determine if spatial segregation varies over time, we replicated analyses described above incorporating time (Appendix). For the kernel regression analysis, we tested if spatial segregation differed among three intervals (2005-2007, 2008-2010, and 2011-2014) and calculated a kernel-based estimate of spatial segregation for each interval. Statistical significance of this test was established using 999 Monte Carlo replications. We evaluated the impact of time in the multinomial GAM by adding year to the model as a continuous variable, testing the effect of year as both a linear term and as a smoothed term using a thin plate spline.

Exploratory Risk Factor Analysis

We explored potential drivers of segregation by testing the association of risk factors included on the DOH case report form with each lineage compared to the reference lineage Ib. Using multinomial GAMs with lineage Ib as reference and adjusting for sex, age, year, and latitude and longitude as a thin plate spline bivariate smoother, we tested each risk factor (Appendix Table S9). In addition to the statewide analyses, region-specific analyses were conducted for the three regions with the highest STEC O157:H7 incidence to determine locally important associations. Regions were defined based on important demographic factors, namely population centers and agricultural foci, and observed segregation clusters. Regional models were adjusted for sex, age, and year.

Results

Of the 1160 STEC O157:H7 cases reported to DOH during the study period, 49 were excluded from analysis. Isolates from six cases were biochemically atypical STEC O157:H7. Thirty-three isolates representing 31 PFGE types were not available for typing (Appendix). We SNP-typed 793 isolates, and matched, by extension, another 328 to a known lineage using

PFGE, allowing us to assign 1121 cases to infection with a specific lineage of STEC O157:H7. Ten cases lacked address data and were excluded, leaving a final sample size of 1111.

Lineages Ib, IIa, and IIb were, in descending order, the most common lineages (Table 4.1). Twelve rare lineages were identified, including two not previously described, encompassing 45 unique PFGE types. Lineages IIa and IIb contained an average of seven (standard deviation [SD] = 14) and eight (SD = 25) isolates per PFGE type, respectively, compared to three (SD = 5) for lineage Ib and one (SD = 2) for the rare lineages. Distribution of cases by sex, age group, and HUS status varied by lineage (Table 4.1). Lineages IIa and IIb isolates originated disproportionately from children <5-years-old compared to isolates in lineage Ib. Cases infected with lineage IIb bacteria also had higher frequencies of HUS (10%) than other cases (6%), and no rare STEC O157:H7 lineage isolates caused HUS.

Spatial Segregation

Diggle's kernel estimation test of spatial segregation was statistically significant ($p = 0.001$) (Figure 4.1). Areas of statistically significant spatial segregation were identified for lineages Ib, IIa, and IIb. Isolates from the southwest and coastal region were marked by segregation of lineage IIb and correspondingly less Ib self-association than expected. Spatial segregation of Ib was observed in the northwest corner of the State and of IIa in the south-central region, both areas lacking in lineage IIb self-association. Sensitivity analysis of alternate bandwidths corroborated these results (Appendix).

Consistent with the kernel regression results, the adjusted GAM risk surface of lineage IIb varied significantly from that of Ib ($p < 0.001$). The incidence of lineage IIb was greater than Ib in the southwest region and diminished as latitude and longitude increased (Figure 4.2). This spatial pattern was also observed in the kernel estimation map of lineage IIb (Figure 4.1). The risk surfaces of lineage IIa and the rare lineage group did not differ significantly from that of Ib (Appendix Table S6). In a series of sensitivity analyses designed to gauge the robustness of

results to model assumptions, the spatial risk surface of lineage IIb consistently varied significantly from the risk surface of lineage Ib (Appendix Table S6). The spatial risk surface of lineage IIa also varied significantly from the risk surface of lineage Ib in some sensitivity analyses, similar to the spatial distribution observed in the kernel estimation lineage IIa map.

Significant differences were also found in lineage by age of the infected case. The odds of being 20-59 years old or ≥60 years old (vs. <5 years old) were lower among IIa-infected cases than among Ib-infected cases [20-59 year odds ratio (OR) 0.65, 95% confidence interval (CI): 0.44, 0.96; ≥60 year OR=0.49, 95% CI: 0.28, 0.85). The odds of being 20-59 years vs. <5 years was also lower among lineage IIb-infected cases than among Ib-infected cases (OR=0.44, 95% CI: 0.28, 0.69). Thus, adults comprised a smaller proportion of cases infected with lineage IIa or IIb STEC O157:H7 than of those infected with lineage Ib. No significant differences were found by sex.

Temporal Variation

The incidence of STEC O157:H7 during the study period did not vary meaningfully (1.73 cases per 100,000 annually for 2005-2006, and 1.82 per 100,000 for 2013-2014), but the composition of the STEC O157:H7 population shifted over time (Figure 4.3). Specifically, the proportion of lineage Ib isolates fell from 59% (2005-2006) to 41 % (2013-2014). In the GAM analysis, incidence relative to lineage Ib increased over time for lineage IIa (OR=1.26, 95% CI: 1.19, 1.34), lineage IIb (OR=1.10, 95% CI: 1.03, 1.17), and rare lineages (OR=1.13, 95% CI: 1.02, 1.26).

A peak of lineage IIb incidence was observed during the middle of the study period in the southwest and Seattle-Tacoma regions (Figure 4.3). Using kernel regression, we identified statistically significant temporal variation in spatial segregation across intervals ($p = 0.001$). Statistically significant overall spatial segregation was observed only during the 2008-2010 interval ($p = 0.001$). Lineage IIb was segregated during all intervals, and lineages Ib and IIa

were segregated during 2008-2010 and 2011-2014 (Appendix Figures S2-S4). Cross-validated log-likelihood bandwidths used in these analyses ranged from 0.73 to 1.0. In sensitivity analysis, a lower bandwidth yielded statistically significant spatial segregation during all periods (Appendix). Latitude and longitude remained significant predictors of Ib in GAMs that included year (Appendix Table S6).

Alternate Methods & Bias Detection

Alternate analytic approaches confirmed the results of our primary analyses. Dixon's test for spatial segregation identified statistically significant spatial segregation overall and for lineages Ib, IIa, and IIb (Appendix Tables S7-S8). Three clusters identified using multinomial spatial scan statistics paralleled areas of segregation found in the kernel regression analysis and were consistent with the southwest trend toward proportionally greater IIb observed in the multinomial GAM (Appendix Figures S5-S6).

We also sought to understand the impact of person-to-person transmission (as determined by DOH investigators) on the results (Appendix). After discounting secondary transmission, we observed spatial segregation using the kernel estimation method ($p = 0.002$). The risk surface of lineage IIb still varied significantly from that of Ib ($p < 0.001$). The trend toward greater IIb relative to Ib risk in the southwest was consistent with the analysis of all cases, but relative IIb risk was substantially lower in the northeast than that observed in the primary analysis. This pattern suggests that IIb-infected cases in the northeast, but not the southwest, may be disproportionately due to secondary transmission compared to Ib-infected cases. Finally, we found no evidence of case ascertainment bias that could independently explain our results (Appendix).

Exploratory Risk Factor Analysis

Statewide, cases infected with lineage IIa STEC O157:H7 were more likely to have consumed raw fruits or vegetables than those infected with lineage Ib bacteria (OR=1.81, 95% CI: 1.05, 3.11). This trend was consistent, although not statistically significant, in all three regions. Cases in the southwest were also more likely to be infected with IIa than with Ib if they had consumed untreated or unchlorinated water (OR=4.49, 95% CI: 1.48, 13.57). This association was opposite in the south-central region (OR=0.16, 95% CI: 0.04, 0.63), suggesting the influence of local factors.

In the southwest region, cases infected with lineage IIb STEC O157:H7 were also more likely than those infected with lineage Ib bacteria to have consumed untreated or unchlorinated water (OR=3.76, 95% CI: 1.38, 10.28). However, IIb cases in the southwest region were less likely than Ib cases to have been exposed to recreational water (OR=0.38, 95% CI: 0.16, 0.93). In the statewide analysis, lineage IIb cases were more likely than Ib cases to have consumed raw milk (OR=2.46, 95% CI: 1.15, 5.28). This association was particularly apparent in the southwest region (OR=17.33, 95% CI: 2.05, 146.50). In the northwest region, IIb-infected cases were inversely associated with prime cut or ground beef exposures, and with having visited a zoo, farm, fair, or pet shop, compared to Ib-infected cases (Appendix Table S9). Similarly, cases associated with IIb bacteria in the south-central region were less likely than cases associated with Ib strains to have had any animal contact (OR=0.16, 95% CI: 0.03, 0.88).

Discussion

Our demonstration of geographic differences and temporal trends in the relative frequencies of particular lineages of STEC O157:H7 by spatial segregation identified areas in Washington State that perennially produce cases infected with lineage Ib, IIa, or IIb bacteria far in excess of what is expected based on statewide data. In all analyses, lineage IIb cases were segregated in the southwest region of the State. Suburbs of Portland, Oregon are situated in this region, and the major interstate connecting Seattle to Portland runs through it. Home to 16% of

Washington's population and only 7.1% of its cattle (50, 111), the southwest counties do not have intensive livestock production but do compose part of the range of Roosevelt elk. Elk elsewhere in the country have been identified as STEC carriers (112). Water is also a potential factor in STEC O157:H7 epidemiology in the southwest region, which has abundant coastal and river access. Spatial segregation of lineage IIb in the southwest was strongest from 2009 to 2012, but it was apparent during all periods. The largest recognized IIb outbreak in this region accounted for only 11 cases linked to a particular daycare, such that the segregation we observed is unlikely due to a single point source. Notably, lineage IIb has the greatest overlap with the putatively hypervirulent clade 8 (20), making its segregation of particular concern. Kernel regression also identified an area of IIa segregation in the south-central part of the State and Ib segregation in the north. These findings were supported by some of the sensitivity analyses. However, the Ib and IIa risk surfaces did not differ significantly and segregation during specific time periods was absent, suggesting that these two lineages might often overlap.

The presence of spatially segregated lineages indicate local environmental reservoirs producing infections above and beyond infections caused by widely dispersed sources such as food. Persistent spatial segregation of a particular lineage could reflect a founder effect, in which bacteria have established themselves in a local reservoir and persist, occasionally crossing over to humans. Such a dynamic would explain genetically similar bacteria isolated in the same general geographic region months or years apart from clinical cases, as was seen in two cases from Webster County, Missouri (113). It is also consistent with findings by Jaros et al., who found that geographic location explained a significant portion of variation in STEC O157:H7 strains in New Zealand (114). Localized transmission provides an opportunity for intervention if the reservoirs can be identified. Public health investigators may be able to target potential reservoirs, both wild and domestic, and identify how STEC O157:H7 is being transmitted humans. Longitudinal studies of how STEC bacteria are maintained over time in local reservoirs would shed additional light on potential public health interventions.

This study also shows that the composition of the STEC O157:H7 population shifted over time. While our work is consistent with previous findings that Ib is responsible for the majority of clinical cases in the United States (89), Washington State experienced statistically significant increases in all other lineages relative to Ib during the study period. The increase is most dramatic for lineage IIa, which appears to have emerged in most regions in the latter half of the study period (Figure 4.3). This could reflect the changing epidemiology of STEC O157:H7 discussed by Rivas et al, owing to changes in food sources and consumption, as well as pathogen evolution (92). Lineage IIa STEC O157:H7 have emerged as important causes of disease across the State, suggesting a disseminated driver of infections for this lineage overall. Lineage IIa's observed association with raw fruit and vegetable consumption, as compared to lineage Ib, is consistent with this hypothesis. Wang et al. (115) demonstrated differential resistance to the levels of chlorine used to wash commercial produce across phylogenetic clades, suggesting one potential mechanism for this association. The south-central region, identified in some analyses as an area of IIa segregation, experienced an uptick in IIa cases earlier than in other regions. This area is home to the Yakima Valley, an agriculture center, and a local IIa reservoir in this region could produce the observed segregation independent of statewide trends.

This study also suggests exposures that may be preferentially associated with particular lineages. Of note, we observed associations of lineage IIb with drinking untreated/unchlorinated water and raw milk in the southwest region where this lineage is segregated. There may be a lineage IIb reservoir in animals producing raw milk in this area, or bacteria from environmental reservoirs in the area may spill over into these animals and local water sources. Only one small raw milk outbreak in 2005 was noted on the DOH case report forms, making it unlikely that a single farm is responsible for the association. While exploratory, this hypothesis-generating analysis suggests that not all STEC O157:H7 is transmitted in the same way. It is possible that some lineages may be especially successful in surviving in particular vehicles or environments,

such as raw produce or unpasteurized milk or water. This is an area needing additional study and could expand opportunities for preventing disease.

Our study is limited somewhat by its reliance on SNP data to define phylogenetic lineages. Whole genome sequencing would have supported finer resolution of relatedness, particularly among isolates that were segregated in time and space, and enabled us to trace the history of segregated clusters, though would not necessarily alter our conclusions. However, our use of phylogenetic lineages rather than PFGE profiles is also a strength of the work, as PFGE does not put differences into an evolutionary perspective (116). By basing the analysis on phylogenetic lineages, we captured relatedness among strains and thus acknowledge that STEC O157:H7 continues to evolve as it circulates through its host populations. The lineages also provided a clear comparator for analysis in the historically dominant lineage in the United States, Ib (89), and we were able to incorporate all isolates in a meaningful way by grouping by lineage. The 355 distinct PFGE patterns would have yielded insufficient power for most analyses.

In summary, clusters of spatial segregation by phylogenetic lineage in Washington State suggest local reservoirs that perennially cause human disease. Further exploration of land use, human movements, and social-behavioral factors could elucidate within-region drivers of spatial segregation. Environmental risk assessment and longitudinal studies based on our findings could provide valuable information by identifying pathogen reservoirs that have not been identified by traditional public health surveillance and which could be mitigated by public health or environmental measures. The makeup of the STEC O157:H7 population in the State is also shifting. To manage emerging lineages, attention is needed to the heterogeneity in risk factors and virulence across the phylogenetic tree.

Figures & Tables

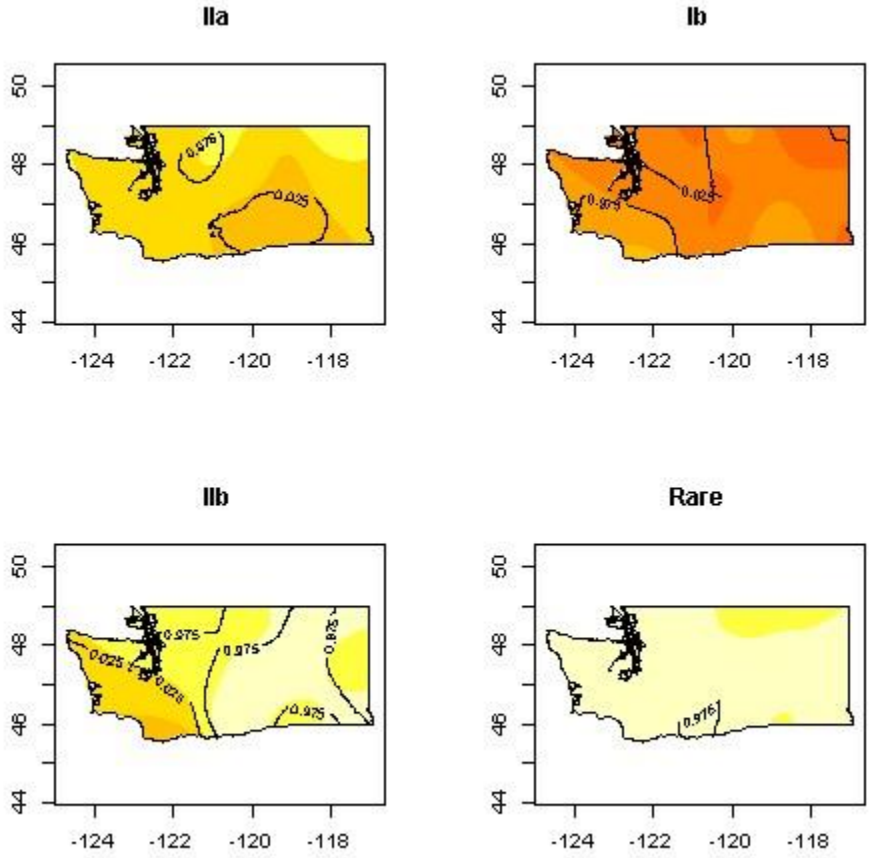

Figure 4.1. Kernel-based estimation of spatial segregation by lineage. Darker hues indicate greater segregation. Contour lines marked 0.025 define areas in which the given lineage is statistically significantly segregated. Contour lines marked 0.975 define areas in which the given lineage is statistically significantly less likely to be found in proximity to itself than to other lineages. N=1,111.

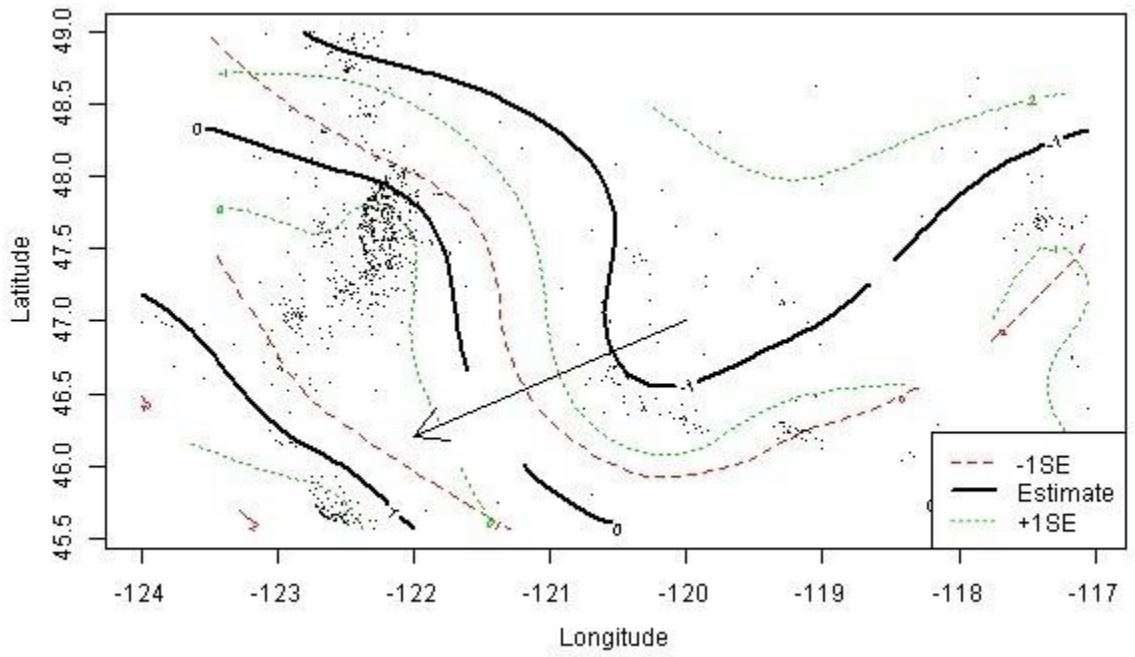

Figure 4.2. Risk surface of lineage IIb compared to lineage Ib. Risk surface was generated using a multinomial generalized additive model and a bivariate thin plate smooth function for longitude and latitude. The black contour lines indicate increased proportional incidence of lineage IIb toward the southwest corner of the area as compared to lineage Ib ($p<0.001$). The arrow indicates the general direction of the trend from higher Ib risk to higher IIb risk. N=1,100.

Abbreviation: SE, standard error

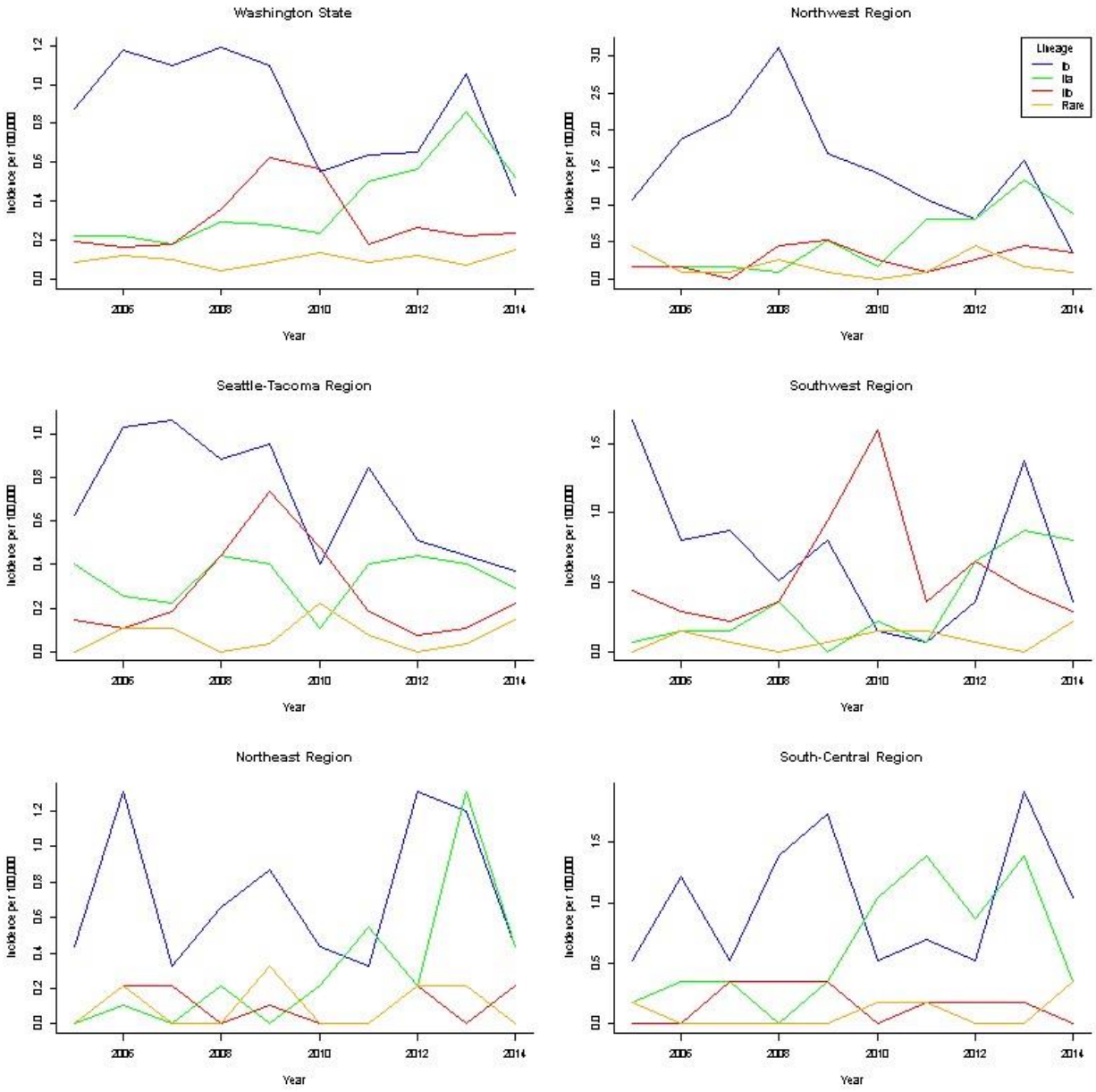

Figure 4.3. Annual incidence per 100,000 people of reported STEC O157:H7 cases by phylogenetic lineage, shown for Washington State and by region, 2005-2014. Regions were defined according to important demographic characteristics and patterns of segregation observed in analyses for the whole period. The Northwest Region experienced the highest peak incidence. The Seattle-Tacoma Region and the Northeast Region experienced the lowest incidences.

Table 4.1. Frequency of Reported STEC O157:H7 Case Characteristics by Phylogenetic Lineage, Washington State, 2005-2014

Variable	Lineage Ib	Lineage IIa	Lineage IIb	Rare lineage*
Total	586 (52.7%)	260 (23.4%)	199 (17.9%)	66 (5.9%)
PFGE† types‡	210 (65.8%)	38 (11.9%)	26 (8.2%)	45 (14.1%)
Mean isolates per PFGE type (SD†)	2.8 (5.3)	6.8 (14.3)	7.7 (24.7)	1.5 (1.7)
Sex				
Female	333 (56.8%)	163 (62.7%)	105 (52.8%)	33 (50.0%)
Male	244 (41.6%)	97 (37.3%)	94 (47.2%)	32 (48.5%)
Unknown	9 (1.5%)	0	0	1 (1.5%)
Age group				
<5 years	119 (20.3%)	72 (27.7%)	63 (31.7%)	10 (15.2%)
5-9 years	81 (13.8%)	32 (12.3%)	33 (16.6%)	12 (18.2%)
10-19 years	97 (16.6%)	51 (19.6%)	31 (15.6%)	6 (9.1%)
20-59 years	207 (35.3%)	81 (31.2%)	49 (24.6%)	29 (43.9%)
≥60 years	81 (13.8%)	24 (9.2%)	23 (11.6%)	9 (13.6%)
Unknown	1 (0.2%)	0	0	0
HUS†				
Yes	37 (6.3%)	18 (6.9%)	20 (10.0%)	0 (0%)
No	526 (89.2%)	236 (90.1%)	173 (86.1%)	67 (98.5%)
Unknown	27 (4.6%)	8 (3.1%)	8 (4.0%)	1 (1.5%)

*"Rare lineage" includes 12 different lineages.

†Abbreviations: HUS, hemolytic uremic syndrome; PFGE, pulsed field gel electrophoresis; SD, standard deviation.

‡PFGE type percentages indicate the proportion of PFGE types with an assigned lineage (n=355) belonging to each lineage.

References

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, & Griffin PM (2011) Foodborne illness acquired in the United States--major pathogens. *Emerging infectious diseases* 17(1):7-15.
2. Economic Research Service (2014) *Cost Estimates of Foodborne Illness*, (U.S. Department of Agriculture).
3. Karmali MA, Steele BT, Petric M, & Lim C (1983) Sporadic cases of haemolytic-uraemic syndrome associated with faecal cytotoxin and cytotoxin-producing Escherichia coli in stools. *Lancet* 1(8325):619-620.
4. Vanaja SK, Jandhyala DM, Mallick EM, Leong JM, & Balasubramanian S (2013) Chapter 5 - Enterohemorrhagic and other Shigatoxin-producing Escherichia coli. *Escherichia Coli (Second Edition)*, ed Donnenberg MS (Academic Press, Boston), pp 121-182.
5. Tarr PI, Gordon CA, & Chandler WL (2005) Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome. *Lancet* 365(9464):1073-1086.
6. Scheiring J, Andreoli SP, & Zimmerhackl LB (2008) Treatment and outcome of Shiga-toxin-associated hemolytic uremic syndrome (HUS). *Pediatric nephrology (Berlin, Germany)* 23(10):1749-1760.
7. Centers for Disease Control and Prevention (1993) Update: Multistate Outbreak of Escherichia coli O157:H7 Infections from Hamburgers -- Western United States, 1992-1993. *MMWR. Morbidity and mortality weekly report* 42(14):258-263.
8. National Center for Health Statistics (2012) Chapter 10: Food Safety. *Healthy People 2010 Final Review*, Hyattsville, MD).
9. Adams NL, Byrne L, Smith GA, Elson R, Harris JP, Salmon R, Smith R, O'Brien SJ, Adak GK, & Jenkins C (2016) Shiga Toxin-Producing Escherichia coli O157, England and Wales, 1983-2012. *Emerging infectious diseases* 22(4):590-597.
10. Pennington H (2010) Escherichia coli O157. *The Lancet* 376(9750):1428-1435.
11. Centers for Disease Control and Prevention (2011) Vital signs: incidence and trends of infection with pathogens transmitted commonly through food--foodborne diseases active surveillance network, 10 U.S. sites, 1996-2010. *MMWR. Morbidity and mortality weekly report* 60(22):749-755.
12. Healthy People 2020 (2016) Food Safety. (U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion, Washington, DC).
13. Centers for Disease Control and Prevention (2016) Summary of Notifiable Diseases, 2014. *MMWR. Morbidity and mortality weekly report* 63(54):1-154.
14. Marder EP, Cieslak PR, Cronquist AB, Dunn J, Lathrop S, Rabatsky-Ehr T, Ryan P, Smith K, Tobin D'Angelo M, Vugia DJ, Zansky S, Holt KG, Wolpert BJ, Lynch M, Tauxe R, & Geissler AL (2017) Incidence and Trends of Infections with Pathogens Transmitted Commonly Through Food and the Effect of Increasing Use of Culture-Independent Diagnostic Tests on Surveillance — Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2013–2016. *MMWR. Morbidity and mortality weekly report* 66(15):397-403.
15. Gould LH, Mody RK, Ong KL, Clogher P, Cronquist AB, Garman KN, Lathrop S, Medus C, Spina NL, Webb TH, White PL, Wymore K, Gierke RE, Mahon BE, & Griffin PM (2013) Increased recognition of non-O157 Shiga toxin-producing Escherichia coli infections in the United States during 2000-2010: epidemiologic features and comparison with E. coli O157 infections. *Foodborne pathogens and disease* 10(5):453-460.
16. Hadler JL, Clogher P, Hurd S, Phan Q, Mandour M, Bemis K, & Marcus R (2011) Ten-year trends and risk factors for non-O157 Shiga toxin-producing Escherichia coli found

through Shiga toxin testing, Connecticut, 2000-2009. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 53(3):269-276.
17. Centers for Disease Control (2017) *Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet 2015 Surveillance Report (Final Data)* (Atlanta, Georgia), (U.S. Department of Health and Human Services C).
18. Rangel JM, Sparling PH, Crowe C, Griffin PM, & Swerdlow DL (2005) Epidemiology of Escherichia coli O157:H7 outbreaks, United States, 1982-2002. *Emerging infectious diseases* 11(4):603-609.
19. Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, Adak B, Willshaw G, Cheasty T, Green J, Dougan G, Parkhill J, & Wain J (2013) Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. *Journal of clinical microbiology* 51(1):232-237.
20. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, Mladonicky JM, Somsel P, Rudrik JT, Dietrich SE, Zhang W, Swaminathan B, Alland D, & Whittam TS (2008) Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks. *Proceedings of the National Academy of Sciences of the United States of America* 105(12):4868-4873.
21. Iyoda S, Manning SD, Seto K, Kimata K, Isobe J, Etoh Y, Ichihara S, Migita Y, Ogata K, Honda M, Kubota T, Kawano K, Matsumoto K, Kudaka J, Asai N, Yabata J, Tominaga K, Terajima J, Morita-Ishihara T, Izumiya H, Ogura Y, Saitoh T, Iguchi A, Kobayashi H, Hara-Kudo Y, Ohnishi M, Arai R, Kawase M, Asano Y, Asoshima N, Chiba K, Furukawa I, Kuroki T, Hamada M, Harada S, Hatakeyama T, Hirochi T, Sakamoto Y, Hiroi M, Takashi K, Horikawa K, Iwabuchi K, Kameyama M, Kasahara H, Kawanishi S, Kikuchi K, Ueno H, Kitahashi T, Kojima Y, Konishi N, Obata H, Kai A, Kono T, Kurazono T, Matsumoto M, Matsumoto Y, Nagai Y, Naitoh H, Nakajima H, Nakamura H, Nakane K, Nishi K, Saitoh E, Satoh H, Takamura M, Shiraki Y, Tanabe J, Tanaka K, Tokoi Y, & Yatsuyanagi J (2014) Phylogenetic Clades 6 and 8 of Enterohemorrhagic Escherichia coli O157:H7 With Particular stx Subtypes are More Frequently Found in Isolates From Hemolytic Uremic Syndrome Patients Than From Asymptomatic Carriers. *Open forum infectious diseases* 1(2):ofu061.
22. Soderlund R, Jernberg C, Ivarsson S, Hedenstrom I, Eriksson E, Bongcam-Rudloff E, & Aspan A (2014) Molecular typing of Escherichia coli O157:H7 isolates from Swedish cattle and human cases: population dynamics and virulence. *Journal of clinical microbiology* 52(11):3906-3912.
23. Bono JL, Smith TP, Keen JE, Harhay GP, McDaneld TG, Mandrell RE, Jung WK, Besser TE, Gerner-Smidt P, Bielaszewska M, Karch H, & Clawson ML (2012) Phylogeny of Shiga toxin-producing Escherichia coli O157 isolated from cattle and clinically ill humans. *Molecular biology and evolution* 29(8):2047-2062.
24. Yokoyama E, Hirai S, Hashimoto R, & Uchimura M (2012) Clade analysis of enterohemorrhagic Escherichia coli serotype O157:H7/H- strains and hierarchy of their phylogenetic relationships. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 12(8):1724-1728.
25. Eppinger M, Mammel MK, Leclerc JE, Ravel J, & Cebula TA (2011) Genomic anatomy of Escherichia coli O157:H7 outbreaks. *Proceedings of the National Academy of Sciences of the United States of America* 108(50):20142-20147.
26. Eppinger M, Mammel MK, Leclerc JE, Ravel J, & Cebula TA (2011) Genome signatures of Escherichia coli O157:H7 isolates from the bovine host reservoir. *Applied and environmental microbiology* 77(9):2916-2925.
27. Jung WK, Bono JL, Clawson ML, Leopold SR, Shringi S, & Besser TE (2013) Lineage and genogroup-defining single nucleotide polymorphisms of Escherichia coli O157:H7. *Applied and environmental microbiology* 79(22):7036-7041.

28. Garg AX, Suri RS, Barrowman N, Rehman F, Matsell D, Rosas-Arellano MP, Salvadori M, Haynes RB, & Clark WF (2003) Long-term Renal Prognosis of Diarrhea-Associated Hemolytic Uremic Syndrome: A Systematic Review, Meta-analysis, and Meta-regression. *Jama* 290(10):1360-1370.
29. Gould LH, Demma L, Jones TF, Hurd S, Vugia DJ, Smith K, Shiferaw B, Segler S, Palmer A, Zansky S, & Griffin PM (2009) Hemolytic uremic syndrome and death in persons with Escherichia coli O157:H7 infection, foodborne diseases active surveillance network sites, 2000-2006. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 49(10):1480-1485.
30. Mody RK, Gu W, Griffin PM, Jones TF, Rounds J, Shiferaw B, Tobin-D'Angelo M, Smith G, Spina N, Hurd S, Lathrop S, Palmer A, Boothe E, Luna-Gierke RE, & Hoekstra RM (2015) Postdiarrheal hemolytic uremic syndrome in United States children: clinical spectrum and predictors of in-hospital death. *The Journal of pediatrics* 166(4):1022-1029.
31. Crim SM, Griffin PM, Tauxe R, Marder EP, Gilliss D, Cronquist AB, Cartter M, Tobin D'Angelo M, Blythe D, Smith K, Lathrop S, Zansky S, Cieslak PR, Dunn J, Holt KG, Wolpert B, & Henao OL (2015) Preliminary Incidence and Trends of Infection with Pathogens Transmitted Commonly Through Food — Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2006–2014. *MMWR. Morbidity and mortality weekly report* 64(18):495-498.
32. Crim SM, Iwamoto M, Huang JY, Griffin PM, Gilliss D, Cronquist AB, Cartter M, Tobin-D'Angelo M, Blythe D, Smith K, Lathrop S, Zansky S, Cieslak PR, Dunn J, Holt KG, Lance S, Tauxe R, Henao OL, Centers for Disease C, & Prevention (2014) Incidence and trends of infection with pathogens transmitted commonly through food--Foodborne Diseases Active Surveillance Network, 10 U.S. sites, 2006-2013. *MMWR. Morbidity and mortality weekly report* 63(15):328-332.
33. Davis TK, McKee R, Schnadower D, & Tarr PI (2013) Treatment of Shiga toxin-producing Escherichia coli infections. *Infectious disease clinics of North America* 27(3):577-597.
34. Ahn CK, Holt NJ, & Tarr PI (2009) Shiga-Toxin Producing Escherichia coli and the Hemolytic Uremic Syndrome: What Have We Learned in the Past 25 Years? 634:1-17.
35. Holtz LR, Neill MA, & Tarr PI (2009) Acute Bloody Diarrhea: A Medical Emergency for Patients of All Ages. *Gastroenterology* 136(6):1887-1898.
36. Ardissino G, Possenti I, Tel F, Testa S, & Paglialonga F (2014) Time to change the definition of hemolytic uremic syndrome. *European journal of internal medicine* 25(2):e29.
37. Balestracci A, Martin SM, & Toledo I (2015) Hemoconcentration in hemolytic uremic syndrome: time to review the standard case definition? *Pediatric nephrology (Berlin, Germany)* 30(2):361.
38. Wong CS, Mooney JC, Brandt JR, Staples AO, Jelacic S, Boster DR, Watkins SL, & Tarr PI (2012) Risk factors for the hemolytic uremic syndrome in children infected with Escherichia coli O157:H7: a multivariable analysis. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 55(1):33-41.
39. Wong CS, Jelacic S, Habeeb RL, Watkins SL, & Tarr PI (2000) The risk of the hemolytic-uremic syndrome after antibiotic treatment of Escherichia coli O157:H7 infections. *The New England journal of medicine* 342(26):1930-1936.
40. Ake JA, Jelacic S, Ciol MA, Watkins SL, Murray KF, Christie DL, Klein EJ, & Tarr PI (2005) Relative nephroprotection during Escherichia coli O157:H7 infections: association with intravenous volume expansion. *Pediatrics* 115(6):e673-680.

41. Klein EJ, Stapp JR, Clausen CR, Boster DR, Wells JG, Qin X, Swerdlow DL, & Tarr PI (2002) Shiga toxin-producing Escherichia coli in children with diarrhea: a prospective point-of-care study. *The Journal of pediatrics* 141(2):172-177.
42. Freedman SB, Eltorki M, Chui L, Xie J, Feng S, MacDonald J, Dixon A, Ali S, Louie M, Lee BE, Osterreicher L, & Thull-Freedman J (2017) Province-Wide Review of Pediatric Shiga Toxin-Producing Escherichia coli Case Management. *The Journal of pediatrics* 180:184-190 e181.
43. Bielaszewska M, Kock R, Friedrich AW, von Eiff C, Zimmerhackl LB, Karch H, & Mellmann A (2007) Shiga toxin-mediated hemolytic uremic syndrome: time to change the diagnostic paradigm? *PloS one* 2(10):e1024.
44. Bielaszewska M, Friedrich AW, Aldick T, Schürk-Bulgrin R, & Karch H (2006) Shiga Toxin Activatable by Intestinal Mucus in *Escherichia coli* Isolated from Humans: Predictor for a Severe Clinical Outcome. *Clinical Infectious Diseases* 43:1160-1167.
45. Council of State and Territorial Epidemiologists (2009) Public Health Reporting and National Notification for Hemolytic Uremic Syndrome (post-diarrheal). (Atlanta, Georgia).
46. Mody RK, Luna-Gierke RE, Jones TF, Comstock N, Hurd S, Scheftel J, Lathrop S, Smith G, Palmer A, Strockbine N, Talkington D, Mahon BE, Hoekstra RM, & Griffin PM (2012) Infections in pediatric postdiarrheal hemolytic uremic syndrome: factors associated with identifying shiga toxin-producing Escherichia coli. *Archives of pediatrics & adolescent medicine* 166(10):902-909.
47. Ong KL, Apostal M, Comstock N, Hurd S, Webb TH, Mickelson S, Scheftel J, Smith G, Shiferaw B, Boothe E, & Gould LH (2012) Strategies for surveillance of pediatric hemolytic uremic syndrome: Foodborne Diseases Active Surveillance Network (FoodNet), 2000-2007. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 54 Suppl 5:S424-431.
48. Meites S & Buffone GJ (1989) *Pediatric Clinical Chemistry: Reference (Normal) Values* (AACC Press, Washington, D.C.).
49. Anonymous (2009) *The Harriet Lane Handbook: A Manual for Pediatric House Officers* (Elsevier Mosby, Philadelphia, PA) 18th Ed.
50. U. S. Census Bureau (2012) 2012 TIGER/Line Shapefiles (machine-readable data files). (U.S. Census Bureau).
51. Dundas S, Todd WT, Stewart AI, Murdoch PS, Chaudhuri AK, & Hutchinson SJ (2001) The central Scotland Escherichia coli O157:H7 outbreak: risk factors for the hemolytic uremic syndrome and death among hospitalized patients. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 33(7):923-931.
52. Smith KE, Wilker PR, Reiter PL, Hedican EB, Bender JB, & Hedberg CW (2012) Antibiotic treatment of Escherichia coli O157 infection and the risk of hemolytic uremic syndrome, Minnesota. *The Pediatric infectious disease journal* 31(1):37-41.
53. R Core Team (2015) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).
54. Davis TK, Van De Kar NC, & Tarr PI (2014) Shiga Toxin/Verocytotoxin-Producing Escherichia coli Infections: Practical Clinical Perspectives. *Microbiol Spectr* 2(4):EHEC-0025-2014.
55. Tarr PI, Neill MA, Allen J, Siccardi CJ, Watkins SL, & Hickman RO (1989) The increasing incidence of the hemolytic-uremic syndrome in King County, Washington: lack of evidence for ascertainment bias. *American journal of epidemiology* 129(3):582-586.
56. Tarr PI & Hickman RO (1987) Hemolytic uremic syndrome epidemiology: a population-based study in King County, Washington, 1971 to 1980. *Pediatrics* 80(1):41-45.

57. Freedman SB, Xie J, Neufeld MS, Hamilton WL, Hartling L, Tarr PI, Alberta Provincial Pediatric Enteric Infection T, Nettel-Aguirre A, Chuck A, Lee B, Johnson D, Currie G, Talbot J, Jiang J, Dickinson J, Kellner J, MacDonald J, Svenson L, Chui L, Louie M, Lavoie M, Eltorki M, Vanderkooi O, Tellier R, Ali S, Drews S, Graham T, & Pang XL (2016) Shiga Toxin-Producing Escherichia coli Infection, Antibiotics, and Risk of Developing Hemolytic Uremic Syndrome: A Meta-analysis. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 62(10):1251-1258.
58. Chandler WL, Jelacic S, Boster DR, Ciol MA, Williams GD, Watkins SL, Igarashi T, & Tarr PI (2002) Prothrombotic coagulation abnormalities preceding the hemolytic-uremic syndrome. *The New England journal of medicine* 346(1):23-32.
59. Broadstock M, Michie S, & Marteau T (2000) Psychological consequences of predictive genetic testing: a systematic review. *European Journal of Human Genetics* 8:731-738.
60. Fulda KG & Lykens K (2006) Ethical issues in predictive genetic testing: a public health perspective. *J Med Ethics* 32(3):143-147.
61. Joly Y, Ngueng Feze I, & Simard J (2013) Genetic discrimination and life insurance: a systematic review of the evidence. *BMC medicine* 11(25).
62. Banatvala N, Griffin PM, Greene KD, Barrett TJ, Bibb WF, Green JH, & Wells JG (2001) The United States National Prospective Hemolytic Uremic Syndrome Study: microbiologic, serologic, clinical, and epidemiologic findings. *The Journal of infectious diseases* 183(7):1063-1070.
63. Friedrich AW, Bielaszewska M, Zhang WL, Pulz M, Kuczius T, Ammon A, & Karch H (2002) Escherichia coli harboring Shiga toxin 2 gene variants: frequency and association with clinical symptoms. *The Journal of infectious diseases* 185(1):74-84.
64. Persson S, Olsen KE, Ethelberg S, & Scheutz F (2007) Subtyping method for Escherichia coli shiga toxin (verocytotoxin) 2 variants and correlations to clinical manifestations. *Journal of clinical microbiology* 45(6):2020-2024.
65. Eklund M, Leino K, & Siitonen A (2002) Clinical Escherichia coli strains carrying stx genes: stx variants and stx-positive virulence profiles. *Journal of clinical microbiology* 40(12):4585-4593.
66. Luna-Gierke RE, Griffin PM, Gould LH, Herman K, Bopp CA, Strockbine N, & Mody RK (2014) Outbreaks of non-O157 Shiga toxin-producing Escherichia coli infection: USA. *Epidemiology and infection* 142(11):2270-2280.
67. Abu-Ali GS, Ouellette LM, Henderson ST, Lacher DW, Riordan JT, Whittam TS, & Manning SD (2010) Increased adherence and expression of virulence genes in a lineage of Escherichia coli O157:H7 commonly associated with human infections. *PloS one* 5(4):e10167.
68. Neupane M, Abu-Ali GS, Mitra A, Lacher DW, Manning SD, & Riordan JT (2011) Shiga toxin 2 overexpression in Escherichia coli O157:H7 strains associated with severe human disease. *Microbial pathogenesis* 51(6):466-470.
69. Pianciola L, Chinen I, Mazzeo M, Miliwebsky E, Gonzalez G, Muller C, Carbonari C, Navello M, Zitta E, & Rivas M (2014) Genotypic characterization of Escherichia coli O157:H7 strains that cause diarrhea and hemolytic uremic syndrome in Neuquen, Argentina. *International journal of medical microbiology : IJMM* 304(3-4):499-504.
70. Amigo N, Zhang Q, Amadio A, Zhang Q, Silva WM, Cui B, Chen Z, Larzabal M, Bei J, & Cataldi A (2016) Overexpressed Proteins in Hypervirulent Clade 8 and Clade 6 Strains of Escherichia coli O157:H7 Compared to E. coli O157:H7 EDL933 Clade 3 Strain. *PloS one* 11(11):e0166883.
71. Haugum K, Brandal LT, Lobersli I, Kapperud G, & Lindstedt BA (2011) Detection of virulent Escherichia coli O157 strains using multiplex PCR and single base sequencing for SNP characterization. *Journal of applied microbiology* 110(6):1592-1600.

72. Shringi S, Schmidt C, Katherine K, Brayton KA, Hancock DD, & Besser TE (2012) Carriage of stx2a differentiates clinical and bovine-biased strains of Escherichia coli O157. *PloS one* 7(12):e51572.
73. Rogers MF, Rutherford GW, Alexander SR, DiLiberti JH, Foster L, Schonberger LB, & Hurwitz ES (1986) A population-based study of hemolytic-uremic syndrome in Oregon, 1979-1982. *American journal of epidemiology* 123(1):137-142.
74. Taylor DN, Echeverria P, Sethabutr O, Pitarangsi C, Leksomboon U, Blacklow NR, Rowe B, Gross R, & Cross J (1988) Clinical and microbiologic features of Shigella and enteroinvasive Escherichia coli infections detected by DNA hybridization. *Journal of clinical microbiology* 26(7):1362-1366.
75. Ostroff SM, Kobayashi JM, & Lewis JH (1989) Infections with Escherichia coli O157:H7 in Washington State. The first year of statewide disease surveillance. *Jama* 262(3):355-359.
76. Reiss G, Kunz P, Koin D, & Keeffe EB (2006) Escherichia coli O157:H7 infection in nursing homes: review of literature and report of recent outbreak. *Journal of the American Geriatrics Society* 54(4):680-684.
77. Rowe PC, Orrbine E, Wells GA, & McLaine PN (1991) Epidemiology of hemolytic-uremic syndrome in Canadian children from 1986 to 1988. The Canadian Pediatric Kidney Disease Reference Centre. *The Journal of pediatrics* 119(2):218-224.
78. Rivas M, Sosa-Estani S, Rangel J, Caletti MG, Valles P, Roldan CD, Balbi L, Marsano de Mollar MC, Amoedo D, Miliwebsky E, Chinen I, Hoekstra RM, Mead P, & Griffin PM (2008) Risk factors for sporadic Shiga toxin-producing Escherichia coli infections in children, Argentina. *Emerging infectious diseases* 14(5):763-771.
79. Al-Jader L, Salmon RL, Walker AM, Williams HM, Willshaw GA, & Cheasty T (1999) Outbreak of Escherichia coli O157 in a nursery: lessons for prevention. *Archives of disease in childhood* 81(1):60-63.
80. Imai K, Keele L, & Yamamoto T (2010) Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science* 25(1):51-71.
81. Robins JM & Greenland S (1992) Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology (Cambridge, Mass.)* 3(2):143-155.
82. Pearl J (2001) Direct and Indirect Effects. in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco), pp 411-420.
83. Cole SR & Hernán MA (2002) Fallibility in estimating direct effects. *International journal of epidemiology* 31:163-165.
84. Richiardi L, Bellocco R, & Zugna D (2013) Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology* 42(5):1511-1519.
85. Imai K, Keele L, & Tingley D (2010) A general approach to causal mediation analysis. *Psychol Methods* 15(4):309-334.
86. Tingley D, Yamamoto T, Hirose K, Keele L, & Imai K (2014) mediation: **R** Package for Causal Mediation Analysis. *Journal of Statistical Software* 59(5):1-38.
87. Ogura Y, Mondal SI, Islam MR, Mako T, Arisawa K, Katsura K, Ooka T, Gotoh Y, Murase K, Ohnishi M, & Hayashi T (2015) The Shiga toxin 2 production level in enterohemorrhagic Escherichia coli O157:H7 is correlated with the subtypes of toxin-encoding phage. *Sci Rep* 5:16663.
88. Amigo N, Mercado E, Bentancor A, Singh P, Vilte D, Gerhardt E, Zotta E, Ibarra C, Manning SD, Larzabal M, & Cataldi A (2015) Clade 8 and Clade 6 Strains of Escherichia coli O157:H7 from Cattle in Argentina have Hypervirulent-Like Phenotypes. *PloS one* 10(6):e0127710.
89. Mellor GE, Fegan N, Gobius KS, Smith HV, Jennison AV, D'Astek BA, Rivas M, Shringi S, Baker KN, & Besser TE (2015) Geographically Distinct Escherichia coli O157 Isolates

Differ by Lineage, Shiga Toxin Genotype, and Total Shiga Toxin Production. *Journal of clinical microbiology* 53(2):579-586.
90. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, Dopfer D, Fazil A, Fischer-Walker CL, Hald T, Hall AJ, Keddy KH, Lake RJ, Lanata CF, Torgerson PR, Havelaar AH, & Angulo FJ (2015) World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis. *PLoS Med* 12(12):e1001921.
91. Majowicz SE, Scallan E, Jones-Bitton A, Sargeant JM, Stapleton J, Angulo FJ, Yeung DH, & Kirk MD (2014) Global incidence of human Shiga toxin-producing Escherichia coli infections and deaths: a systematic review and knowledge synthesis. *Foodborne pathogens and disease* 11(6):447-455.
92. Rivas M, Chinen I, Miliwebsky E, & Masana M (2014) Risk Factors for Shiga Toxin-Producing Escherichia coli-Associated Human Diseases. *Microbiol Spectr* 2(5).
93. Denno DM, Keene WE, Hutter CM, Koepsell JK, Patnode M, Flodin-Hursh D, Stewart LK, Duchin JS, Rasmussen L, Jones R, & Tarr PI (2009) Tri-county comprehensive assessment of risk factors for sporadic reportable bacterial enteric infection in children. *The Journal of infectious diseases* 199(4):467-476.
94. Luffman I & Tran L (2014) Risk Factors for E. coli O157 and Cryptosporidiosis Infection in Individuals in the Karst Valleys of East Tennessee, USA. *Geosciences* 4(3):202-218.
95. Liebana E, Smith RP, Batchelor M, McLaren I, Cassar C, Clifton-Hadley FA, & Paiba GA (2005) Persistence of Escherichia coli O157 isolates on bovine farms in England and Wales. *Journal of clinical microbiology* 43(2):898-902.
96. LeJeune JT, Besser TE, Rice DH, Berg JL, Stilborn RP, & Hancock DD (2004) Longitudinal Study of Fecal Shedding of Escherichia coli O157:H7 in Feedlot Cattle: Predominance and Persistence of Specific Clonal Types despite Massive Cattle Population Turnover. *Applied and environmental microbiology* 70(1):377-384.
97. Cernicchiaro N, Pearl DL, McEwen SA, Harpster L, Homan HJ, Linz GM, & Lejeune JT (2012) Association of wild bird density and farm management factors with the prevalence of E. coli O157 in dairy herds in Ohio (2007-2009). *Zoonoses and public health* 59(5):320-329.
98. Rosales-Castillo JA, Vazquez-Garciduenas MS, Alvarez-Hernandez H, Chassin-Noria O, Varela-Murillo AI, Zavala-Paramo MG, Cano-Camacho H, & Vazquez-Marrufo G (2011) Genetic diversity and population structure of Escherichia coli from neighboring small-scale dairy farms. *Journal of microbiology (Seoul, Korea)* 49(5):693-702.
99. Herbert LJ, Vali L, Hoyle DV, Innocent G, McKendrick IJ, Pearce MC, Mellor D, Porphyre T, Locking M, Allison L, Hanson M, Matthews L, Gunn GJ, Woolhouse ME, & Chase-Topping ME (2014) E. coli O157 on Scottish cattle farms: evidence of local spread and persistence using repeat cross-sectional data. *BMC veterinary research* 10:95.
100. Widgren S, Soderlund R, Eriksson E, Fasth C, Aspan A, Emanuelson U, Alenius S, & Lindberg A (2015) Longitudinal observational study over 38 months of verotoxigenic Escherichia coli O157:H7 status in 126 cattle herds. *Preventive veterinary medicine* 121(3-4):343-352.
101. Saxena T, Kaushik P, & Krishna Mohan M (2015) Prevalence of E. coli O157:H7 in water sources: an overview on associated diseases, outbreaks and detection methods. *Diagnostic microbiology and infectious disease* 82(3):249-264.
102. Barker J, Humphrey TJ, & Brown MWR (1999) Surivival of *Escherichia coli* O157 in a soil protozoan: implications for disease. *FEMS microbiology letters* 173:291-295.
103. Gargiulo A, Russo TP, Schettini R, Mallardo K, Calabria M, Menna LF, Raia P, Pagnini U, Caputo V, Fioretti A, & Dipineto L (2014) Occurrence of enteropathogenic bacteria in urban pigeons (Columba livia) in Italy. *Vector borne and zoonotic diseases (Larchmont, N.Y.)* 14(4):251-255.

104. U. S. Census Bureau (2015) 2010 Census Urban and Rural Classification and Urban Area Criteria.
105. Dixon PM (2002) Nearest-neighbor contingency table analysis of spatial segregation for several species. *Ecoscience* 9(2):142-151.
106. Diggle PJ, Zheng P, & Durr P (2005) Nonparametric estimation of spatial segregation in amultivariate point process: bovine tuberculosis in Cornwall, UK. *Appl Statist* 54(Part 3):645-658.
107. Zheng P & Diggle PJ (2013) spatialkernel: Nonparametric estimation of spatial segregation in a multivariate point process.).
108. Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1):3-36.
109. Wood SN (2003) Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* 65(1):95-114.
110. Jung I, Kulldorff M, & Richard OJ (2010) A spatial scan statistic for multinomial data. *Statistics in medicine* 29(18):1910-1918.
111. United States Department of Agriculture (2014) *2012 Census of Agriculture: Washington* (National Agricultural Statistics Service), (Agriculture USDo).
112. Franklin AB, Vercauteren KC, Maguire H, Cichon MK, Fischer JW, Lavelle MJ, Powell A, Root JJ, & Scallan E (2013) Wild ungulates as disseminators of Shiga toxin-producing Escherichia coli in urban areas. *PloS one* 8(12):e81512.
113. Turabelidze G, Lawrence SJ, Gao H, Sodergren E, Weinstock GM, Abubucker S, Wylie T, Mitreva M, Shaikh N, Gautom R, & Tarr PI (2013) Precise dissection of an Escherichia coli O157:H7 outbreak by single nucleotide polymorphism analysis. *Journal of clinical microbiology* 51(12):3950-3954.
114. Jaros P, Cookson AL, Campbell DM, Duncan GE, Prattley D, Carter P, Besser TE, Shringi S, Hathaway S, Marshall JC, & French NP (2014) Geographic Divergence of Bovine and Human Shiga Toxin–ProducingEscherichia coliO157:H7 Genotypes, New Zealand1. *Emerging infectious diseases* 20(12):1980-1989.
115. Wang S, Deng K, Zaremba S, Deng X, Lin C, Wang Q, Tortorello ML, & Zhang W (2009) Transcriptomic response of Escherichia coli O157:H7 to oxidative stress. *Applied and environmental microbiology* 75(19):6110-6123.
116. Leopold SR, Magrini V, Holt NJ, Shaikh N, Mardis ER, Cagno J, Ogura Y, Iguchi A, Hayashi T, Mellmann A, Karch H, Besser TE, Sawyer SA, Whittam TS, & Tarr PI (2009) A precise reconstruction of the emergence and constrained radiations of Escherichia coli O157 portrayed by backbone concatenomic analysis. *Proceedings of the National Academy of Sciences of the United States of America* 106(21):8713-8718.
117. VanderWeele TJ (2016) Mediation Analysis: A Practitioner's Guide. *Annual review of public health* 37:17-32.
118. Petersen ML, Sinisi SE, & van der Laan MJ (2006) Estimation of direct causal effects. *Epidemiology (Cambridge, Mass.)* 17(3):276-284.
119. Sofrygin O, van der laan MJ, & Neugebauer R (2016) simcausal: Simulating Longitudinal Data with Causal Inference Applications).
120. de la Cruz M (2008) Metodos para analizar datos puntuales. *Introduccion al Analisis Espacial de Datos en Ecologia y Ciencias Ambientales: Metodos y Aplicaciones*, eds Maestre FT, Escudero A, & Bonet A (Asociacion Espanola de Ecologia Terrestre, Universidad Rey Juan Carlos and Caja de Ahorros del Mediterraneo, Madrid), pp 76-127.
121. Kulldorff M & Information Management Services I (2009) SaTScan™ v8.0: Software for the spatial and space-time scan statistics).
122. Communicable Disease Epidemiology Section (2014) *Communicable Disease Report 2014* (Washington State Department of Health, Shoreline, WA), (Health WSDo).

123. Communicable Disease Epidemiology Section (2009) *Communicable Disease Report 2009* (Washington State Department of Health, Shoreline, WA), (Health WSDo).
124. Brundsdon C & Chen H (2014) GISTools: Some further GIS capabilities for R).

Appendix

Supplemental Methods & Results

Chapter 2: Hospital Records for HUS Validation

To validate HUS status, we sought records from all hospitals listed on the CRF for the STEC O157:H7 visit. Many cases were admitted at multiple hospitals, and charts from more than one institution were reviewed in some situations. Washington residents hospitalized in Oregon were reported through their Washington local health jurisdiction of residence and were included in this review.

STEC O157:H7 hospitalizations occurred across 71 facilities, with each institution treating between 1 and 87 cases. Children's hospitals, as identified by the Children's Hospital Association (www.childrenshospitals.org), treated 136 of the cases.

Chapter 2: Definition Validity and Cases with Discrepant HUS Status

Relative to the SCD, CSTE definition sensitivity was 96% [95% confidence interval (CI): 89%, 99%] and specificity was 77% (95% CI 72%, 81%) (Table 2.3). Among the three STEC O157:H7 cases meeting the SCD but not the CSTE definition for HUS (Appendix, Table S2), two were two years of age and did not meet the 1.0 mg/dL serum creatinine criteria and had no urinalysis documented to establish hematuria or proteinuria. One had no documented peripheral blood smear and HUS developed >3 weeks after diarrhea onset. They would have met the probable CSTE definition if only one of the microangiopathic changes or timing criteria were not met, but not if neither were met (Table 2.1) (45). Eighty-one STEC O157:H7 cases were defined as HUS by CSTE criteria but not by SCD. The largest share (52 or 64%) of these discrepant cases would not have been considered HUS if the CSTE definition included criteria of <150,000 platelets/mm^3 for thrombocytopenia. Of the remaining 29 discrepant cases, 26 had serum creatinine concentrations that did not exceed the age-specific normal values (48) but had

hematuria and/or proteinuria. Two had elevated serum creatinine concentrations but did not meet the SCD criteria for anemia, and one met all criteria but on different days.

The HCD had the lowest sensitivity but highest specificity (Table 2.3), identifying only 74% of HUS cases (95% CI 62%, 83%) but 99% (95% CI 98%, 100%) of non-HUS cases. The 20 SCD-based HUS cases not identified using this definition (Appendix, Table S2) either lacked smear evidence of hemolysis or did not meet the necessary creatinine serum concentrations. Two cases were defined as HUS by the HCD but not by the SCD.

Sensitivity of the CRF was 86% (95% CI 76%, 93%) (Table 2.3). Eight of the 10 missed HUS cases (Appendix, Table S2) were diagnosed by their hospital clinician as having HUS. Specificity was 94% (95% CI 91%, 96%), and there was strong agreement between the CRF and hospital diagnosis.

Hospital diagnosis exhibited the high sensitivity (97%; 95% CI 91%, 100%) (Table 3). Only two HUS cases were missed by hospital providers (Appendix, Table S2). However, specificity of the hospital diagnosis was only 94% (95% CI 91%, 97%). Nineteen STEC O157:H7 cases were diagnosed with HUS without fulfilling the SCD criteria. Twelve (63%) of these patients were <10 years-old and did not meet age-specific criteria for creatinine serum concentrations (48).

Reflecting the low specificity of the CSTE definition but accruing more true positive HUS cases, combining the CSTE definition and hospital diagnosis to emulate common surveillance approaches yielded the highest sensitivity (99%; 95% CI 93%, 100%) but lowest specificity (76%; 95% CI 71%, 80%) (Table 2.3).

Chapter 2: SCD Sensitivity to Changes in Serum Creatinine Concentration Criteria

Using creatinine normal values from the Harriet Lane Handbook (49) instead of Meites (48) changed the classification of three cases, all ≤2 years-old and diagnosed with HUS during

their hospitalization. Two 2-year-old SCD-defined HUS cases with serum creatinine concentrations above 0.60 (the Meites cutoff) but ≤0.70 (the Harriet Lane Handbook cutoff) became non-cases. One 1-year-old who failed to meet the Meites cutoff of 0.60 for the SCD became a case using the Harriet Lane Handbook cutoff of 0.40. Using the alternative criteria, HUS incidence remained similar, and all dialysis cases were identified as HUS (Appendix, Table S3).

Chapters 3 & 4: Assigning Phylogenetic Lineage to Non-SNP-typed Isolates

In previous studies analyzing patterns associated with STEC O157:H7 phylogenetic classification, it has been common to use a single representative isolate from each PFGE subtype (20, 21, 69). This practice masks the variability among isolates with the same PFGE fingerprint (e.g. variability in demographics, location, etc.). Further, estimation of effects at the population level is compromised, because the isolates being analyzed are not reflective of the STEC O157:H7 case population distribution. To accurately make inference at the population level, we sought to include all reported cases during the study period. Because we did not have sufficient resources to SNP-type all isolates, we leveraged the assumption inherent in the single-representative-isolate approach, although not generally made explicit: isolates with the same PFGE fingerprint belong to the same phylogenetic grouping.

Our sample contained 1160 isolates reflecting 355 unique PFGE patterns. We SNP-typed 793 of these isolates, covering 319 PFGE subtypes. The 36 PFGE subtypes not SNP-typed were either biochemically atypical or they were not present in the isolate bank. Atypical isolates were exclusively from 2013 and 2014, the last two years of sampling. Missing isolates were predominantly (82%) from 2005 and 2006, the first two years of sampling. Of the 793 SNP-type isolates, 570 belonged to a PFGE subtype with multiple SNP-typed isolates. Among these 570,

we examined which phylogenetic lineages the isolates had been assigned via SNP-typing. All but one PFGE subtype were assigned a consistent lineage. The one variable PFGE subtype was EXHX01.0047. It encompassed 82 isolates: 21 were not typed, 59 were typed to lineage IIa, and 2 were typed to lineage Ib. In other words, only 2 of 570 isolates, or 0.4%, showed aberrant lineage assignment. With this, we felt that the assumption that isolates of the same PFGE subtype would be in the same lineage held adequately well to use the SNP-typing results to assign lineage to non-SNP-typed isolates. We were able to assign lineage to 328 additional isolates using this approach.

Chapter 3: Causal Mediation Analysis

The parametric estimation algorithm for causal mediation analysis (85) is simulation-based, which provides greater flexibility in model choice than analytic regression-based mediation analysis (117). Separate models are fit for the mediator and outcome variables. Through a given number of simulations, model parameters are drawn from their sampling distributions. The causal mediation effects are calculated for each simulation and then combined into summary statistics (85). The causal mediation effect is given by (80):

$$\delta_i(t) \equiv Y_i\big(t, M_i(1)\big) - Y_i(t, M_i(0))$$

where t is a given level of the exposure, and $M_i(t)$ is the potential value of the mediator at the given exposure level. The causal mediation effect $\delta_i(t)$ is also referred to as the natural indirect effect. The average causal mediation effect (ACME) is the expected value of $\delta_i(t)$ across all subjects.

The direct effect is given as (80):

$$\zeta_i(t) \equiv Y_i\big(1, M_i(t)\big) - Y_i\big(0, M_i(t)\big)$$

Averaged across all subjects, the average direct effect (ADE) is the comparison of potential outcomes for different exposure levels (e.g. lineage IIb vs. lineage Ib) when the mediator takes

on its potential value at *t*. This is distinct from what are called controlled direct effects (82, 118), in which the mediator takes on specific observed values for the entire population, thus blocking all effects of the mediator on the outcome. In the ADE, only the effect of the exposure on the mediator is blocked, providing a better estimate of the direct effect observed in nature.

Chapter 3: Modification of Causal Mediation Effects

Because of the potential for age to modify the effect of lineage on HUS, we were also interested in testing whether age modified the role of *stx* genes as mediators. We separately estimated the ACME and ADE for 2-year-olds and 40-year-olds and tested whether the values were different across ages. Ages 2 and 40 are the approximate midpoints of the two largest age groups. The *mediation* package (86) was used to estimate the age-specific ACME and ADE and generate summary statistics for the difference and 95% CI based on 10,000 simulations.

Chapter 3: Case-Cohort Sample and SBI Typing

The case-cohort sampling approach entailed sampling a random subcohort from the full STEC O157:H7 cohort and supplementing the sample with all remaining cases of HUS. Of the 76 HUS cases, 16 could not be SBI-typed, preventing their inclusion in the case group of the case-cohort sample. Five were not present in the isolate bank and four were biochemically atypical STEC O157:H7 isolates and were excluded from typing. SBI typing was done in concert with SNP typing, so four isolates that had been SNP-typed for a previous project were not SNP-typed again and thus not SBI-typed either. SBI typing was conducted prior to completion of the HUS validation, and the final three non-SBI-typed HUS cases were not noted as HUS cases on the CRF.

Chapter 3: Selection Bias Estimation

To assess the effect on the observed OR if the restriction to reported cases produced selection bias, we used the *simcausal* package (119) in R. We simulated a case population infected with STEC O157:H7 similar in lineage composition to our population: 55% Ib, 20% IIa, 15% IIb, and 10% rare lineages. Severity, reporting frequency, and HUS frequency were modeled to obtain expected characteristics from our population. Based on Scallan et al. (1), we expected 11.7% of cases to be reported. We also assumed all HUS cases would be reported. Among those reported, we expected 7.2% total frequency of HUS, with proportions by lineage of 5%, 10%, 10%, and 1% for lineage Ib, IIa, IIb, and the rare lineages, respectively. Under these conditions, in the case of a true OR of 2.01, selection bias would produce an OR of 1.86, showing a slight attenuation of the true effect.

Chapter 4: Spatial Segregation by Diggle's Kernel Estimation Method

Diggle's kernel estimation provides smoothed estimates of spatial segregation that take into account multiple neighbors of each case. Diggle's method assumes an underlying Poisson point process for each phylogenetic lineage. The degree of smoothing is dependent on the choice of a bandwidth. Using the *spatialkernel* R package (107), a cross-validated log-likelihood function can be employed to calculate the bandwidth (106). This bandwidth is used for all lineages within a given analysis but is recalculated for different subsets of the data (e.g. when restricting the analysis to particular years). To identify the sensitivity of the kernel estimation results to the bandwidth of 0.6472 selected, alternate bandwidths were tested: 0.02, 0.1, 0.2, 0.4, and 0.9. All yielded $p = 0.001$ for the overall test for spatial segregation. The segregation maps for individual lineages grew predictably smoother as the bandwidth was increased and identified statistically significant areas of segregation consistent with the primary result from a bandwidth of 0.6472.

Temporal variation in segregation was tested across three intervals: 2005-2007, 2008-2010, and 2011-2014. The slightly longer last interval is not expected to affect the validity of the

results. However, because of greater number of cases in this interval, greater precision was expected. For the overall test of variation of spatial segregation across time intervals using the kernel regression method, a bandwidth of 0.8236 was chosen using the cross-validated log-likelihood function. The bandwidths chosen using this method for each of the individual intervals were 1.0000 for 2005-2007, 0.7256 for 2008-2010, and 0.9314 for 2011-2014. Not unexpectedly given the high degree of smoothing in the first and last period, only the middle period had detectable overall spatial segregation ($p = 0.001$). However, all periods displayed some statistically significant spatial segregation for individual lineages (Appendix Figures S2-S4). A bandwidth of 0.4 was also tested for each of the intervals, resulting in statistically significant tests for overall spatial segregation in each interval (2005-2007 $p = 0.037$, 2008-2010 $p = 0.001$, 2011-2014 $p = 0.014$).

Chapter 4: Multinomial Generalized Additive Model

The multinomial GAM provides a smoothed risk surface relative to Ib, the most common lineage. Unlike the direct measures of spatial segregation, the GAM captures spatial trends without selecting a specific distance or number of neighbors across which to smooth. It does this through a flexible spline function. The GAM also supports adjustment for covariates, providing some assurance that the associations observed are not due to factors such as the distribution of cases by age. Results of the GAM multinomial models must be interpreted conditional on having a reported STEC O157:H7 illness. As such, odds ratios presented estimate risk proportional to that in the most common lineage, Ib.

We tested multiple aspects of the GAM specification. Latitude and longitude were specified individually and jointly to allow interaction. The basis dimension of the penalized regression smoother was altered to improve the effective degrees of freedom. Age and sex covariates were removed, and the form of the spline smoother was altered. Lineage IIa was used as the comparison lineage. These sensitivity analyses are summarized in Appendix Table S6.

None of the model perturbations meaningfully changed the primary model results. In the set of GAMs incorporating year, a trivariate smooth of latitude, longitude, and year was also tested and found to be statistically significant for lineages IIa and IIb (Appendix Table S6).

Chapter 4: Spatial Segregation by Dixon's Nearest-neighbor Method

Another measure of spatial segregation, Dixon's nearest-neighbor method considers only the closest neighbor of each case. It conducts no smoothing and can be expected to be sensitive to clustered outbreaks. This method does not indicate areas in which spatial segregation exists but does provide an overall test of spatial segregation, as well as for segregation of individual lineages and pairwise segregation tests. We created a 4x4 contingency table of nearest-neighbor counts for each lineage group. A χ-square with 12 degrees of freedom was used to test overall spatial segregation, and segregation was tested for each individual lineage group (Appendix Table S7). We calculated Dixon's segregation index for each nearest-neighbor combination (e.g., from Ib to IIa; Appendix Table S8). Dixon's pairwise segregation index is defined according to (105) as:

$$S_{ij} = \log \frac{N_{ij}/(N_i - N_{ij})}{\mathrm{E}N_{ij}/(N_i - \mathrm{E}N_{ij})} = \log \frac{N_{ij}/(N_i - N_{ij})}{N_i/(N - N_j - 1)}$$

where *i* and *j* in this analysis are phylogenetic lineages. A positive value of *S* indicates association, and a negative value indicates segregation. Z-scores for each combination were calculated by comparing the observed nearest-neighbor count in each cell to the expected count. A p-value based on the Z-scores was calculated assuming an asymptotic normal distribution. The *dixon* R package was used for this analysis (120).

We used Dixon χ-square tests for segregation to indicate statistically significant segregation overall ($p < 0.001$) and for lineages Ib, IIa, and IIb ($p = 0.046$, $p = 0.002$, and $p < 0.001$, respectively), but not for the group of rare lineages (Appendix Table S7). This is consistent with the findings of the kernel estimation method, which found statistically

significant overall spatial segregation and identified areas of segregation for lineages Ib, IIa, and IIb. Dixon's method also tests associations between individual lineages. Pairwise nearest-neighbor comparisons showed statistically significant positive association from each of lineages Ib, IIa, and IIb to itself. Segregation was observed from Ib to IIa, IIa to the rare lineages, IIb to all other lineages, and the rare lineages to Ib (Appendix Table S8).

Spatial segregation was examined using Dixon's method for the three intervals analyzed with the kernel estimation method. Spatial segregation was found to be statistically significant with $p < 0.001$ during all three periods, contrasting with Diggle's method, which only identified statistically significant overall segregation during the 2008-2010 period. However, the two spatial segregation tests were consistent in identifying spatial segregation of lineage IIb during all intervals ($p < 0.001$ for Dixon's method during all intervals). Additionally, Dixon's method identified segregation of lineage IIa during the 2005-2007 period ($p < 0.001$) and segregation of lineage Ib during the 2008-2010 ($p < 0.001$) and 2011-2014 ($p = 0.005$) periods.

Chapter 4: Multinomial Spatial Scan Statistics

We used multinomial spatial scan statistics (110) in SaTScan (121) to identify clusters within which the distribution of lineages differed significantly from the distribution of lineages outside the cluster. The spatial scan statistics are designed to identify clusters of disease. In the multinomial framework used here, the clusters reflect areas within which the distribution of cases by lineage is skewed as compared to the area outside the cluster. These are similar to the areas of segregation identified by the kernel regression method. However, the scan statistics look at the distribution of all four lineages simultaneously and not individually, thus allowing detection of clusters in which multiple lineages may be out of proportion. Like the multinomial GAM models, the multinomial spatial scan statistics must be interpreted conditionally on having a reported STEC O157:H7 illness.

For the primary spatial scan statistic model we used a maximum cluster size of 20% of cases. Statistical significance of the clusters was determined based on Monte Carlo replications under the null. Relative risks presented estimate risk of one's infection being from the given lineage inside the cluster compared to the risk outside that cluster.

We identified three statistically significant clusters in which the distribution of cases by phylogenetic lineage varied from the distribution in the rest of the State (Appendix Figure S5). The first cluster ($p = 0.001$) contained 203 cases, was centered in the southwest region of the State, and was characterized by a higher proportion of lineage IIb cases than observed elsewhere in the State [relative risk (RR) 2.59). The second cluster ($p = 0.001$), encompassing the sparsely-populated northern reaches of the State, contained 185 cases and had somewhat more Ib (RR 1.37) and rare lineage (RR 1.88) cases and fewer IIb cases (RR 0.29). The final significant cluster ($p = 0.006$) contained 79 cases in the south-central region of the State: lineage IIa was more common than elsewhere in the State (RR 1.70), IIb was uncommon (RR 0.13), and cases due to rare lineages were nearly absent (RR 0). The first cluster, dominated by IIb, and third cluster, dominated by IIa, recapitulate the results of the kernel estimation maps and, for IIb, the GAM-generated risk surface. The second cluster, dominated by lineage Ib, is larger and centered somewhat further east than the area of segregation identified for Ib by the kernel estimation method, though still similar.

Altering the parameters of the analysis to allow fewer or greater percentages of the cases to be included in clusters did not meaningfully affect the position of the clusters identified. We tested allowing clusters up to 50% of cases and 10% of cases. From the former, the main IIb-dominant and Ib/rare-dominant clusters were identified, but the IIa-dominated cluster was not. Limiting clusters to 10% of cases, all three clusters identified in the primary analysis were identified but with smaller numbers of included cases.

We detected variant clusters using multinomial spatio-temporal scan statistics, using year as the time scale and allowing up to 50% of the study period in a cluster, as well as purely

spatial clusters. We identified three statistically significant clusters (Appendix Figure S6). The first ($p = 0.001$) contained 76 cases reported 2009 to 2012 in the southwest region of the State and had an elevated risk of lineage IIb (RR 4.45). The second cluster ($p = 0.001$) included 107 cases across the northeast region during the years 2005 to 2009. The Ib (RR 1.61) and rare (RR 1.88) lineages were elevated. The third cluster ($p = 0.002$) included only 46 cases reported during 2009 and 2010, with a predominance of lineage IIb (RR 3.63) and near-absence of IIa (RR 0.09). This cluster included part of Seattle, Washington's largest urban area, and areas immediately south and east.

Chapter 4: Secondary Cases

To separate the effect of person-to-person transmission from other potential environmental factors that may result in segregation, sensitivity analyses were conducted after excluding known secondary cases. To be excluded, the most likely source of the infection had to have been identified during the public health investigation as person-to-person, or the notes had to indicate that another individual in the household or childcare situation had previously been diagnosed. Based on these criteria, 82 secondary cases were excluded. No meaningful changes in the results were observed. The overall test of spatial segregation was statistically significant using the kernel estimation method ($p = 0.002$) and the nearest-neighbor method ($p < 0.001$). The latitude/longitude smooth of lineage IIb from the multinomial GAM is statistically significantly different than that of lineage Ib ($p < 0.001$). However, the cluster identified in the southwest region of the State, dominated by lineage IIb, through multinomial spatial scan statistics moved somewhat northward and decreased in size without the secondary cases.

Chapter 4: Reporting Bias

We assessed potential reporting bias by county. Reporting of cases who have tested positive is considered near 100% (1), but testing intensity may vary by provider. STEC O157 is

most often detected by stool culture, a test that also detects *Campylobacter*, *Salmonella*, and *Shigella*. If providers in an area have heightened awareness of STEC O157 and are more likely to test for it than in other areas, we would expect that detections of these other pathogens would also be higher. There is overlap in the epidemiology of STEC O157, *Campylobacter*, and *Salmonella*, so some correlation is expected. However, risk factors for *Shigella* are generally different (93). If there were reporting bias, we would expect this to most greatly impact the observed incidence of milder STEC O157 strains.

Case counts by county for 2005-2014 for Campylobacteriosis, Salmonellosis, and Shigellosis were obtained from the Washington State Communicable Disease Reports for 2009 and 2014 (each contained five years of data) (122, 123). Incidence rates were calculated using county populations as reported in 2010 U.S. Census TIGER/Line Shapefiles (50). Using the *GISTools* (124) package in R, we mapped the incidence quintile of each of the four pathogens at the county level for the study period to assess the potential for reporting bias (Appendix Figure S7). Two counties, Yakima and Grant, appear in the uppermost quintile of incidence for each of the four diseases. However, incidence of rare lineage STEC O157:H7 in this region is remarkably low (Figure 4.1, Appendix Figure S5). Infections caused by these bacteria are generally milder (Table 4.1) and would be the type whose numbers would be exaggerated in the presence of heightened testing. Thus, it is unlikely that reporting bias is responsible for the observed results.

Supplemental Figures & Tables

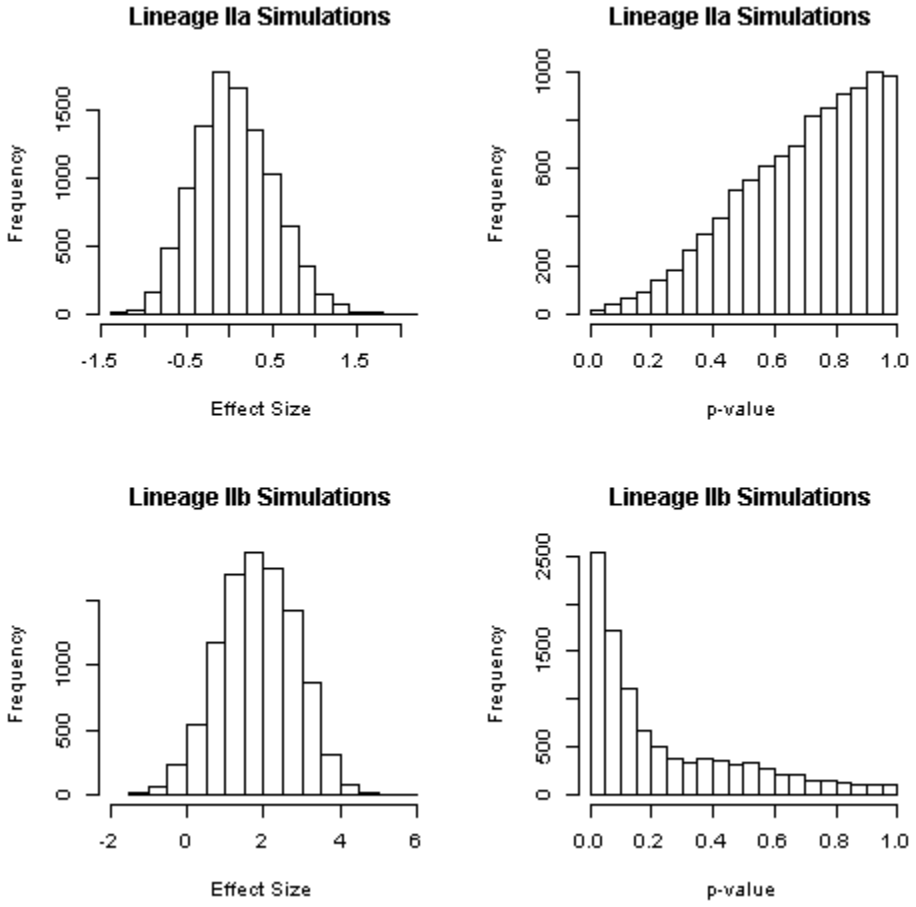

Figure S1. Histograms of results from 10,000 simulations of randomly selecting one isolate per PFGE-defined strain. Effect size and p-value were obtained from GEE logistic regression of HUS on lineage, adjusted for age (continuous) and sex. Effect sizes can be exponentiated to obtain ORs. The lineage IIa OR exceeded 1 (effect size 0) in 53% of simulations, and $p<0.05$ in 0.11%. The lineage IIb OR exceeded 1 in 97% of simulations, and $p<0.05$ in 25%.

Abbreviations: GEE, generalized estimating equations; HUS, hemolytic uremic syndrome; OR, odds ratio; PFGE, pulsed field gel electrophoresis

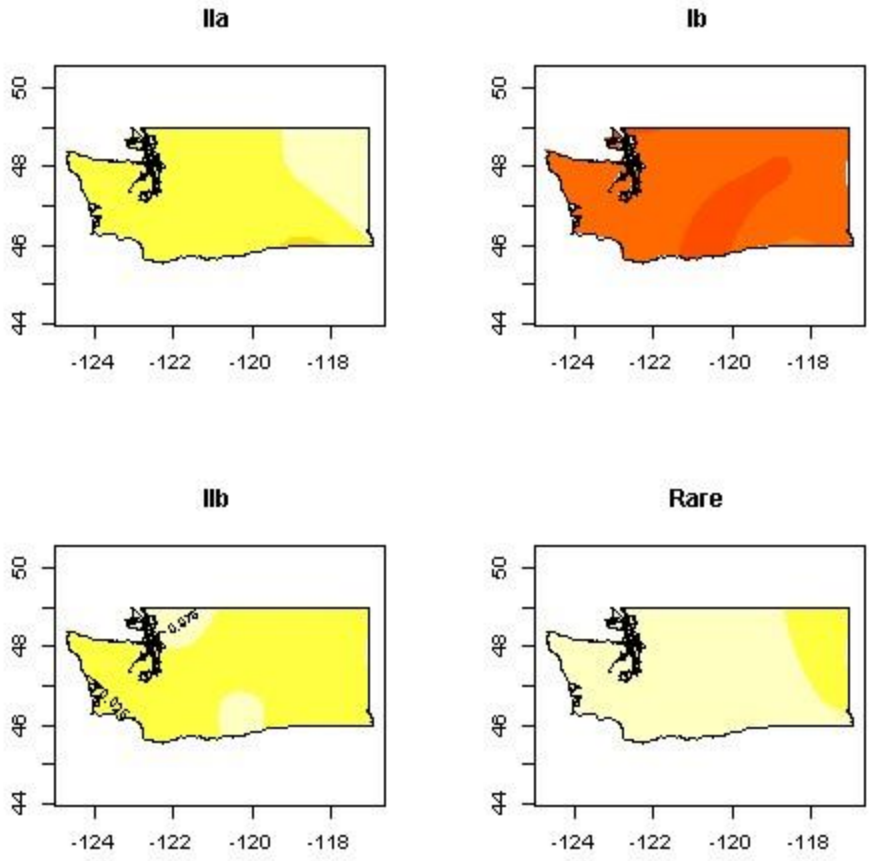

Figure S2. Kernel-based estimation of spatial segregation by lineage, 2005-2007, n = 305, bandwidth = 1.0000. Overall spatial segregation p = 0.769. Darker hues indicate greater segregation. Contour lines marked 0.025 define areas in which the given lineage is statistically significantly segregated. Contour lines marked 0.975 define areas in which the given lineage is statistically significantly less likely to be found in proximity to itself than to other lineages.

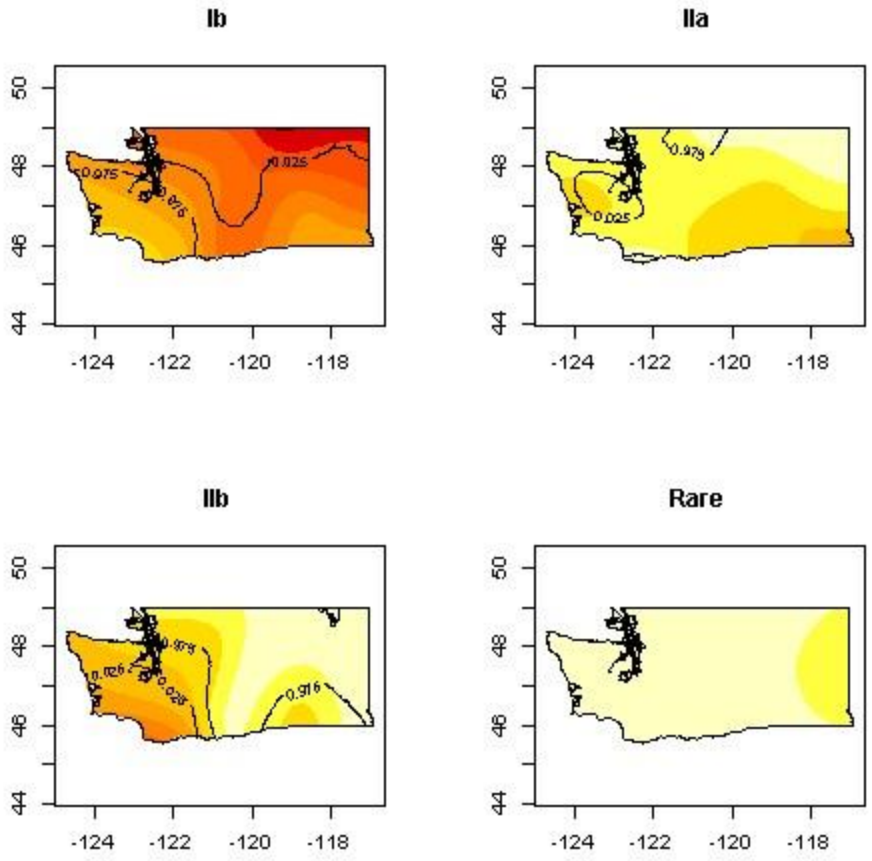

Figure S3. Kernel-based estimation of spatial segregation by lineage, 2008-2010, n = 367, bandwidth = 0.7256. Overall spatial segregation p = 0.001. Darker hues indicate greater segregation. Contour lines marked 0.025 define areas in which the given lineage is statistically significantly segregated. Contour lines marked 0.975 define areas in which the given lineage is statistically significantly less likely to be found in proximity to itself than to other lineages.

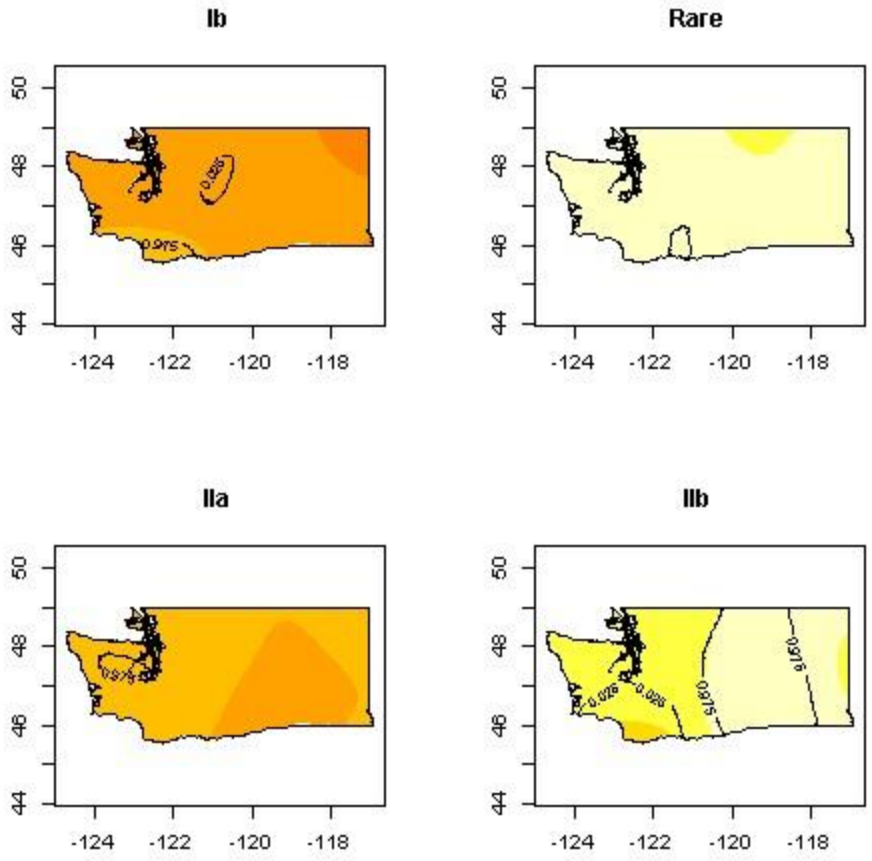

Figure S4. Kernel-based estimation of spatial segregation by lineage, 2011-2014, n = 439, bandwidth = 0.9314. Overall spatial segregation p = 0.138. Darker hues indicate greater segregation. Contour lines marked 0.025 define areas in which the given lineage is statistically significantly segregated. Contour lines marked 0.975 define areas in which the given lineage is statistically significantly less likely to be found in proximity to itself than to other lineages.

Figure S5. Statistically significant clusters of variant phylogenetic lineage.

Multinomial spatial scan statistics were used to identify clusters in which the distribution of lineages varied from that of the rest of the state. Clusters were restricted to a maximum of 20% of cases. Cluster 1: 203 cases; Ib relative risk (RR) = 0.66, IIa RR = 0.94, IIb RR = 2.59, Rare RR = 0.80; p = 0.001. Cluster 2: 185 cases; Ib RR = 1.37, IIa RR = 0.65, IIb RR = 0.29, Rare RR = 1.88; p = 0.001. Cluster 3: 79 cases; Ib RR = 1.14, IIa RR = 1.70, IIb RR = 0.13, Rare RR = 0; p = 0.006.

Figure S6. Statistically significant space-time clusters of variant phylogenetic lineage. Multinomial spatio-temporal scan statistics were used to identify clusters in which the distribution of lineages varied from that of the rest of the state during years outside the cluster. Clusters were restricted to a maximum of 20% of cases and 50% of the study window. Cluster 1: 2009-2012; 76 cases; Ib relative risk (RR) = 0.28, IIa RR = 0.49, IIb RR = 4.45, Rare RR = 1.36; p = 0.001. Cluster 2: 2005-2009; 107 cases; Ib RR = 1.61, IIa RR = 0.22, IIb RR = 0.19, Rare RR = 1.88; p = 0.001. Cluster 3: 2009-2010; 46 cases; Ib RR = 0.65, IIa RR = 0.09, IIb RR = 3.63, Rare RR = 0.72; p = 0.002.

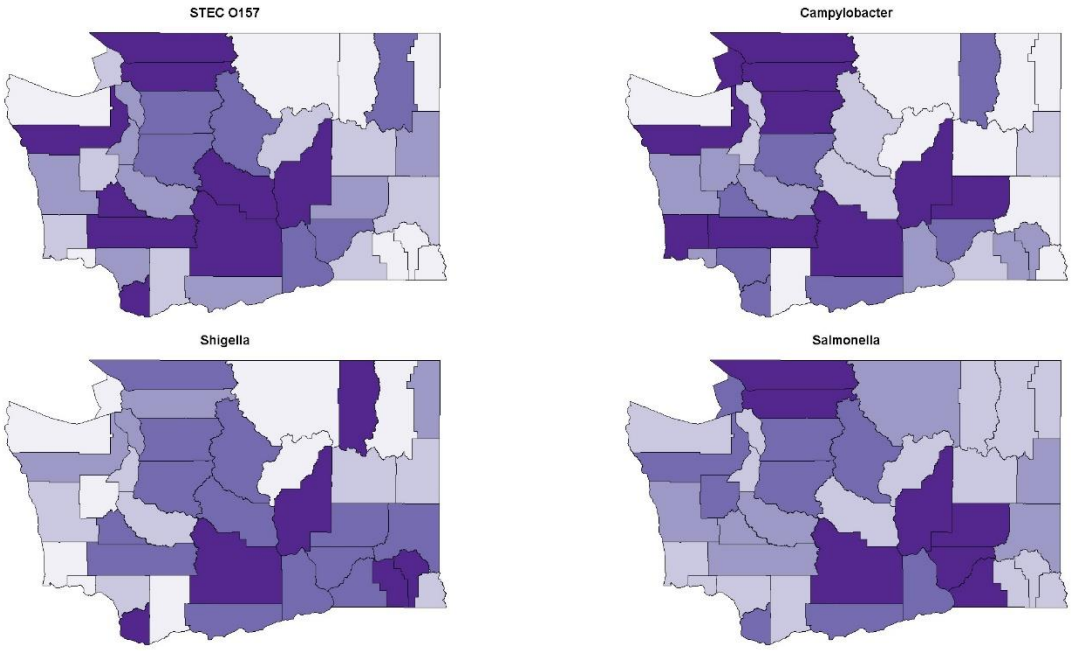

Figure S7. Incidence rate quintiles by county of reported STEC O157, Campylobacter, Shigella, and Salmonella, 2005-2014. These four pathogens are routinely tested for simultaneously, and uniformly high rates may suggest higher testing intensity in a county.

Table S1. Fatal Hospitalized STEC O157:H7 Cases, Washington, 2005-2014

Case	Age (years)	Platelets cells/mm³ (day)	Hematocrit % (day)	Smear	Creatinine mg/dL (day)	Hematuria or Proteinuria	Dialysis	Meets Stringent Clinical Definition	Meets CSTE Definition	Meets Alternative Literature Definition	Case Report Form HUS	Hospital HUS Diagnosis	Combined Definition
1	≥60	173 (28)	29.2 (28)	normal	0.90 (28) 1.50 (34)	Yes	No	No	No	No	No	No	No
2	≥60	142 (8)	28.5 (8)	ND	2.70 (8)	ND	No	Yes	Yes	No	No	No	Yes
3	<5	12 (12)	25.4 (12)	4+ schistocytes	1.40 (12)	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
4	<5	58 (8)	23.0 (8)	1+ schistocytes	1.96 (8)	ND	Yes	Yes	Yes	Yes	Yes	Yes	Yes
5	≥60	96 (7)	37.8 (7)	normal	0.80 (7)	Yes	No	No	Yes	No	No	No	Yes

Day refers to the day of illness after diarrhea onset. Values for multiple days are shown when relevant.

Abbreviation: CSTE, Council for State and Territorial Epidemiologists; HUS, hemolytic uremic syndrome; ND, not documented

Table S2. Clinical Characteristics of Discrepant Hospitalized STEC O157:H7 Cases, Washington, 2005-2014

Case	Age (years)	Platelets cells/mm^3 (day)	Hematocrit % (day)	Smear	Creatinine mg/dL (day)	Hematuria or Proteinuria	Dialysis	Meets Stringent Clinical Definition	Meets CSTE Definition	Meets Hematology-focused Confirmed Definition	Case Report Form HUS	Hospital HUS Diagnosis	Combined Definition
1	77	90 (24)	29.2 (24)	ND	4.10 (24)	ND	No	Yes	No	No	No	No	No
2	2	32 (10)	22.1 (10)	ND	0.80 (10)	ND	No	Yes	No	No	No	Yes	Yes
3	2	47 (11)	21.8 (11)	1+ schistocytes	0.64 (11)	ND	No	Yes	No	No	Yes	Yes	Yes
4	≥90	142 (8)	28.5 (8)	ND	2.70 (8)	ND	No	Yes	Yes	No	No	No	Yes
5	6	37 (6)	29.2 (6)	ND	2.00 (6)	ND	Yes	Yes	Yes	No	No	Yes	Yes
6	3	19 (7)	24.5 (7)	3+ schistocytes	0.80 (7)	Yes	No	Yes	Yes	No	No	Yes	Yes
7	85	109 (11)	29.0 (11)	1+ schistocytes	1.42 (11)	Yes	No	Yes	Yes	No	No	Yes	Yes
8	6	22 (9)	24.0 (9)	ND	1.30 (9)	ND	No	Yes	Yes	No	--	Yes	Yes
9	7	43 (9)	21.9 (9)	ND	1.80 (9)	ND	No	Yes	Yes	No	--	Yes	Yes
10	3	49 (5)	26.9 (5)	ND	3.70 (5)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
11	10	74 (6)	29.9 (6)	ND	2.40 (6)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
12	12	35 (10)	29.1 (10)	ND	1.00 (10)	Yes	No	Yes	Yes	No	Yes	Yes	Yes
13	2	14 (10)	24.8 (10)	ND	3.70 (10)	ND	Yes	Yes	Yes	No	Yes	Yes	Yes

14	1	112 (8)	29.2 (8)	normal	1.70 (8)	Yes	No	Yes	Yes	No	Yes	Yes	Yes
15	8	37 (14)	28.9 (14)	normal	1.70 (14)	No	Yes	Yes	Yes	No	Yes	Yes	Yes
16	6	35 (7)	20.9 (7)	4+ schistocytes	0.80 (7)	Yes	No	Yes	Yes	No	Yes	Yes	Yes
17	2	39 (8)	24.1 (8)	3+ schistocytes	0.80 (8)	Yes	No	Yes	Yes	No	Yes	Yes	Yes
18	1	140 (4)	23.6 (4)	3+ schistocytes	0.90 (4)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
19	1	41 (5)	26.1 (5)	3+ schistocytes	0.71 (5)	Yes	No	Yes	Yes	No	Yes	Yes	Yes
20	2	65 (7)	26.6 (7)	1+ schistocytes	0.70 (7)	Yes	No	Yes	Yes	No	Yes	Yes	Yes
21	2	46 (12)	19.1 (12)	RBC fragments	1.40 (12)	Yes	No	Yes	Yes	Yes	No	Yes	Yes
22	9	23 (5)	26.9 (5)	1+ schistocytes	3.80 (5)	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
23	8	41 (9)	29.8 (9)	2+ schistocytes	1.20 (9)	Yes	No	Yes	Yes	Yes	No	Yes	Yes
24	28	69 (10)	28.4 (10)	1+ schistocytes	1.40 (10) 1.60 (11)	Yes	No	Yes	Yes	Yes	No	Yes	Yes
25	12	37 (10)	26.2 (10)	few schistocytes	1.00 (10)	Yes	No	No	Yes	Yes	Yes	Yes	Yes
26	6	11 (6) 39 (11)	19.4 (11) 23.7 (14)	2+ schistocytes	1.10 (6) 0.63 (11)	ND	No	No	Yes	Yes	Yes	Yes	Yes
27	12	84 (11)	21.8 (11)	ND	0.86 (11)	Yes	No	No	Yes	No	Yes	Yes	Yes
28	4	26 (8)	20.2 (8)	3+ schistocytes	0.50 (8)	Yes	No	No	Yes	No	Yes	Yes	Yes
29	2	83 (5)	30.0 (5)	RBC fragments	0.60 (5)	Yes	No	No	Yes	No	Yes	Yes	Yes

30	2	33 (9)	19.9 (9)	1+ schistocytes	0.40 (9)	Yes	No	No	Yes	No	Yes	Yes	Yes
31	9	29 (7)	28.2 (7)	RBC fragments	0.50 (7)	Yes	No	No	Yes	No	Yes	Yes	Yes
32	6	54 (10)	26.3 (10)	RBC fragments	0.60 (10)	Yes	No	No	Yes	No	Yes	Yes	Yes
33	1	109 (12)	25.5 (12)	1+ schistocytes	0.34 (12)	Yes	No	No	Yes	No	Yes	Yes	Yes
34	4	78 (9)	24.4 (9)	1+ schistocytes	0.52 (9)	Yes	No	No	Yes	No	Yes	Yes	Yes
35	57	43 (9)	26.6 (9)	1+ schistocytes	1.10 (8)	Yes	No	No	Yes	No	Yes	Yes	Yes
36	15	73 (12)	29.5 (12)	1+ schistocytes	1.10 (11)	Yes	No	No	Yes	No	No	Yes	Yes
37	22	51 (11)	38.0 (11)	RBC fragments	1.30 (11)	Yes	No	No	Yes	No	No	Yes	Yes
38	2	66 (9)	22.4 (11)	RBC fragments	0.40 (8)	Yes	No	No	Yes	No	No	No	Yes
39	5	41 (8)	25.8 (8)	1+ schistocytes	0.50 (8)	Yes	No	No	Yes	No	No	No	Yes
40	7	192 (10)	27.0 (10)	normal	0.40 (10)	Yes	No	No	Yes	No	No	No	Yes
41	29	233 (2)	37.4 (2)	normal	0.70 (2)	Yes	No	No	Yes	No	No	No	Yes
42	6	352 (5)	33.8 (5)	normal	0.40 (5)	Yes	No	No	Yes	No	No	No	Yes
43	8	137 (8)	32.4 (8)	normal	0.60 (8)	Yes	No	No	Yes	No	No	No	Yes
44	18	208 (3) 133 (9)	34.6 (3)	normal	0.80 (3)	Yes	No	No	Yes	No	No	No	Yes
45	11	219 (4)	34.5 (4)	normal	0.40 (4)	Yes	No	No	Yes	No	No	No	Yes
46	6	111 (8)	30.4 (8)	normal	0.60 (8)	Yes	No	No	Yes	No	No	No	Yes

47	10	253 (6)	32.1 (6)	ND	0.40 (6)	Yes	No	No	Yes	No	No	No	Yes
48	8	237 (5)	34.2 (5)	ND	0.50 (5)	Yes	No	No	Yes	No	No	No	Yes
49	69	164 (3)	30.4 (3)	ND	0.80 (3)	Yes	No	No	Yes	No	No	No	Yes
50	38	208 (6)	32.4 (6)	normal	0.60 (6)	Yes	No	No	Yes	No	No	No	Yes
51	59	221 (7)	31.2 (7)	ND	0.77 (7)	Yes	No	No	Yes	No	No	No	Yes
52	19	217 (2)	35.7 (2)	ND	0.50 (2)	Yes	No	No	Yes	No	No	No	Yes
53	69	176 (5)	33.5 (5)	ND	0.90 (5)	Yes	No	No	Yes	No	No	No	Yes
54	36	192 (5)	30.7 (5)	normal	5.50 (5)	No	No	No	Yes	No	No	No	Yes
55	54	259 (8)	32.0 (8)	rare schistocytes	0.90 (8)	Yes	No	No	Yes	No	No	No	Yes
56	73	216 (6)	31.9 (6)	ND	0.60 (6)	Yes	No	No	Yes	No	No	No	Yes
57	8	240 (7)	28.8 (7)	normal	0.40 (7)	Yes	No	No	Yes	No	No	No	Yes
58	70	205 (4)	32.8 (4)	ND	1.60 (4)	No	No	No	Yes	No	No	No	Yes
59	69	103 (9)	29.6 (9)	normal	0.50 (9)	Yes	No	No	Yes	No	No	No	Yes
60	9	230 (5)	34.7 (5)	normal	0.40 (5)	Yes	No	No	Yes	No	No	No	Yes
61	73	187 (7)	29.0 (7)	normal	0.40 (7)	Yes	No	No	Yes	No	No	No	Yes
62	69	162 (5)	33.9 (5)	ND	1.00 (5)	Yes	No	No	Yes	No	No	No	Yes
63	34	191 (8)	29.7 (8)	normal	0.60 (8)	Yes	No	No	Yes	No	No	No	Yes
64	16	210 (3)	33.3 (3)	normal	0.70 (3)	Yes	No	No	Yes	No	No	No	Yes
65	20	108 (8)	33.5 (8)	normal	0.90 (8)	Yes	No	No	Yes	No	No	No	Yes
66	62	96 (7)	37.8 (7)	normal	0.80 (7)	Yes	No	No	Yes	No	No	No	Yes
67	18	226 (5)	40.4 (5)	normal	0.60 (5)	Yes	No	No	Yes	No	No	No	Yes
68	49	179 (5)	30.7 (5)	ND	0.90 (5)	Yes	No	No	Yes	No	No	No	Yes
69	8	248 (6)	33.2 (6)	ND	0.50 (6)	Yes	No	No	Yes	No	No	No	Yes
70	49	242 (4)	28.9 (4)	ND	0.70 (4)	Yes	No	No	Yes	No	No	No	Yes

71	18	279 (5)	35.0 (7)	ND	0.90 (7)	Yes	-	No	Yes	No	No	No	Yes
72	66	162 (6)	37.1 (6)	normal	1.60 (4)	Yes	No	No	Yes	No	No	No	Yes
73	30	185 (7)	28.2 (7)	ND	0.60 (7)	Yes	-	No	Yes	No	No	No	Yes
74	21	139 (6)	35.5 (6)	normal	0.70 (6)	Yes	-	No	Yes	No	No	No	Yes
75	47	216 (6)	33.4 (8)	normal	0.90 (5)	Yes	-	No	Yes	No	No	No	Yes
76	≥90	201 (5)	33.1 (5)	normal	1.60 (3)	Yes	-	No	Yes	No	No	No	Yes
77	15	242 (6)	35.3 (5)	ND	0.52 (4)	Yes	-	No	Yes	No	No	No	Yes
78	63	180 (9)	32.0 (9)	normal	3.60 (9)	Yes	No	No	Yes	No	No	No	Yes
79	60	182 (4)	32.6 (4)	normal	0.70 (4)	Yes	No	No	Yes	No	No	No	Yes
80	84	201 (7)	32.0 (7)	ND	1.30 (7)	Yes	No	No	Yes	No	No	No	Yes
81	48	212 (7)	33.5 (7)	ND	0.60 (6)	Yes	No	No	Yes	No	No	No	Yes
82	75	356 (6)	30.6 (6)	ND	0.60 (4)	Yes	No	No	Yes	No	No	No	Yes
83	81	184 (5)	31.3 (5)	ND	0.80 (5)	Yes	No	No	Yes	No	No	No	Yes
84	18	195 (7)	34.6 (7)	ND	0.70 (7)	Yes	No	No	Yes	No	No	No	Yes
85	2	184 (5)	31.2 (5)	ND	0.20 (5)	Yes	No	No	Yes	No	No	No	Yes
86	20	147 (4)	32.1 (4)	ND	0.50 (4)	Yes	No	No	Yes	No	No	No	Yes
87	65	216 (3)	31.5 (3)	ND	0.70 (3)	Yes	No	No	Yes	No	No	No	Yes
88	79	179 (6)	39.7 (6)	normal	2.10 (6)	ND	No	No	Yes	No	No	No	Yes
89	59	354 (8)	29.5 (8)	ND	0.80 (8)	Yes	No	No	Yes	No	No	No	Yes
90	68	276 (17)	34.8 (17)	ND	2.10 (15)	Yes	No	No	Yes	No	No	No	Yes
91	3	320 (10)	33.4 (10)	ND	0.30 (10)	Yes	No	No	Yes	No	No	No	Yes
92	18	182 (5)	33.6 (5)	ND	0.72 (5)	Yes	No	No	Yes	No	No	No	Yes
93	41	142 (7)	31.0 (7)	ND	6.20 (5)	ND	Yes	No	Yes	No	No	No	Yes
94	58	148 (4)	35.0 (5)	ND	0.80 (5)	Yes	No	No	Yes	No	No	No	Yes

95	71	188 (8)	33.0 (8)	normal	1.50 (8)	Yes	No	No	Yes	No	No	No	Yes
96	64	246 (15)	30.0 (15)	ND	0.80 (15)	Yes	No	No	Yes	No	No	No	Yes
97	65	275 (6)	30.0 (6)	ND	0.60 (6)	Yes	No	No	Yes	No	No	No	Yes
98	66	221 (6)	35.2 (6)	ND	1.30 (4)	Yes	No	No	Yes	No	No	No	Yes
99	<1	52 (21)	17.2 (21)	1+ schistocytes	0.20 (21)	Yes	No	No	Yes	No	--	No	Yes
100	12	142 (10)	31.6 (10)	normal	0.60 (10)	Yes	No	No	Yes	No	-	No	Yes
101	71	265 (6)	31.6 (6)	normal	0.50 (6)	Yes	No	No	Yes	No	-	No	Yes
102	66	370 (10)	35.9 (10)	ND	10.30 (10)	ND	Yes	No	Yes	No	-	No	Yes
103	14	97 (10)	32.4 (10)	normal	0.90 (10)	Yes	No	No	Yes	No	Yes	Yes	Yes
104	63	76 (8)	31.9 (8)	normal	1.20 (7)	Yes	No	No	Yes	No	Yes	Yes	Yes
105	2	439 (7)	30.3 (7)	rare schistocytes	0.20 (7)	Yes	--	No	Yes	No	Yes	--	No
106	5	21 (8)	19.2 (9)	1+ schistocytes	0.70 (8)	ND	No	No	No	No	Yes	Yes	Yes
107	2	47 (8)	25.8 (8)	rare schistocytes	0.58 (8)	ND	No	No	No	No	Yes	Yes	Yes
108	17	175 (5)	35.4 (5)	normal	0.70 (5)	ND	No	No	No	No	Yes	No	No
109	36	187 (2)	33.5 (4)	ND	ND	ND	--	No	No	No	Yes	--	No
110	8	242 (unk)	40.9 (unk)	ND	ND	No	--	No	No	No	Yes	--	No
111	1	46 (6)	25.4 (6)	2+ schistocytes	0.60 (6) 0.80 (8)	ND	No	No	No	No	No	Yes	Yes

112	2	64 (10)	28.9 (10)	1+ schistocytes	0.60 (10)	No	No	No	No	No	No	Yes	Yes

Day refers to the day of illness after diarrhea onset. Dashes indicate that HUS or dialysis status was not reported on the case report form or in the hospital chart. Values for multiple days are shown when the source of discordance is criteria met on different days.

Abbreviation: CSTE, Council for State and Territorial Epidemiologists; HUS, hemolytic uremic syndrome; ND, not documented; RBC, red blood cells; unk, unknown

Table S3. Distribution of Clinical Variables by Modified HUS Definitions

Variable	Stringent Clinical Definition		CSTE Definition	Hematology-focused Confirmed Definition
Modified criteria	Age-specific serum creatinine from Harriet Lane Handbook (49)	Criteria do not need to be met on the same day	Platelet count >150,000 mm^{-3} required	Microangiopathic changes not required
Number of HUS cases	75	78	102	69
Incidence per 100,000				
<18 years-old	0.40	0.42	0.51	0.37
All ages	0.11	0.12	0.15	0.10
Bloody diarrhea (%)	73 (99%)	76 (99%)	100 (99%)	68 (100%)
Missing	1	1	1	1
Vomiting (%)	65 (90%)	67 (89%)	80 (82%)	59 (88%)
Missing	3	3	4	2
Days hospitalized, median (IQR)	13 (9.5, 19.5)	13 (10, 19)	11 (6, 16)	13 (9, 21)
Missing	0	0	1	0
Urine output				
Anuria (%)	29 (45%)	29 (43%)	30 (34%)	29 (47%)
<1.0 ml/kg/hr (%)	55 (85%)	55 (82%)	65 (75%)	52 (84%)
Missing	10	11	15	7
Received dialysis (%)	39 (53%)	39 (51%)	40 (40%)	39 (57%)
Missing	1	1	2	1

Each column describes data for only those patients considered to have HUS according to the stated definition.

Abbreviation: HUS, hemolytic uremic syndrome; SD, standard deviation

Table S4. Sensitivity and Specificity of Probable HUS Definitions, Using the Stringent Clinical Definition as Comparator

	Modified CSTE Definition	Modified Hematology-focused Confirmed Definition
All cases (n)	429	429
Sensitivity (95% CI)	96% (89%, 99%)	87% (77%, 94%)
Specificity (95% CI)	92% (88%, 94%)	99% (98%, 100%)

Abbreviation: CSTE, Council for State and Territorial Epidemiologists; HUS, hemolytic uremic syndrome

Table S5. Modification by Age of *stx2a* Genotype Mediation of Lineage-HUS Association

	Age 2 Estimate (95% CI)	Age 40 Estimate (95% CI)	Difference (95% CI)
Lineage IIa			
Total effect	0.14 (-0.16, 0.60)	-0.083 (-0.18, 0.13)	
ACME	0.18 (-0.076, 0.46)	0.093 (-0.042, 0.31)	0.13 (-0.13, 0.42)
ADE	-0.048 (-0.40, 0.49)	-0.18 (-0.42, 0.041)	0.085 (-0.17, 0.58)
Proportion mediated	0.53 (-14.4, 13.6)	-0.68 (-4.37, 2.31)	
Lineage IIb			
Total effect	0.10 (-0.036, 0.25)	-0.005 (-0.12, 0.14)	
ACME	0.18 (-0.076, 0.40)	0.13 (-0.10, 0.38)	0.093 (-0.17, 0.33)
ADE	-0.073 (-0.32, 0.23)	-0.14 (-0.41, 0.14)	0.017 (-0.27, 0.31)
Proportion mediated	1.48 (-11.8, 16.0)	-0.39 (-35.9, 32.4)	

Modification of mediation of the lineage-HUS association by the *stx2a* genotype was tested at age 2 and age 40. Each mediation analysis was conducted for lineages IIa and IIb separately vs. lineage Ib. Genotype is compared against all non-IIa- and -IIb-dominant genotypes: no *stx*, *stx1*, *stx1-stx2a*, *stx1-stx2c*, *stx1-stx2a-stx2c*, and *stx2c*. Under the assumption of no interaction, the ACME, ADE, and proportion mediated shown are the average of the estimates for Ib and the lineage being analyzed. Mediation analyses were conducted using Imai et al.'s parametric estimation algorithm (85) with 10,000 simulations and robust standard errors. Tests of differences were conducted using 10,000 simulations of mediation models at age 2 vs. 40 by the mediation package in R (86).

Abbreviations: ACME, average causal mediated effect; ADE, average direct effect; CI, confidence interval; HUS, hemolytic uremic syndrome; *stx*, Shiga toxin gene

Table S6. Multinomial Generalized Additive Model Sensitivity Analysis

Model	Latitude/Longitude *p*-value	AIC
Bivariate thin plate regression spline model for latitude/longitude, age and sex covariates*	IIa: 0.127 IIb: <0.001 Rare: 0.692	1337
Intercept only	N/A	1396
Univariate thin plate regression spline models for latitude and longitude	IIa latitude: 0.022 IIa longitude: 0.967 IIb latitude: <0.001 IIb longitude: <0.001 Rare latitude: 0.399 Rare longitude: 0.734	1338
Bivariate thin plate regression spline model for latitude/longitude	IIa: 0.071 IIb: <0.001 Rare: 0.688	1340
Bivariate thin plate regression spline model for latitude/longitude, age and sex covariates, basis dimension doubled	IIa: 0.127 IIb: <0.001 Rare: 0.691	1336
Cubic regression spline models for latitude and longitude, age and sex covariates	IIa latitude: 0.042 IIa longitude: 0.845 IIb latitude: <0.001 IIb longitude: <0.001 Rare latitude: 0.425 Rare longitude: 0.646	1336
Bivariate tensor product spline model for latitude/longitude, age and sex covariates	IIa: 0.077 IIb: <0.001 Rare: 0.860	1338
Bivariate thin plate regression spline model for latitude/longitude, age and sex covariates, using lineage IIa as the comparator instead of Ib	Ib: 0.127 IIb: <0.001 Rare: 0.189	1969
Bivariate thin plate regression spline model for latitude/longitude; age, sex, and year covariates	IIa: 0.104 IIb: <0.001 Rare: 0.739	1273
Thin plate regression spline models for latitude/longitude (bivariate) and year (univariate), age and sex covariates	IIa: 0.116 IIb: <0.001 Rare: 0.730	1237
Trivariate thin plate regression spline model for latitude/longitude/year, age and sex covariates	IIa latitude/longitude/year: <0.001 IIb latitude/longitude/year: <0.001 Rare latitude/longitude/year: 0.475	1174

*Primary model

Table S7. Dixon Nearest-neighbor Contingency Table Analysis of Spatial Segregation

Lineage	df	χ-square	*p*-value
Overall	12	96.19	<0.001
Ib	3	8.02	0.046
IIa	3	15.08	0.002
IIb	3	75.61	<0.001
Rare	3	4.04	0.257

Abbreviation: df, degrees of freedom

Table S8. Pairwise Segregation of Lineages Using Dixon's Nearest-neighbor Contingency Table Method

From	To	Observed Count	Expected Count	*S*	*Z*-score	*p*-value
Ib	Ib	343	308.84	0.10	2.61	0.009
Ib	IIa	115	137.26	-0.10	-2.19	0.028
Ib	IIb	92	105.06	-0.07	-1.44	0.150
Ib	Rare	36	34.84	0.02	0.21	0.832
IIa	Ib	122	137.26	-0.10	-1.80	0.072
IIa	IIa	90	60.67	0.24	3.61	<0.001
IIa	IIb	40	46.61	-0.08	-1.08	0.280
IIa	Rare	8	15.46	-0.30	-2.00	0.046
IIb	Ib	80	105.06	-0.22	-3.42	<0.001
IIb	IIa	24	46.61	-0.35	-3.80	<0.001
IIb	IIb	91	35.50	0.59	8.50	<0.001
IIb	Rare	4	11.83	-0.49	-2.39	0.017
Rare	Ib	43	34.84	0.22	1.98	0.047
Rare	IIa	11	15.46	-0.18	-1.30	0.195
Rare	IIb	9	11.83	-0.14	-0.91	0.362
Rare	Rare	3	3.86	-0.12	-0.36	0.717

Table S9. Association of Known Risk Factors with Phylogenetic Lineage

Variable	Statewide Frequency	Statewide OR (95% CI)	Southwest Region (n = 234) OR (95% CI)	Northwest Region (n = 289) OR (95% CI)	South-Central Region (n = 109) OR (95% CI)
Hispanic ethnicity (vs. non-Hispanic)					
Lineage Ib	46/372	Ref	Ref	Ref	Ref
Lineage IIa	32/197	1.13 (0.67, 1.91)	0.3 (0.03, 2.86)	2.79 (0.66, 11.83)	0.87 (0.33, 2.25)
Lineage IIb	19/152	1.13 (0.61, 2.11)	0.99 (0.3, 3.33)	3.24 (0.62, 16.86)	0.73 (0.12, 4.37)
Rare lineage	6/42	1.21 (0.46, 3.15)	8.15 (0.89, 75.06)	1.98 (0.18, 21.31)	0 (0, Inf)†
American Indian (vs. white race)‡					
Lineage Ib	5/377	Ref	Ref	Ref	Ref
Lineage IIa	7/196	3.82 (1.13, 12.95)§	-	-	-
Lineage IIb	0/148	0 (0, Inf)†	-	-	-
Rare lineage	0/40	0 (0, Inf)†	-	-	-
Asian race (vs. white race)‡					
Lineage Ib	24/377	Ref	Ref	Ref	Ref
Lineage IIa	7/196	0.53 (0.22, 1.28)	-	-	-
Lineage IIb	19/148	2.03 (1.02, 4.01)§	-	-	-
Rare lineage	2/40	0.72 (0.16, 3.22)	-	-	-
Black race (vs. white race)‡					
Lineage Ib	12/377	Ref	Ref	Ref	Ref
Lineage IIa	5/196	0.81 (0.27, 2.43)	-	-	-
Lineage IIb	5/148	1.02 (0.34, 3.06)	-	-	-

Rare lineage	0/40	0 (0, Inf)†	-	-	-
Other/multiple race (vs. white race)‡					
Lineage Ib	16/377	Ref	Ref	Ref	Ref
Lineage IIa	9/196	0.94 (0.39, 2.23)	-	-	-
Lineage IIb	11/148	1.59 (0.69, 3.68)	-	-	-
Rare lineage	1/40	0.55 (0.07, 4.32)	-	-	-
Contact with a lab-confirmed case					
Lineage Ib	59/531	Ref	Ref	Ref	Ref
Lineage IIa	39/228	1.34 (0.84, 2.15)	0.88 (0.3, 2.6)	1.48 (0.63, 3.49)	0.99 (0.25, 3.96)
Lineage IIb	43/176	1.96 (1.21, 3.16)¶	2.7 (1.15, 6.31)§	2.03 (0.78, 5.25)	2.74 (0.44, 17.21)
Rare lineage	3/60	0.41 (0.12, 1.37)	0.42 (0.05, 3.82)	0.39 (0.05, 3.24)	0 (0, Inf)†
Epidemiologic link to a confirmed or probable case					
Lineage Ib	74/522	Ref	Ref	Ref	Ref
Lineage IIa	41/221	1.25 (0.80, 1.96)	1.07 (0.37, 3.05)	0.97 (0.42, 2.25)	0.99 (0.24, 3.98)
Lineage IIb	51/172	1.94 (1.24, 3.03)¶	2.17 (0.94, 4.98)	1.41 (0.56, 3.55)	4.72 (0.85, 26.07)
Rare lineage	3/60	0.32 (0.10, 1.06)	0.33 (0.04, 2.95)	0.29 (0.04, 2.39)	0 (0, Inf)†
Underlying illness					
Lineage Ib	66/530	Ref	Ref	Ref	Ref
Lineage IIa	27/233	1.20 (0.70, 2.06)	2.87 (0.86, 9.61)	0.83 (0.2, 3.37)	4.07 (0.5, 33.02)
Lineage IIb	19/184	1.11 (0.61, 2.01)	1.17 (0.36, 3.77)	0.73 (0.15, 3.59)	6.07 (0.33, 111.66)
Rare lineage	2/62	0.19 (0.04, 0.85)§	0.59 (0.06, 5.84)	0.42 (0.05, 3.73)	0 (0, Inf)†
Contact with diapered or incontinent child or adult					

Lineage Ib	122/545	Ref	Ref	Ref	Ref
Lineage IIa	65/231	1.10 (0.75, 1.61)	0.91 (0.37, 2.22)	0.94 (0.42, 2.1)	1.43 (0.54, 3.79)
Lineage IIb	60/187	1.28 (0.86, 1.91)	1.57 (0.76, 3.26)	1.58 (0.67, 3.73)	0.82 (0.13, 5.17)
Rare lineage	8/62	0.53 (0.24, 1.16)	1.44 (0.36, 5.72)	0.19 (0.02, 1.52)	0.69 (0.06, 7.7)
Attends childcare or preschool					
Lineage Ib	39/523	Ref	Ref	Ref	Ref
Lineage IIa	22/235	DNC	2.7 (0.68, 10.64)	1.7 (0.42, 6.86)	1.19 (0.21, 6.56)
Lineage IIb	27/181	DNC	3.17 (1.03, 9.7)§	2.16 (0.55, 8.57)	0 (0, Inf)†
Rare lineage	0/59	DNC	0 (0, Inf)†	0 (0, Inf)†	0 (0, Inf)†
Employed as a health care worker					
Lineage Ib	17/525	Ref	Ref	Ref	Ref
Lineage IIa	8/232	DNC	3.06 (0.44, 21.55)	0.7 (0.06, 8.42)	0 (0, Inf)†
Lineage IIb	7/182	DNC	0.71 (0.06, 8.23)	1.52 (0.15, 15.38)	2.41 (0.18, 33.1)
Rare lineage	1/62	DNC	0 (0, Inf)†	0 (0, Inf)†	0 (0, Inf)†
Employed as a food worker					
Lineage Ib	18/539	Ref	Ref	Ref	Ref
Lineage IIa	12/244	1.64 (0.74, 3.59)	1.4 (0.1, 19.6)	1.58 (0.45, 5.56)	0 (0, Inf)†
Lineage IIb	4/188	0.74 (0.24, 2.28)	1.61 (0.21, 12.62)	0.53 (0.06, 4.44)	0 (0, Inf)†
Rare lineage	2/60	0.99 (0.22, 4.41)	0 (0, Inf)†	1.11 (0.13, 9.77)	0 (0, Inf)†
Works with animals or animal products					
Lineage Ib	24/524	Ref	Ref	Ref	Ref
Lineage IIa	5/196	0.46 (0.16, 1.27)	0 (0, Inf)†	0.31 (0.04, 2.57)	0.87 (0.12, 6.08)

Lineage IIb	5/163	0.84 (0.30, 2.40)	0.77 (0.11, 5.45)	0 (0, Inf)†	2.17 (0.19, 24.56)
Rare lineage	3/53	1.14 (0.33, 4.00)	2.87 (0.25, 33.45)	1.73 (0.32, 9.22)	0 (0, Inf)†
Any contact with animals					
Lineage Ib	300/521	Ref	Ref	Ref	Ref
Lineage IIa	115/200	0.81 (0.57, 1.15)	0.84 (0.34, 2.09)	0.56 (0.26, 1.24)	0.48 (0.18, 1.3)
Lineage IIb	90/167	0.78 (0.54, 1.14)	1.9 (0.89, 4.06)	0.48 (0.2, 1.15)	0.16 (0.03, 0.88)§
Rare lineage	27/52	0.8 (0.44, 1.45)	Inf (0, Inf)†	0.73 (0.24, 2.26)	0.3 (0.02, 3.63)
Contact with cattle, cows, or calves					
Lineage Ib	63/471	Ref	Ref	Ref	Ref
Lineage IIa	30/188	1.06 (0.64, 1.78)	1.11 (0.32, 3.81)	0.68 (0.26, 1.78)	1.06 (0.36, 3.12)
Lineage IIb	13/151	0.59 (0.3, 1.14)	1.04 (0.38, 2.84)	0.14 (0.02, 1.07)	0 (0, Inf)†
Rare lineage	7/49	1.19 (0.5, 2.81)	0.95 (0.1, 8.81)	0.92 (0.24, 3.54)	0 (0, Inf)†
Case or household member lives or works on a farm or dairy					
Lineage Ib	67/526	Ref	Ref	Ref	Ref
Lineage IIa	24/191	0.86 (0.50, 1.46)	0.47 (0.09, 2.5)	0.96 (0.37, 2.44)	1.06 (0.36, 3.13)
Lineage IIb	15/169	0.67 (0.35, 1.27)	1.62 (0.58, 4.48)	0 (0, Inf)†	0.33 (0.04, 2.95)
Rare lineage	7/53	1.08 (0.47, 2.52)	1.35 (0.14, 13)	0.99 (0.26, 3.82)	1.39 (0.11, 17.56)
Visited a zoo, farm, fair, or pet shop					
Lineage Ib	99/526	Ref	Ref	Ref	Ref
Lineage IIa	49/200	1.31 (0.86, 2)	1.59 (0.61, 4.17)	0.93 (0.41, 2.12)	1 (0.28, 3.53)
Lineage IIb	25/166	0.59 (0.35, 1)§	0.88 (0.4, 1.94)	0.17 (0.04, 0.78)§	0 (0, Inf)†
Rare lineage	11/53	1.11 (0.53, 2.33)	0.52 (0.06, 4.65)	0.6 (0.16, 2.32)	2.65 (0.2, 34.74)

Recreational water exposure					
Lineage Ib	130/548	Ref	Ref	Ref	Ref
Lineage IIa	57/229	0.96 (0.65, 1.41)	0.51 (0.18, 1.45)	0.53 (0.24, 1.17)	0.79 (0.25, 2.56)
Lineage IIb	38/174	0.82 (0.53, 1.27)	0.38 (0.16, 0.93)§	0.44 (0.16, 1.24)	6.39 (1.09, 37.47)§
Rare lineage	12/60	0.79 (0.40, 1.57)	0.66 (0.13, 3.41)	0.77 (0.22, 2.73)	1.41 (0.12, 16.12)
Drank untreated/unchlorinated water					
Lineage Ib	61/531	Ref	Ref	Ref	Ref
Lineage IIa	29/219	0.96 (0.58, 1.57)	4.49 (1.48, 13.57)¶	0.89 (0.27, 2.87)	0.16 (0.04, 0.63)¶
Lineage IIb	26/169	1.27 (0.74, 2.16)	3.76 (1.38, 10.28)¶	1.5 (0.44, 5.15)	0.27 (0.03, 2.38)
Rare lineage	7/53	1.14 (0.49, 2.66)	1.68 (0.29, 9.69)	2.14 (0.41, 11.07)	0 (0, Inf)†
Well is source of drinking water					
Lineage Ib	136/559	Ref	Ref	Ref	Ref
Lineage IIa	59/236	0.91 (0.62, 1.32)	1.1 (0.47, 2.54)	1.1 (0.5, 2.41)	0.47 (0.19, 1.17)
Lineage IIb	35/186	0.77 (0.48, 1.21)	1.06 (0.52, 2.12)	0.7 (0.24, 2)	0.08 (0.01, 0.72)§
Rare lineage	14/62	0.87 (0.46, 1.65)	0.49 (0.11, 2.09)	1.13 (0.33, 3.84)	0.18 (0.02, 1.73)
Consumed food from a restaurant					
Lineage Ib	384/505	Ref	Ref	Ref	Ref
Lineage IIa	166/216	1.22 (0.81, 1.83)	1.82 (0.69, 4.81)	0.93 (0.4, 2.17)	0.66 (0.25, 1.72)
Lineage IIb	132/171	1.09 (0.7, 1.68)	1.09 (0.5, 2.39)	0.72 (0.29, 1.78)	Inf (0, Inf)†
Rare lineage	43/54	1.23 (0.61, 2.49)	0.74 (0.19, 2.82)	0.82 (0.24, 2.79)	1.61 (0.15, 17.53)
Consumed food from a group meal					
Lineage Ib	144/531	Ref	Ref	Ref	Ref

Lineage IIa	65/227	1.1 (0.77, 1.59)	0.53 (0.19, 1.48)	1.56 (0.72, 3.39)	0.73 (0.28, 1.92)
Lineage IIb	59/179	1.24 (0.84, 1.82)	1.18 (0.58, 2.4)	2.45 (1.06, 5.71)§	0.27 (0.03, 2.52)
Rare lineage	17/58	1.16 (0.64, 2.13)	0.58 (0.12, 2.86)	3.1 (1.02, 9.4)§	0.46 (0.04, 4.8)
Handled raw meat					
Lineage Ib	122/542	Ref	Ref	Ref	Ref
Lineage IIa	43/226	0.86 (0.54, 1.38)	1.21 (0.4, 3.64)	0.75 (0.3, 1.88)	1.5 (0.37, 6.14)
Lineage IIb	31/182	0.92 (0.55, 1.53)	1.41 (0.55, 3.61)	0.23 (0.05, 1.08)	0.51 (0.07, 3.9)
Rare lineage	15/62	1.09 (0.54, 2.17)	1.47 (0.33, 6.47)	0.62 (0.15, 2.49)	2.14 (0.17, 27.6)
Consumed meat					
Lineage Ib	314/521	Ref	Ref	Ref	Ref
Lineage IIa	138/223	1.09 (0.77, 1.53)	1.09 (0.48, 2.47)	1.33 (0.59, 2.99)	1.45 (0.58, 3.62)
Lineage IIb	106/175	1.07 (0.74, 1.55)	1.25 (0.64, 2.44)	1.83 (0.63, 5.37)	1.3 (0.28, 6.08)
Rare lineage	31/56	0.75 (0.43, 1.33)	0.59 (0.16, 2.13)	0.46 (0.15, 1.42)	0.77 (0.11, 5.23)
Consumed ground beef					
Lineage Ib	331/539	Ref	Ref	Ref	Ref
Lineage IIa	132/229	0.85 (0.61, 1.18)	0.94 (0.39, 2.3)	0.88 (0.43, 1.8)	0.82 (0.32, 2.09)
Lineage IIb	103/180	0.85 (0.59, 1.22)	0.87 (0.43, 1.76)	0.27 (0.11, 0.65)¶	0.82 (0.18, 3.78)
Rare lineage	31/57	0.76 (0.44, 1.34)	1.52 (0.29, 8.01)	0.6 (0.22, 1.69)	0.31 (0.04, 2.14)
Consumed intact beef					
Lineage Ib	283/462	Ref	Ref	Ref	Ref
Lineage IIa	116/185	1.07 (0.74, 1.56)	0.55 (0.2, 1.5)	0.87 (0.4, 1.89)	2.81 (0.85, 9.3)
Lineage IIb	90/156	0.86 (0.58, 1.28)	0.96 (0.42, 2.17)	0.35 (0.14, 0.87)§	1.36 (0.22, 8.52)

Rare lineage	29/46	1.17 (0.61, 2.27)	2.77 (0.3, 25.43)	1.54 (0.43, 5.51)	0.31 (0.01, 7.79)
Consumed venison or other wild game meat					
Lineage Ib	15/521	Ref	Ref	Ref	Ref
Lineage IIa	3/195	0.37 (0.08, 1.68)	0 (0, Inf)†	0 (0, Inf)†	0 (0, Inf)†
Lineage IIb	10/169	1.97 (0.81, 4.79)	1.35 (0.4, 4.58)	1.22 (0.13, 11.12)	0 (0, Inf)†
Rare lineage	5/53	3.56 (1.23, 10.32)§	1.56 (0.16, 14.98)	3.56 (0.58, 21.96)	34.96 (1.03, 1187.37)§
Consumed raw milk					
Lineage Ib	16/551	Ref	Ref	Ref	Ref
Lineage IIa	6/232	0.82 (0.3, 2.23)	4.04 (0.22, 75.92)	0.38 (0.04, 3.72)	0 (0, Inf)†
Lineage IIb	18/183	2.46 (1.15, 5.28)§	17.33 (2.05, 146.5)¶	0 (0, Inf)†	24.32 (0.81, 726.95)
Rare lineage	1/60	0.63 (0.08, 4.88)	0 (0, Inf)†	0 (0, Inf)†	0 (0, Inf)†
Consumed unpasteurized juice					
Lineage Ib	11/496	Ref	Ref	Ref	Ref
Lineage IIa	3/219	0.34 (0.09, 1.27)	0.8 (0.11, 6.04)	0 (0, Inf)†	0 (0, Inf)†
Lineage IIb	7/163	1.53 (0.55, 4.29)	0.6 (0.09, 4.03)	7.08 (0.37, 137.1)	5.9 (0.35, 100.4)
Rare lineage	3/55	2.31 (0.61, 8.78)	2.39 (0.21, 27.47)	23.08 (1.52, 351.69)§	0 (0, Inf)†
Consumed raw fruits or vegetables					
Lineage Ib	435/514	Ref	Ref	Ref	Ref
Lineage IIa	184/205	1.81 (1.05, 3.11)§	6.88 (0.84, 56.67)	2.55 (0.52, 12.41)	1.34 (0.43, 4.16)
Lineage IIb	144/170	1.25 (0.74, 2.1)	1.51 (0.62, 3.64)	0.78 (0.23, 2.6)	1.97 (0.2, 19.15)
Rare lineage	43/48	1.5 (0.57, 4)	Inf (0, Inf)†	2.11 (0.25, 17.82)	0.37 (0.02, 5.85)
Consumed sprouts					

Lineage Ib	22/537	Ref	Ref	Ref	Ref
Lineage IIa	12/231	1.45 (0.68, 3.11)	1.87 (0.23, 15.21)	2.98 (0.57, 15.62)	Inf (0, Inf)†
Lineage IIb	12/180	2 (0.94, 4.27)	1.11 (0.17, 7.45)	5.17 (1.04, 25.74)§	0.5 (0, Inf)
Rare lineage	4/57	1.94 (0.64, 5.94)	0 (0, Inf)†	7.32 (1.11, 48.28)§	0.24 (0, Inf)
Consumed fresh herbs					
Lineage Ib	102/524	Ref	Ref	Ref	Ref
Lineage IIa	44/216	0.83 (0.54, 1.27)	0.95 (0.32, 2.79)	0.88 (0.37, 2.1)	0.19 (0.04, 0.77)§
Lineage IIb	35/178	1.01 (0.64, 1.6)	0.78 (0.29, 2.13)	1.51 (0.59, 3.85)	0.39 (0.04, 3.57)
Rare lineage	9/56	0.7 (0.32, 1.55)	0 (0, Inf)†	1.11 (0.29, 4.3)	0.39 (0.03, 4.47)
Traveled outside the state, the country, or usual routine					
Lineage Ib	143/571	Ref	Ref	Ref	Ref
Lineage IIa	52/246	0.78 (0.53, 1.13)	0.45 (0.17, 1.19)	0.37 (0.14, 1)	1.09 (0.34, 3.54)
Lineage IIb	54/197	1.08 (0.74, 1.59)	0.86 (0.44, 1.7)	1.71 (0.73, 4)	1.53 (0.26, 9.01)
Rare lineage	26/64	2.03 (1.17, 3.50)§	0.66 (0.16, 2.65)	3.72 (1.27, 10.87)§	7.45 (1.03, 54.07)§

All analyses are multinomial logistic regression, using lineage Ib as the reference group, adjusted for age, sex, and year. The statewide analysis was conducted using a generalized additive model to additionally adjust for latitude and longitude using a thin plate spline bivariate smooth.

Abbreviations: CI, confidence interval; DNC, did not converge; Inf, infinity; OR, odds ratio; Ref, reference

†Odds ratios of 0 are reported where 0 cases of the lineage under analysis existed in the category. Odds ratios of infinity are reported where 0 cases of the reference lineage (Ib) existed in the category. Confidence intervals were not estimated for these ORs, indicated by (0, Inf).

‡Analyses with a dash could not be performed or were considered unreliable because of sparse data in these categories. Not all models converged because of sparse data in some categories.

§$p < 0.05$

¶$p < 0.01$