Distribution Agreement

Signature:

______________________ ____________________
Karen Ellis August 1, 2018

Systematic Review and Meta-Analysis of the Global Distribution and Diversity of Norovirus GII.4 Variants

By

Karen Ellis

MPH

Global Epidemiology

Benjamin A. Lopman, PhD

Faculty Thesis Advisor

Systematic Review and Meta-Analysis of the Global Distribution and Diversity of Norovirus GII.4 Variants

By

Karen Ellis
Bachelor of Science
University of Wisconsin-Madison
2012

Faculty Thesis Advisor: Benjamin A. Lopman, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Epidemiology
2018

Abstract

Systematic Review and Meta-Analysis of the Global Distribution and Diversity of Norovirus GII.4 Variants

By Karen Ellis

Noroviruses are the leading cause of outbreaks and cases of non-bacterial acute gastroenteritis (AGE) worldwide. They pose an increasing threat to morbidity and mortality globally, causing 200,000 deaths annually, and contribute a significant burden on health systems in both high- and low-income countries. Noroviruses can be divided into seven genogroups, three of which are associated with human disease and can be further broken down into over 30 genotypes. The vast majority of outbreaks (including all pandemic outbreaks) and sporadic cases of norovirus, however, can be attributed to a single, rapidly-evolving genotype- GII.4, and its many variants. This goal of this study was to describe the global distribution and diversity of GII.4 variants over the time period that they have predominated human norovirus infection (early 1990s-present), and to quantify the effects of these variables on GII.4 diversity. To do this, we conducted a systematic review and meta-analysis of genotyping studies spanning the years of 1995-2016. 41 studies from 26 countries together provided data on 59 GII.4 variants, which were classified into eight pandemic-causing variant types in this study. Simpson's and Shannon's diversity indices were used as metrics for GII.4 variant diversity quantification. Linear regression techniques were performed on these indices to model any association between geographic, demographic, and temporal data (collected in the systematic review), and GII.4 diversity. The results of our descriptive analyses corroborated previous observations of the 2-3-year emergence of novel pandemic-causing GII.4 variants and demonstrated a varied pattern of their global distribution. Linear regression analyses indicated a weak positive relationship between GII.4 diversity and age, and a strong positive relationship between GII.4 diversity and time. The results of this review have strong implications for the future of norovirus vaccine development and implementation, specifically indicating possible benefits of targeting certain age groups and the importance of monitoring evolving pandemic variants. Additionally, they highlight the need for systematic norovirus genotype reporting, particularly in low-income environments.

Systematic Review and Meta-Analysis of the Global Distribution and Diversity of Norovirus GII.4 Variants

By

Karen Ellis
Bachelor of Science
University of Wisconsin-Madison
2012

Faculty Thesis Advisor: Benjamin A. Lopman, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Global Epidemiology
2018

Acknowledgements

I would like to acknowledge and thank the following people who have supported me throughout the completion of this Master's thesis: my thesis advisor, Dr. Ben Lopman, for patiently and attentively providing guidance and peerless expertise throughout this process; Molly Steele, for helping guide my research questions with thoughtfulness and knowledge; the numerous authors that took the pains of providing me with usable data; and lastly, Cory Arrouzet, for not only being a patient, intelligent, and hard-working colleague, but a stellar friend in the end.

Table of Contents

CHAPTER 1

Literature Review

Norovirus Burden

Noroviruses are the leading cause of outbreaks and cases of non-bacterial acute gastroenteritis (AGE) worldwide. Their estimated prevalence in all gastroenteritis cases globally is 18% [1] and they cause one fifth of all diarrheal cases, making them the most common cause of diarrhea worldwide [2]. Other symptoms of norovirus often include nausea, vomiting and abdominal pain and are generally self-limiting. Nonetheless, norovirus contributes over 200,000 deaths annually [2]. This burden is particularly high in low-income countries, where diarrhea is among the leading causes of death among children under 5 years of age.

Rotavirus has also historically been a significant contributor of AGE cases globally, particularly among children under 5. However, with the development and increasing ubiquity of rotavirus vaccines, norovirus is likely to remain in the lead [2]. Morbidity and mortality due to noroviruses are especially high in groups such as young children, the elderly, and the immunocompromised, however they are known to affect people among all age groups in both developed and developing countries. Additionally, they can incur high economic costs as they are a common healthcare acquired infection and are difficult to control. This has led to the closure of wards, increased length of hospitalization among inpatients, the hiring of extra personnel, and the requirement for extra supplies [3].

General Norovirus Epidemiology

In the U.S., norovirus is the leading cause of foodborne illness, causing 58% of all cases [4]. Noroviruses are transmitted via the fecal-oral route, person-to-person, and via contaminated food or water. They are highly infectious- the estimated mean infectious dose is around 18 particles [5]- and are easily spread in healthcare facilities, nursing homes, and schools, commonly leading to outbreaks in these settings. Though clinical symptoms generally only last between two and three days, viral shedding of greater than $1x10^{10}$ RNA copies per gram of stool [6] can last for several weeks beyond the symptomatic phase [7]. Norovirus infection does not induce immunity to reinfection by other genotypes, increasing the burden of disease [8].

The first recorded norovirus outbreak occurred in an elementary school Norwalk, Ohio in 1968. However, illness due to the virus was described as early as 1929 and referred to as “winter vomiting disease” because of its apparent winter seasonality [8]. Before this event, an etiologic agent for infectious AGE had never been determined [9]. Since its initial characterization, norovirus has displayed and maintained its widespread global distribution. Understanding its distribution and diversity has major potential implications for vaccine development.

Molecular Epidemiology of Norovirus

1. Genetic Structure

Norovirus is a 7.5 kb single-stranded RNA virus of the *Caliciviridae* family. Structurally, it is comprised of three open reading frames (ORF1, ORF2, and ORF3). The ORF1 contains non-structural proteins that are used in replication, including the RNA polymerase. ORF2 encodes the major capsid protein (VP1) and ORF3 encodes a minor structural capsid protein (VP2) linked to VP1 stability [10]. ORF1 and ORF2 are most commonly used for Norovirus genotyping and are generally referred to throughout the literature as the “polymerase” and “capsid” regions, respectively. Due to the increasing occurrence of recombinant genotypes between these two regions, it is important to note which region is used for genotyping when reporting Norovirus genotypes. In recent years, it has become increasingly common to genotype the ORF1/ORF2 overlap region to identify these genetic recombinants.

2. Genotype Diversity and Distribution

Noroviruses have historically been divided into six genogroups (GI-GVI), with a seventh (GVII) recently proposed [11]. However, only three are known to infect humans (genogroups GI, GII, and GIV), with genogroup GII estimated to be the cause of greater than 95% of all human Norovirus infections[12]. The two major genogroups associated with human disease, GI and GII, can be broken down into 28 genotypes [13], with the GII.4 genotype being further divided into variants and subvariants.

GII.4 Variant Evolution

1. GII.4 Distribution

Though the vast genotypic diversity is largely responsible for high norovirus prevalence and persistence around the globe, the majority of both sporadic cases and outbreaks worldwide can be attributed to the GII.4 genotype and its variants, which are estimated to account for >80% of all norovirus infections at any given time [3]. Additionally, it is the only genotype associated with global pandemics of gastroenteritis [12]. GII.4 occurrence was first documented in 1987 and became recognized as a major epidemic strain in the mid-1990s [14]. Since then, pandemic GII.4 variants have emerged every 2-3 years, replacing existing strains and causing recurring pandemic outbreaks [3].

As no cell culture system exists for Norovirus, it is not currently possible to segregate them using biologically-relevant criteria [15, 16]. This has made it difficult to establish a unified classification and naming scheme for GII.4 variants and subvariants since the strain's emergence. Though the use of RT-PCR and sequencing methods is widespread, lack of consensus regarding typing schemes for Norovirus genotypes and variants has resulted in confusion, disagreement and misclassification, and has prevented the development of a standardized classification scheme for GII.4 variants. Norovirus genotypes were initially classified based on the complete sequence of the VP1 region. New genotypes were assigned when amino acids in this region differed by more than 20% compared to other genotypes [17]. However, as Norovirus genotypes have become more diverse with increasing rapidity in recent years, it has been proposed that this divergence be reduced to 14.1% to account for within-genotype diversity [18]. For GII.4 specifically, it has been proposed that variants have >5% amino acid divergence [19] and that subvariants have up to 2.8% divergence [20]. Due to the propensity for GII.4 Norovirus to cause widespread, rapidly emerging pandemics, it is important that a unified classification and naming scheme be established. A study published in 2013 by Annelies Kroneman et al. proposes a phylogenetic, rather than pairwise distance cutoff approach (as described above) to classifying GII.4 variants. Based on this criteria, Kroneman et al. classified GII.4 pandemic variants into the following eight groups: US95_96, Farmington_Hills_2002, Asia_2003, Hunter_2004, Yerseke_2006a, Den_Haag_2006b, New_Orleans_2009, Sydney_2012 [16]. Due to the increasing occurrence and relevance of recombinant genotypes, Kroneman et al. additionally suggest that a dual nomenclature, based on both VP1 and ORF1 sequences, be used to further classify Norovirus genotypes.

2. GII.4 Evolution

The ability of GII.4 strains to evolve rapidly has led to the emergence of numerous pandemic variants over the last two decades in quick succession of one another. It has been observed that highly transmissible viruses that cause acute illness and short-lived epidemics often display the most complex global behavior as the result of a three-way interplay between transmission, host herd immunity, and viral adaptation [21]. The evolution of the GII.4 lineage has therefore likely been driven by a number of mechanisms producing both antigenic drift and antigenic shift.

2.1. Antigenic Drift

High GII.4 virulence is thought to be attributed to its ability to bind to a wider group of histo-blood group antigens (HBGAs), which are composed of antigenically-distinct carbohydrate core structures that are believed to facilitate this binding [22]. This provides a larger susceptible population for GII.4 Norovirus compared to other Norovirus genotypes. Early GII.4 pandemic variants have demonstrated mutations within the protruding (P) domain of the capsid region, perhaps as a result of higher replication and mutation rates compared to other genotypes [23]. Significant variations between five evolving blockade epitopes (A-E) of different GII.4 variants within the P2 capsid domain demonstrate that antigenic variation is a probable contributor to the epochal emergence of novel GII.4 variant. Accumulation of mutations at these five sites allow GII.4 Noroviruses to rapidly produce new variants that are able to escape host immunity. Therefore, there is strong evidence that GII.4 Noroviruss, and specifically their P2 capsid region, are under strong selective pressure from host immunity [12]. This theory explains the reemergence of norovirusel pandemic variants every 2-3 years as newly evolved variants displace the previous pandemic variant by escaping the current herd immunity.

2.2. Antigenic Shift

The two most recent pandemic variants, New_Orleans_2009 and Sydney_2012, differ from the previous five variants in that they show not only capsid evolution, but intragenomic recombination at the ORF1/ORF2 overlap region, demonstrating evolution via processes of both antigenic drift (point mutations, deletions, etc.) and antigenic shift (recombination). Recombinant genotypes are nearly ubiquitous among non-GII.4 genotypes [12], indicating not only how prominent RNA recombination is as an evolutionary mechanism for viruses, but that it may be the main source of future pandemic variants.

Future Pandemic Variants

Though the GII.4 genotype has predominated as the perpetrator of global outbreaks and sporadic cases over the last twenty years, numerous reports have documented the emergence of GII.17 (namely the Kawasaki 308 variant) as a dominant genotype in outbreaks and sporadically-detected cases occurring in the 2014-2016 seasons [24, 25]. This variant is different from other GII.17 variants in circulation as a result of two amino acid insertions that are surface-exposed in the major capsid (VP1) region [26]. One of these insertions took place at a region that correlates to the GII.4 VPI region that comprises the HGBA-interface, having potential implications for shifts in its host-binding preference and therefore its ability to escape host immunity. Thus, it is postulated that this GII.17 variant may spread globally, replacing GII.4 variants as the predominating genotype in some parts of the world [24, 27].

Global Serotype Diversity/Host Immune Factors

Susceptibility to norovirus varies in the general population depending on heterogeneity in host immune factors. Variation in the degree to which norovirus infection affects different individuals was first observed in the 1970s by Parrino et. al., when they posed challenge studies on 12 different individuals and observed that only six became ill after initial infection and immediate reinfection [28]. Since this initial study, research has shown that innate resistance to norovirus in some individuals and not others is due to variation in histo-blood group antigens (HBGAs) which are expressed on the mucosal lining of the gastrointestinal tract and are known for their binding ability to caliciviruses. HGBAs are oligosaccharide epitopes that vary in structure between individuals, the determination of which is mediated by fucosyltransferases FUT1 and FUT2 [8]. The *FUT2* gene is mostly expressed in the mucosal epithelial cells that line the GI tract and controls the secretion of HGBAs at the gut surface. Among the human population, 70%-80% are thought to be "secretors" of the *FUT2* gene. The remaining 20%-30% of individuals are referred to as "non-secretors" and have demonstrated resistance to norovirus infection, particularly among certain genotypes which include GII.4 [29]. Variation both in HBGAs among the human population and in viral structure among noroviruses has unsurprisingly led to the observation that susceptibility to certain strains of norovirus differ among the population. This is also affected by acquired immunity to certain norovirus strains over an individual's lifetime. Host/pathogen variation imposes a strong selective pressure on noroviruses, largely explaining their continual rapid evolution and diversification.

Diversity Indices

Diversity among organisms, whether species, chemical, or genetic diversity, is difficult to quantify becaue there is no single measure that fully captures the concept [30]. Certain measures are commonly used throughout the field of biology, including Richness (S), which is simply the number species/attributes in a sample, and Shannon's (H') and Simpson's (D) diversity indices, which account for the presence of both richness and proportional abundance [31].

These indices, while both representative of diversity, differ in the theory from which they are derived as well as their direct interpretations. Shannon's index comes from information science and represents the uncertainty about the identity of an unknown individual. For example, in a sample that is highly diverse and evenly distributed, predictions about the identity of a given individual have high uncertainty, and vice versa for populations with low diversity and uneven distribution [30]. Values of Shannon's diversity index range from 0 to the log of 1/number of categories ([31]), with higher numbers representing greater uncertainty and therefore greater diversity. Simpson's diversity index is the probability that any two randomly-chosen individuals in a sample belong to the same group [30]. Values of this index range from 0 to 1-1/number of categories, approaching 1 as the number of categories increases. Like Shannon's index, larger numbers imply greater diversity [31].

Both indices have their advantages and disadvantages. Shannon’s index tends to be affected more by species richness and less abundant species, making it sensitive to smaller changes in diversity. Simpson’s index, however, puts more weight on dominant species and evenness. It is common, therefore, that both indices are used in parallel.

CHAPTER 2

Systematic Review and Meta-Analysis of the Global Distribution and Diversity of Norovirus GII.4 Variants

Authors: Karen Ellis, Cory Arrouzet, Molly Steele, Benjamin Lopman

Abstract

Noroviruses are the leading cause of outbreaks and cases of non-bacterial acute gastroenteritis (AGE) worldwide. They pose an increasing threat to morbidity and mortality globally, causing 200,000 deaths annually, and contribute a significant burden on health systems in both high- and low-income countries. Noroviruses can be divided into seven genogroups, three of which are associated with human disease and can be further broken down into over 30 genotypes. The vast majority of outbreaks (including all pandemic outbreaks) and sporadic cases of norovirus, however, can be attributed to a single, rapidly-evolving genotype- GII.4, and its many variants. This goal of this study was to describe the global distribution and diversity of GII.4 variants over the time period that they have predominated human norovirus infection (early 1990s-present), and to quantify the effects of these variables on GII.4 diversity. To do this, we conducted a systematic review and meta-analysis of genotyping studies spanning the years of 1995-2016. 41 studies from 26 countries together provided data on 59 GII.4 variants, which were classified into eight pandemic-causing variant types in this study. Simpson's and Shannon's diversity indices were used as metrics for GII.4 variant diversity quantification. Linear regression techniques were performed on these indices to model any association between geographic, demographic, and temporal data (collected in the systematic review), and GII.4 diversity. The results of our descriptive analyses corroborated previous observations of the 2-3-year emergence of novel pandemic-causing GII.4 variants and demonstrated a varied pattern of their global distribution. Linear regression analyses indicated a weak positive relationship between GII.4 diversity and age, and a strong positive relationship between GII.4 diversity and time. The results of this review have strong implications for the future of norovirus vaccine development and implementation, specifically indicating possible benefits of targeting certain age groups and the importance of monitoring evolving pandemic variants. Additionally, they highlight the need for systematic norovirus genotype reporting, particularly in low-income environments.

Introduction

Noroviruses are the leading cause of non-bacterial outbreaks and cases of acute gastroenteritis (AGE) worldwide. Their estimated prevalence in all gastroenteritis cases globally is 18% [1] and they cause one fifth of all diarrheal cases, making them the most common cause of diarrhea worldwide [2]. Other symptoms of norovirus often include nausea, vomiting and abdominal pain and are generally self-limiting. Nonetheless, norovirus contributes over 200,000 deaths annually [2]. This burden is particularly high in low-income countries, where diarrhea is among the leading causes of death among children under 5 years of age. Additionally, noroviruses can incur high economic costs as they are a common healthcare acquired infection and are difficult to control. This has led to the closure of wards, increased length of hospitalization among inpatients, the hiring of extra personnel, and the requirement for extra supplies [3]. In the U.S., norovirus is the leading cause of foodborne illness, causing 58% of all cases [4].

Noroviruses are transmitted via the fecal-oral route, person-to-person, and via contaminated food or water. They are highly infectious- the estimated mean infectious dose is around 18 particles [5]- and are easily spread in healthcare facilities, nursing homes, and schools, commonly leading to outbreaks in these settings. Immunity to norovirus does not last long and generally does not protect against other genotypes, increasing the burden of disease [8].

The vast diversity of noroviruses can be attributed, in large part, to its structural make-up. Norovirus is a 7.5 kb single-stranded RNA virus of the *Caliciviridae* family. Its genome is comprised of three open reading frames (ORF1, ORF2, and ORF3). ORF1 and ORF2 are most commonly used for norovirus genotyping and are generally referred to throughout the literature as the “polymerase” and “capsid” regions, respectively. Due to the increasing occurrence of recombinant genotypes between these two regions, typing region is important to note when reporting norovirus genotypes. In recent years, it has become increasingly common to genotype the ORF1/ORF2 overlap region to identify these genetic recombinants.

Noroviruses have historically been divided into six genogroups (GI-GVI), with a seventh (GVII) recently proposed [11]. However, only three are known to infect humans (genogroups GI, GII, and GIV), with genogroup GII estimated to be the cause of greater than 95% of all human Norovirus infections[12]. These genogroups can be broken down into 38 genotypes [16], with the GII.4 genotype being further divided into variants and subvariants. Though this vast genotypic diversity is largely responsible for high norovirus prevalence and persistence around the globe, the majority of both sporadic cases and outbreaks worldwide can be attributed to the GII.4 genotype and its variants, which are estimated to account for >80% of all norovirus infections at any given time [3]. Additionally, it is the only genotype associated with global pandemics of gastroenteritis [12]. GII.4 variants emerge every 2-3 years, replacing existing strains and causing recurring pandemic outbreaks [3]. The evolution of the GII.4 lineage has likely been driven by a number of factors producing both antigenic drift (the accumulation of mutations) and antigenic shift (recombination), leading to the emergence of numerous pandemic variants over the last two decades in quick succession.

As no cell culture system exists for norovirus, it is not currently possible to segregate them using biologically-relevant criteria [15, 16]. This has made it difficult to establish a unified classification and naming scheme for GII.4 variants and subvariants since the strain's emergence. A study published in 2013 by Annelies Kroneman et al. proposes a phylogenetic, rather than pairwise distance cutoff approach (as has commonly been used throughout recent decades) to classifying GII.4 variants. Based on this classification scheme, GII.4 pandemic variants were classified into the following eight groups: US_95_96, Farmington_Hills_2002, Asia_2003, Hunter_2004, Yerseke_2006a, Den_Haag_2006b, New_Orleans_2009, Sydney_2012 [16].

Though the GII.4 genotype has predominated as the as the perpetrator of global outbreaks and sporadic cases over the last twenty years, numerous reports have documented the emergence of GII.17 (namely the Kawasaki 308 variant) as a dominant genotype in outbreaks and sporadically-detected cases occurring in the 2014-2016 seasons [24, 25]. This variant is different from other GII.17 variants in circulation as a result of two amino acid insertions that are surface-exposed in the major capsid (VP1) region [26]. One of these insertions took place at a region that correlates to the GII.4 VPI region that comprises the HGBA-interface, having potential implications for shifts in its host-binding preference and therefore its ability to escape host immunity. Thus, it is postulated that this GII.17 variant may spread globally, replacing GII.4 variants as the predominating genotype in some parts of the world [24, 27].

Susceptibility to noroviruses varies in the general population depending upon differences in certain physiological factors between individuals. Research has shown that innate resistance to norovirus in some individuals and not others is due to variation in histo-blood group antigens (HBGAs) which are expressed on the mucosal lining of the gastrointestinal tract and are known for their binding ability to caliciviruses. Secretion of HGBAs at the gut surface is mainly regulated by the *FUT2* gene. Among the human population, 70%-80% are thought to be "secretors" of the *FUT2* gene. The remaining 20%-30% of individuals are referred to as "non-secretors" and have demonstrated resistance to norovirus infection, particularly among certain genotypes which include GII.4 [29]. Variation both in HBGAs among the human population and in viral structure among noroviruses have unsurprisingly led to the observation that susceptibility to certain strains of norovirus differ among the population. This is also affected by acquired immunity to certain norovirus strains over an individual's lifetime. Host/pathogen variation imposes a strong selective pressure on noroviruses, largely explaining their continual rapid evolution and diversification.

Diversity among organisms, whether species, chemical, or genetic diversity, is difficult to quantify because there is no single measure that fully captures the concept [30]. Certain measures are commonly used throughout the field of biology, including Richness (S), which is simply the number species/attributes in a sample, and Shannon's (H') and Simpson's (D) diversity indices, which account for the presence of both richness and proportional abundance [31]. These indices, while both representative of diversity, differ in the theory from which they are derived as well as their direct interpretations. Shannon's index represents the uncertainty about the identity of an unknown individual, whereas Simpson's index is the probability that any two randomly-chosen individuals in a sample belong to the same group [30]. Because these measures express different interpretations of diversity, it is common that both indices are used in parallel.

Using both descriptive and analytic methods, the aim of this study is to characterize the patterns of GII.4 diversity on a global scale over the two-and-a-half-decade period that it has been in circulation. We performed a systematic review and meta-analysis to describe the distribution of GII.4 and its variants temporally, geographically, and demographically, and to assess the significance of these factors' relationships with quantitative measures of variant diversity using both Shannon's and Simpson's diversity indices. The results of this study have potential implications for future norovirus prevention and control interventions, particularly in the area of vaccine development and implementation.

Methods

a. Search strategy

We conducted a systematic search through the PubMed database for studies that contained information relevant to our research question (the distribution of GII.4 variant prevalence among human populations) using a combination of the term "norovirus" coupled with any of the terms "surveillance", "genotype", or "strain". All relevant papers further included in the screening process were published in English and during or after the year 1995.

b. Study selection criteria

In the first step of the screening process, any articles that were extraneous to our research question were excluded. Such studies included animal or environmental studies, non-primary studies, nosocomial studies, and single community outbreak reports. Because we were interested in the prevalence, distribution, and diversity of GII.4 variants throughout the general population, nosocomial cases and single outbreak reports could distort this distribution. Hospital cases (inpatient or outpatient) resulting from community, rather than nosocomial infections were included. Articles were not restricted by surveillance or collection method, or by study design, as there is no reason to believe that these factors would significantly skew the data. Subsequent screenings removed articles from which GII.4 variant prevalence numbers were not reported. If relevant data were presented graphs or figures, the authors were emailed and asked for raw variant numbers. Papers further restricted to subsects of the population of interest (e.g. by age, transmission method, or setting) were included and later controlled for in the analysis. Studies were restricted to those that provided three or more years of variant surveillance data in order to capture any shifts in variant distribution over the study period, as a result of the 2-3 year pattern of emergence of pandemic GII.4 variants. This is not a strict requirement for this study in particular, but for a norovirus modeling study that will makes use of the dataset created for this review. Studies from the same country were cross-checked for duplicate data.

c. Data abstraction

Data abstraction was carried out using DistillerSR systematic review software. Articles that passed the abstract screen were uploaded via EndNote X8 to the software where a form we created containing relevant information fields was used to abstract data for each study. Upon review of the full text of each article, additional studies were excluded if they were found not to meet inclusion criteria. Any excluded studies were accounted for in the abstraction form. Data on the following information were collected: first author, title of paper, journal name, year published, country, within-country region, month and year of surveillance start, month and year of surveillance end, reporting unit (outbreaks or sporadic cases), type of study population, age, transmission method, norovirus genotyping method, genotyping region, total number of genotyped GII.4, number of strain-typed GII.4, and the numbers of each reported GII.4 variant. Each observation period (minimum of three) within each study- generally assigned by the authors using either norovirus season or calendar year- was recorded using a separate form. Country-specific information regarding WHO region, UNDP Human Development Index, and Under 5 Mortality Rate were assigned after abstraction. All completed forms were exported to a Microsoft Office Excel spreadsheet.

The FUT2 secretor status dataset was provided by Yingxi Chen and Anita Kambhampti of the Centers for Disease Control and Prevention and was merged with our GII.4 dataset for analysis.

d. GII.4 variant assignment and classification

All GII.4 variants were recorded as they were presented in each study; however, due to the lack of a unified classification and naming system for GII.4 variants, there was a significant need to crosscheck variants among multiple available resources such as published literature, phylogenetic analyses, GenBank accession numbers, and NoroNet databases to avoid misclassification of variants between studies. Using these resources, variants and sub-variants were both consolidated across studies so that each variant type was represented with one name, and where appropriate, grouped into the eight pandemic clusters or separate non-epidemic clusters. Although this was an attempt to rectify redundancy in both the naming and classification of GII.4 variants, granularity of variant types was maintained where possible to preserve the representative diversity of strain evolution.

e. Analysis of data

All data cleaning, calculations, and analysis were performed using SAS version 9.4. All observations (N=261) were assigned to various levels of categorical variables (Table 1) for descriptive analysis. The relative prevalence of pandemic strains for these variables was calculated by dividing the number of each pandemic variant by the total number of pandemic variants per level of each variable. UNDP Development Index categories (used in both the descriptive and statistical analysis) and values (used in the statistical analysis) were most recently assigned by country in 2016 [32]. Under-Five Mortality Rates for each country were also acquired from the 2016 UNDP Human Development Report. These rates were arranged into quartiles for both descriptive and statistical analyses, and crude rates were used for statistical analysis. Observations were stratified into five time periods (1995-2000, 2001-2004, 2005-2008, 2009-2012, 2012-2016) to account for the temporal trends of GII.4 variant evolution observed in the literature.

Linear regression was used to model associations between predictor variables of interest and measures of GII.4 variant diversity (using Simpson's and Shannon's diversity indices). Simple linear regressions were performed on each index for under-five mortality rates, development index values, and secretor status. For each diversity measure, Tukey's method was used to make pairwise comparisons of diversity measure means between different levels of each categorical variable of interest (WHO region, time period, development index category, under-five mortality rate quartile, and age) in order to determine intragroup variation in diversity measure prediction. Finally, to determine an overall predictive model for GII.4 variant diversity, multiple linear regression was performed on each diversity measure as the dependent variable, containing all of the following predictor variables in the full model: under-five mortality rate, development index value, secretor status, age category, WHO region, reporting unit, and time period. The GLMSELECT procedure in SAS was used to select the model of best fit, using backwards elimination in combination with a stepwise procedure, and Mallow's Cp, AICC (Akaike information criterion corrected for small-sample-size), PRESS (predicted residual sum of squares statistic), and the significance level of the F statistic as selection criteria.

f. Diversity indices

Simpson's and Shannon's diversity index values were calculated for each study. Simpson's index (D) was calculated by summing the product of the number of GII.4 variants and the number of variants minus 1, for each variant type, dividing this number by the product of total number of different variants and the number of different variant types minus 1, and subtracting this value from 1.

$$D = 1 - \frac{\sum_{i=1}^{s} n(n-1)}{N(N-1)}$$

Shannon's index (H') was calculated by summing the proportion of a particular GII.4 variant relative to all variants multiplied by the natural log of this number and multiplying this sum by negative 1.

$$H' = -\sum_{i=1}^{s} (p_i)\ [\ln(p_i)]$$

Both indices are a measure the diversity of GII.4 variants over each study period, though they characterize this diversity in slightly different ways. In ecological studies, it is common practice to use both to provide a more comprehensive analysis of species diversity. Both indices were used as the dependent variable in the linear regression analyses.

Results

A total of 2528 articles were identified as the result of our PubMed search. After an initial title screen for eligibility, 943 proceeded to the abstract screen. 456 full articles were then screened and 41 underwent data abstraction to be included in the final analyses (Figure 1). 4 of these articles included observations from separate populations and were therefore split into separate studies for analysis, resulting in 45 studies overall. Together, the 41 included articles contributed 261 individual observation periods.

a. Description of Studies

Studies from 26 countries were represented in our analyses. Though every WHO region was represented, the majority of studies (n=22) came from the Western Pacific region (largely due to the overrepresentation of studies from China) and the fewest from the Eastern Mediterranean region (n=1). Similarly, most studies came from countries classified by the UNDP as having "Very High Human Development" (n=37), whereas only 1 came from a country with "Low Human Development". Under 5 Morality Rates were distributed similarly, with >75% of studies having rates of 12 or below, and <25% having rates fall in the range of 13-71. Together, all articles spanned the years of 1995 and 2016, with an average study length period of 5.26 years. The majority of studies reported GII.4 prevalence among single cases (n=27), with 18 reporting outbreaks.

The largest proportion of studies did not restrict their genotype surveillance to a specific age group (n=22), however those that did generally tended to restrict to younger age groups. Only two studies focused on populations aged 15 or older. The setting of study populations was generally not well defined among studies (i.e. few authors were clear regarding the types of populations their data were drawn from, and there did not seem to be a systematic way of reporting this.) We were able to define these population settings in the following ways: inpatients, outpatients, community, country-wide surveillance, hospitals and/or nursing homes, childcare centers, and otherwise not specified. These settings often overlapped, as is accounted for in Table 1, but were not controlled for in any of our own analyses.

To assess general temporal patterns in the data, observations were split into one 6-year (the first period, during which the fewest studies were reported) and four 4-year periods: 1995-2000, 2001-2004, 2005-2008, 2009-2012, and 2012-2016. Though both ORF1 and ORF2 regions of the norovirus genome were used throughout each period, there is a clear trend in the increased use of ORF2 compared to ORF1, particularly between the years of 2005 and 2016. During this time, relative proportions of ORF2-use increased, while the opposite occurred for ORF1 use. Additionally, there is an obvious trend in the increasing use of multiple regions, as well as the ORF1/ORF2 overlap region, for norovirus genotyping. The first reporting of the use of the overlap region occurred between the years of 2001 and 2004 but was not used exclusively until the 2009-2012 period.

b. GII.4 Variant Patterns

After reclassification of GII.4 variants, a total of 59 unique variants were reported among all 41 studies (Supplemental Figure). These variants were then classified into eight pandemic variant clusters, and 19 different non-pandemic clusters. Sub-variants of pandemic clusters were classified as a pandemic variant, and any untypeable variants were classified as non-pandemic variants.

Temporally, similar patterns of distribution of GII.4 pandemic variants were observed between outbreaks and sporadic cases (Figure 3a). Shifts in the prominence of subsequent pandemic variants were observed in the same chronological order as previously reported throughout the literature, demonstrating the 2-3 year pattern of GII.4 variant evolution and replacement. Each of the classified pandemic variants, except for Asia_2003, had its own period of dominance that lined up roughly with its assigned pandemic years. The lack of dominance of the Asia_2003 variant is likely explained by inconsistencies in classification schemes over the years. Whereas the other variant names were more commonly and consistently used, the Asia_2003 variant was more recently assigned as a main pandemic-causing variant. It is possible, therefore, that variants of this lineage may have been misclassified as other pandemic variants occurring around this time. There was a noticeable percentage of Den_Haag_2006b variants observed in the 1995-2000 period before this strain reaches pandemic dominance in the 2004-2008 period. This may be attributed to the presence of Den_Haag_2006b pre-epidemic variants that could have been circulating around this time before evolving into its pandemic version.

Proportions of variants vary between regions, though it can be observed that the Den_Haag_2006b, New_Orleans_2009, and Sydney_2012 generally tend to dominate the relative number of infections in each region (not including the Eastern Mediterranean Region, for which only one study was reported and included only one GII.4 variant). This may be due in part to the fact that the majority of studies included in this review were collected over the periods that these strains were most dominant; however, Figure 2a shows the relative dominance of these particular variants even within their own time periods compared to the degree of dominance of other variants, the exception being the US95_96 variant which had the fewest number of studies collected over its dominant period. Figure 2b therefore shows that the degree to which each of these variants caused pandemic outbreaks may vary between region. Reasons for this may be both the geographical proximity of regions, as well as heterogeneity within the host populations of each of these regions.

A similar trend was seen among age groups, in that the same three variants tend to dominate each group (except for the 15 years and older age group for which only two studies reported data). The most diversity among variants appears in the 5 years and under and mixed/all age categories, though these results are difficult to interpret as the 17 years and under category would also include those that are 5 years and under. Additionally, it is impossible to know the make-up of the mixed all category- whether it is evenly divided between all ages or more heavily skewed towards one particular age group.

Slight variations in relative pandemic variant proportions are observed between studies that reported sporadic cases and those that reported outbreaks. The US_95_96 and Den_Haag_2006b variants are the only two that appear to have caused more sporadic cases relative to outbreaks, whereas the Farmington_Hills_2002, Hunter_2004, and Sydney_2012 variants appear to have been involved in more outbreaks. These results indicate that pandemic-causing variants may not all have the same propensity to cause outbreaks, and may have varied transmission mechanisms for reaching pandemic prominence.

c. Linear Regression

The results of the single linear regression analyses indicate that both under-five mortality rate (D: β = 0.0016, R^2 = 0.0065; H': β=0.0017, R^2 = 0.0065) and secretor status (D: β = 0.04, R^2 = 0.0002; H': β=0.15, R^2 = 0.0009) are positively associated with GII.4 variant diversity, though neither of these estimates yield a statistically significant association between predictor and outcome (Table 2). Conversely, development index value ((D: β = -0.12, R^2 = 0.0018; H': β=-0.08, R^2 = 0.0003) is negatively associated with diversity among GII.4 variants increase. This association, however, is also not statistically significant. These trends are the same for both Simpson's and Shannon's diversity indices. Among the Tukey pairwise comparisons, only two were found to be statistically significant at the 0.05 alpha level (Table 4). Both of these comparisons were between different levels of the time period variable. The means of Simpson diversity index values were statistically significantly different between the 2009-2012 and the 2001-2004 periods, as well as between the 2013-2016 and 2009-2012 periods. All comparisons represent the absolute value of the difference between means of each level. The means for the levels of each categorical variables are given in Table 3. The model selection process for the multiple linear regression analyses yielded a final model that included age category and time period as predictor variables, for both Simpson and Shannon indices:

$$Y = \beta_0 + \beta_1(agecat1) + \beta_2(agecat2) + \beta_3(agecat3) + \beta_4(period1) + \beta_5(period2) + \beta_6(period3) + \beta_7(period4) + \beta_8(period5) + \varepsilon$$

Both models have an overall p value of <0.0001, an R^2 of 0.20, a root MSE of 0.19 and 0.30, respectively, and AICC values of -297.4 and -179.0, respectively. The p-values for the partial F tests for age category were 0.06 and 0.04 for Simpson's and Shannon's indices, respectively, indicating that the overall association between age category and GII.4 diversity may be slightly significant at the 0.05 alpha level. The p-values for the partial F tests for time period were highly significant at the 0.05 alpha level (<0.0001 for each index), indicating a strong association between time period and GII.4 diversity. Parameter estimates for each level of each variable are given in Table 5. The first time period (1995-2000) was used as the referent group for this variable, and the Mixed/All category was used as the referent group for the age category variable. Compared to the referant group, diversity of GII.4 variants for all later time periods is increased. Diversity of GII.4 variants, compared to the referant group, is decreased for the 17 years and under and 15 years and older age groups, and increased for the 5 and under age group. For both diversity indices, the only age category level that had a statistically significant parameter estimate was "15 and older". The only time period level that had a statistically significant parameter estimate was 2009-2012.

Discussion

Our study identified associations between time period and age categories and GII.4 variant diversity. As a result of model selection for our multiple linear regression, both were found to have significant associations with variant diversity. The p value for the overall F-test indicates that the proposed relationships are statistically reliable. These associations were consistent between both Simpson's and Shannon's diversity indices, giving them further explanatory power. The low R^2 square value indicates that this is not a good predictive model, but that is of little relevance to the purpose of this study as there is likely to be much more that contributes to variant diversity than was able to be accounted for in this study. The partial F-tests for age category and time period indicate that the latter has greater explanatory power than the former, though both are necessary to include in the model. The significance of age category is more questionable in this model, with p-values for Simpson and Shannon diversity indices bordering the 0.05 alpha level cutoff. This cutoff is arbitrary, however, and at least one of the age category parameter estimates was significant in relation to the mixed/all reference group. This, in combination with evidence from the literature [2, 33] that host immune function varies with age and may therefore affect susceptibility to different variants, influenced our decision to keep age category as an explanatory variable in the model. For example, older children and adults will have a certain degree of acquired immunity, making them more susceptible to newly emerging strains, whereas infants with less-mature immune systems will be more equally susceptible to all strains. The results of the Tukey pairwise comparisons mostly substantiate the results of the MLR, showing significant differences in means between various time periods. However, these comparisons failed to detect differences between age categories, further showing that this variable has less explanatory power than time period.

Overall, both descriptive and statistical analyses elucidated important trends in our data. First, it appears that newer strains of GII.4 are becoming increasingly prevalent-whether this can be explained by virulence factors, by trends of increasing globalization, or both. Secondly, heterogeneity in the relative dominance of GII.4 pandemic variants between regions, age groups, and sporadic cases vs. outbreaks indicates that the degree to which these variants are widespread may differ for each new variant depending upon different characteristics in the population. The ability for norovirus to evolve so rapidly and thereby diversify itself on such a large scale makes it difficult to control and prevent globally, as regions/populations may differ in susceptibility and may not be equally affected with each pandemic. This is further emphasized by the trend that GII.4 variant diversity increases in each time period except for the last (2012-2016), in which it decreased compared to the previous time period (Figure 4). This decrease, however, may be explained by the increasing dominance of GII.17 strains as well as other recombinant strains (not accounted for in this study), thus perhaps signaling an end to the GII.4 era.

Though numerous studies exist that attempt to characterize GII.4 variant diversity within different populations (many of which are primary studies that were included in this meta-analysis), this review is one of few that attempts to compile this widespread data into one general analysis that summarizes patterns of GII.4 variant evolution and diversity on the global scale. Of these studies, it is the first of its kind to reclassify previously reported variants under one classification scheme. Though previous studies have aggregated data based off of the reporting of commonly classified pandemic variants, they have not, to our knowledge, made any attempt to account for variants that may have been misclassified as the result of the lack of a unified classification and naming scheme over the last 2.5 decades. As a result, we created a comprehensive norovirus dataset by compiling genotype data collected over time and throughout the world. Additionally, this study is the first of its kind to quantify variant diversity, which allows not only for better visualization of trends in diversity but also for statistical analysis to be performed on these trends.

This review has its limitations. First, the articles used in this meta-analysis were not uniform in terms of study design, data collection method and presentation, reporting unit, study population characteristics, and genotyping method. Where possible (and necessary), these factors were controlled for in the analysis, however it is acknowledged that the nature of reviews on this scale are often made less generalizable to the overall population. Second, though best efforts were made to thoroughly account for discrepancies among GII.4 variant classification and nomenclature, this process lacks certain objectivity due to the fact that strain data were not presented uniformly and that there is no current widely-accepted standard for doing this. This introduces the potential for misclassification bias, particularly in terms of defining the overall diversity of GII.4 variants as well as the overall impact of pandemic variants. For example, certain variants (particularly those that occurred more recently, such as Den_Haag_2006b, New_Orleans_2009, and Sydney_2012) have been more widely acknowledged as pandemic-causing variants, whereas earlier variants (particularly Asia_2003) have been less agreed upon. This, in turn, makes classification of sub-variants into these larger variants more difficult and generally less doable, as less information is available in the literature. It is therefore difficult to assess whether the abundance of certain pandemic variants relative to others is a result of actual abundance or misclassification. Third, there is an extreme paucity of data from low-income countries. This presents an issue both because it makes our results less interpretable on an overall global scale, and because it is these countries that bear the highest burden of norovirus morbidity and mortality. Additionally, these countries tend to be clustered in certain regions, potentially skewing the geographical distribution of variants as well. Lastly, uncertainty in the calculations of diversity indices was not accounted for in this analysis.

Conclusions

Norovirus infection is widespread, causing more cases and outbreaks of AGE worldwide than any other non-bacterial agent. The majority of these outbreaks can be attributed to a single genotype: GII.4, which generally contributes over half of the global outbreaks in any given season [3, 34]. Our study showed that the distribution of this burden, caused by the numerous GII.4 variants, differs by time, region, age of those infected, and their presence in either outbreaks or sporadic cases. Our study further demonstrated that there is a strong positive relationship between time and GII.4 diversity, indicating that certain selective pressures are causing these variants to evolve and diversify at a steady rate. The increased abundance of these genetic variants may possibly be leading to the increasing prevalence of GII.4 variant recombinants. Additionally, our study showed a weak association between the quantity of GII.4 diversity and age group.

The results of our study have significant public and global health implications. Efforts to develop a norovirus vaccine are underway, but will require knowledge of mechanisms of norovirus evolution and dynamics, as well as data specific to the spatial and temporal distribution of norovirus genotypes. Additionally, information regarding which populations to target with a vaccine will be necessary for the planning of vaccine introduction. This review helps to inform both of these areas, summarizing the distribution of GII.4 variants using available data over the last two and a half decades, and providing insightful preliminary information regarding possible associations between various factors and the degree of norovirus diversity. The observation that genotype diversity has been steadily increasing over time implies that vaccines will need to be developed in a way that they can account for this increasing diversity or may need to be introduced on a cyclic basis, similar to the influenza vaccine. Should this be the case, understanding norovirus evolution will be essential for predicting patterns of novel strain emergence so that vaccines can be rapidly adjusted to target them. Additionally, the knowledge that different age groups may not be equally susceptible to different norovirus genotypes may help to inform who should be the primary beneficiary of any vaccine.

This review also highlighted a number of areas in which future research should be done. Firstly, like many reviews of its kind, it emphasized the need for data collection in less developed countries where resources to do such studies are often the poorest, but the need is greatest. Secondly, it may be worthwhile to direct efforts towards studying variation in norovirus infection among different age groups in more detail. Research directed specifically at this question may further help to inform vaccine development and other prevention strategies. Thirdly, recent studies have indicated that future pandemic norovirus outbreaks may be more increasingly caused by recombinant types and the emerging GII.17 strain, and less by GII.4 variants. Focusing research and data collection on these norovirus strains will be crucial for the continued surveillance of the virus. Lastly, this review has underscored the need to move forward in this field of research with a unified classification scheme and typing method, particularly as non-GII.4 strains gain predominance.

References

1. Ahmed, S.M., et al., *Global prevalence of norovirus in cases of gastroenteritis: a systematic review and meta-analysis.* Lancet Infect Dis, 2014. **14**(8): p. 725-730.

2. Lopman, B.A., et al., *The Vast and Varied Global Burden of Norovirus: Prospects for Prevention and Control.* PLoS Med, 2016. **13**(4): p. e1001999.

3. Siebenga, J.J., et al., *Norovirus illness is a global problem: emergence and spread of norovirus GII.4 variants, 2001-2007.* J Infect Dis, 2009. **200**(5): p. 802-12.

4. Centers for Disease Control and Prevention. *Burden of Norovirus Illness in the U.S.* . 2018 July 16, 2018 [cited 2018 May, 05]; Available from: https://www.cdc.gov/norovirus/trends-outbreaks/burden-US.html.

5. Teunis, P.F., et al., *Norwalk virus: how infectious is it?* J Med Virol, 2008. **80**(8): p. 1468-76.

6. Chan, M.C., et al., *Fecal viral load and norovirus-associated gastroenteritis.* Emerg Infect Dis, 2006. **12**(8): p. 1278-80.

7. Tu, E.T., et al., *Epidemics of gastroenteritis during 2006 were associated with the spread of norovirus GII.4 variants 2006a and 2006b.* Clin Infect Dis, 2008. **46**(3): p. 413-20.

8. Robilotti, E., S. Deresinski, and B.A. Pinsky, *Norovirus.* Clin Microbiol Rev, 2015. **28**(1): p. 134-64.

9. Kapikian, A.Z., et al., *Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis.* J Virol, 1972. **10**(5): p. 1075-81.

10. Bull, R.A., et al., *Norovirus recombination in ORF1/ORF2 overlap.* Emerg Infect Dis, 2005. **11**(7): p. 1079-85.

11. Ludwig-Begall, L.F., A. Mauroy, and E. Thiry, *Norovirus recombinants: recurrent in the field, recalcitrant in the lab - a scoping review of recombination and recombinant types of noroviruses.* J Gen Virol, 2018.

12. White, P.A., *Evolution of norovirus.* Clin Microbiol Infect, 2014. **20**(8): p. 741-5.

13. Parra, G.I., et al., *Static and Evolving Norovirus Genotypes: Implications for Epidemiology and Immunity.* PLoS Pathog, 2017. **13**(1): p. e1006136.

14. Noel, J.S., et al., *Identification of a distinct common strain of "Norwalk-like viruses" having a global distribution.* J Infect Dis, 1999. **179**(6): p. 1334-44.

15. Duizer, E., et al., *Laboratory efforts to cultivate noroviruses.* J Gen Virol, 2004. **85**(Pt 1): p. 79-87.

16. Kroneman, A., et al., *Proposal for a unified norovirus nomenclature and genotyping.* Arch Virol, 2013. **158**(10): p. 2059-68.

17. Vinje, J., et al., *Genetic polymorphism across regions of the three open reading frames of "Norwalk-like viruses".* Arch Virol, 2000. **145**(2): p. 223-41.

18. Zheng, D.P., et al., *Norovirus classification and proposed strain nomenclature.* Virology, 2006. **346**(2): p. 312-23.

19. Zheng, D.P., et al., *Molecular epidemiology of genogroup II-genotype 4 noroviruses in the United States between 1994 and 2006.* J Clin Microbiol, 2010. **48**(1): p. 168-77.

20. Bok, K., et al., *Evolutionary dynamics of GII.4 noroviruses over a 34-year period.* J Virol, 2009. **83**(22): p. 11890-901.

21. Pybus, O.G. and A. Rambaut, *Evolutionary analysis of the dynamics of viral infectious disease.* Nat Rev Genet, 2009. **10**(8): p. 540-50.

22. Shanker, S., et al., *Structural analysis of histo-blood group antigen binding specificity in a norovirus GII.4 epidemic variant: implications for epochal evolution.* J Virol, 2011. **85**(17): p. 8635-45.

23. Bull, R.A., et al., *Rapid evolution of pandemic noroviruses of the GII.4 lineage.* PLoS Pathog, 2010. **6**(3): p. e1000831.

24. Giammanco, G.M., et al., *Norovirus GII.17 as Major Epidemic Strain in Italy, Winter 2015-16.* Emerg Infect Dis, 2017. **23**(7): p. 1206-1208.

25. Sakon, N., et al., *Foodborne Outbreaks Caused by Human Norovirus GII.P17-GII.17-Contaminated Nori, Japan, 2017.* Emerg Infect Dis, 2018. **24**(5): p. 920-923.

26. Chan, M.C., et al., *Rapid emergence and predominance of a broadly recognizing and fast-evolving norovirus GII.17 variant in late 2014.* Nat Commun, 2015. **6**: p. 10061.

27. Pan, L., et al., *The novel norovirus genotype GII.17 is the predominant strain in diarrheal patients in Shanghai, China.* Gut Pathogens, 2016. **8**(1): p. 49.

28. Parrino, T.A., et al., *Clinical immunity in acute gastroenteritis caused by Norwalk agent.* N Engl J Med, 1977. **297**(2): p. 86-9.

29. Currier, R.L., et al., *Innate Susceptibility to Norovirus Infections Influenced by FUT2 Genotype in a United States Pediatric Population.* Clin Infect Dis, 2015. **60**(11): p. 1631-8.

30. Morris, E.K., et al., *Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories.* Ecol Evol, 2014. **4**(18): p. 3514-24.

31. Maryland Sea Grant. *How To Calculate Biodiversity?* 2018 July 17, 2018 [cited 2018 July 17]; Available from: http://ww2.mdsg.umd.edu/interactive_lessons/biofilm/diverse.htm.

32. Jahan, S., *Human Development Report*. 2016, United Nations Development Programme: New York, NY.

33. Romero, C., et al., *Incidence of Norovirus-Associated Diarrhea and Vomiting Disease Among Children and Adults in a Community Cohort in the Peruvian Amazon Basin.* Clin Infect Dis, 2017. **65**(5): p. 833-839.

34. Cannon, J.L., et al., *Genetic and Epidemiologic Trends of Norovirus Outbreaks in the United States from 2013 to 2016 Demonstrated Emergence of Novel GII.4 Recombinant Viruses.* J Clin Microbiol, 2017. **55**(7): p. 2208-2221.

Tables

Table 1
Distribution of extracted data used in meta-analysis

	Number of studies	Number of Observations	Average Study Time Span (years)	Number of different GII.4 types	Number of Pandemic Strains	Number of non-Pandemic Strains
N	*45*[f]	*261*	*5.26*	*59*	*8*	*19*
WHO region						
African	2	16	6.71	12	7	4
American	7	47	5.94	19	8	5
European	11	59	4.69	35	8	9
Eastern Mediterranean	1	5	3.92	1	1	0
South-east Asian	2	12	5.50	13	6	2
Western Pacific	22	122	5.22	28	8	10
Time Period[a]						
1995-2000	5	18	Not Assigned	15	6	4
2001-2004	19	51	Not Assigned	28	8	11
2005-2008	29	79	Not Assigned	33	7	10
2009-2012	27	91	Not Assigned	29	8	6
2013-2016	13	22	Not Assigned	9	4	3
UNDP Development Index[b]						
Very High Human Development	37	202	4.98	48	8	17
High Human Development	4	30	6.71	10	8	1
Medium Human Development	3	18	5.42	11	7	2
Low Human Development	1	11	9.08	9	5	3
Under-Five Mortality Rate[c]						
First quartile (0-2)	11	69	6.23	29	8	7
Second quartile (3-4)	11	63	4.75	28	7	9
Third quartile (5-12)	7	36	4.29	16	7	5
Fourth quartile (13-71)	7	56	7.06	22	8	5
Missing[d]	9	37	4.04	11	5	3
Age						
5 years and under	11	78	6.61	28	8	7
17 years and under	10	43	3.76	21	6	5
15 years and older	2	10	3.83	2	1	1
Mixed/All	22	130	5.39	40	8	11
Setting[e]						
Inpatient	18	99	5.13	52	8	6
Outpatient	11	59	4.87	35	8	5
Community	15	84	5.89	44	8	8
Country-wide Surveillance	9	47	4.52	33	8	8
Hospitals/Nursing Homes	2	9	3.96	10	4	1
Childcare Centers	1	13	11.92	2	1	1
Not Specified	57	51	8.3	12	7	3
Reporting Unit						
Single Case	27	154	5.17	44	8	10
Outbreak	18	107	5.39	35	8	12

[a] Assigned in this study per observation, therefore average study time span could not be assigned

[b] United Nations Development
Programme

[c] per 1,000 live births

[d] No data provided for China and Taiwan

[e] Studies that include, but are not limited to that setting

[f] N studies=45 to account for separate study populations within articles (N articles=42)

Table 2
Results of simple linear regression of Simpson Shannon Diversity Indices on Under 5 Mortality Rate, Human Development Index Value, and Secretor Status.

	Parameter Estimate (95% Confidence Interval)			
	Simpson D.I.		Shannon D.I.	
Under 5 mortality rate	0.0016	(-0.0017, 0.0048)	0.0017	(-0.0035, 0.0069)
Development index value	-0.12	(-0.54, 0.31)	-0.08	(-0.75, 0.59)
Secretor status	0.04	(-0.46, 0.55)	0.15	(-0.66, 0.95)

*Indicates statistical significance at alpha level 0.05

Table 3
Mean values for Simpson and Shannon diversity indices by different levels of each categorical variable (under 5 mortality rate quartiles, HDI categories, WHO region, age category, and time period)

	N	Mean (Simpson)	Mean (Shannon)
Under-Five Mortality Rate			
First quartile (0-2)	46	0.24	0.41
Second quartile (3-4)	56	0.27	0.46
Third quartile (5-12)	10	0.20	0.36
Fourth quartile (13-71)	33	0.23	0.37
Missing	21	0.14	0.21
UNDP Development Index			
Very High Human Development	145	0.23	0.39
High Human Development	11	0.26	0.40
Medium Human Development	10	0.28	0.44
Low Human Development	0	ND	ND
Age			
5 years and under	39	0.26	0.43
17 years and under	36	0.23	0.38
15 years and older	3	0	0
Mixed/All			
WHO region			
African	5	0.27	0.42
American	22	0.25	0.41
European	51	0.23	0.39
Eastern Mediterranean	3	0.00	0.00
South-east Asian	5	0.40	0.64
Western Pacific	80	0.23	0.38
Time Period			
1995-2000	2	0.03	0.08
2001-2004	27	0.13	0.21
2005-2008	55	0.23	0.39
2009-2012	65	0.32	0.53
2013-2016	17	0.10	0.20

ND = Not able to be determined

Table 4
Tukey pairwise comparisons of the difference of means for the Simpson's diversity index (Shannon's diversity index means were excluded, but showed the same statistically significant relationships) between categories for a) Under 5 mortality quartiles b) HDI categories c) WHO Region d) Age category and e) Time period

Under-Five Mortality Rate Quartiles

Simpson	1st	2nd	3rd	4th
1st	.	0.027 (-0.088, 0.142)	0.043 (-0.159, 0.244)	0.012 (-0.120, 0.144)
2nd		.	0.070 (-0.129, 0.268)	0.039 (-0.088, 0.166)
3rd			.	0.030 (-0.178, 0.239)
4th				.

UNDP Development Index

Simpson	Very High	High	Medium	Low
Very High	.	0.032 (-0.125, 0.189)	0.051 (-0.113, 0.215)	ND
High		.	0.019 (-0.200, 0.238)	ND
Medium			.	ND
Low				.

Age

Simpson	5 years and under	17 years and under	15 years and older	Mixed/All
5 years and under	.	0.031 (-0.100, 0.157)	0.257 (-0.070, 0.584)	0.024 (-0.081, 0.129)
17 years and under		.	0.227 (-0.102, 0.555)	0.006 (-0.102, 0.114)
15 years and older			.	0.233 (-0.088, 0.553)
Mixed/All				.

WHO Region

Simpson	African	American	European	Eastern Mediterranean	Sout-east Asian	Western Pacific
African	.	0.017 (-0.283, 0.317)	0.038 (-0.245, 0.322)	0.269 (-0.173, 0.711)	0.129 (-0.254, 0.511)	0.044 (-0.235, 0.323)
American		.	0.021 (-0.133, 0.175)	0.252 (-0.120, 0.624)	0.146 (-0.154, 0.446)	0.027 (-0.119, 0.172)
European			.	0.231 (-0.129, 0.590)	0.167 (-0.117, 0.450)	0.006 (-0.103, 0.114)
Eastern Mediterranean				.	0.400 (-0.044, 0.840)	0.225 (-0.131, 0.581)
South-east Asian					.	0.173 (-0.106, 0.451)
Western Pacific						.

Time Period

Simpson	1996-2000	2001-2004	2005-2008	2009-2012	2013-2016

1996-2000	.	0.140 (-0.487, 0.766)	0.312 (-0.303, 0.928)	0.450 (-0.164, 1.064)	0.123 (-0.516, 0.762)
2001-2004		.	0.173 (-0.028, 0.374)	0.311 (0.115, 0.507)*	0.016 (-0.248, 0.281)
2005-2008			.	0.138 (-0.019, 0.295)	0.189 (-0.048, 0.426)
2009-2012				.	0.327 (0.094, 0.570)*
2013-2016					.

ND = Not able to be determined

Table 5
Predictors of Simpson and Shannon diversity indices

	Simpson		Shannon	
Parameter	**B**	**95% CI**	**B**	**95% CI**
Age				
Mixed/All (referent)	0	.	0	.
5 years and under	0.013	(-0.064, 0.089)	0.018	(-0.102, 0.138)
17 years and under	-0.049	(-.126, 0.029)	-0.084	(-0.206, 0.038)
15 years and older*	-0.277	(-0.505, -0.050)	-0.463	(-0.819, -0.107)
Time Period				
1995-2000 (referent)	0	.	0	.
2001-2004	0.112	(-0.177, 0.401)	0.165	(-0.288, 0.617)
2005-2008	0.229	(-0.053, 0.511)	0.361	(-0.080, 0.803)
2009-2012*	0.309	(0.028, 0.590)	0.496	(0.055, 0.936)
2013-2016	0.085	(-0.207, 0.377)	0.146	(-0.311, 0.603)

Figures

Figure 1
Flow diagram of systematic review process and articles included in analysis

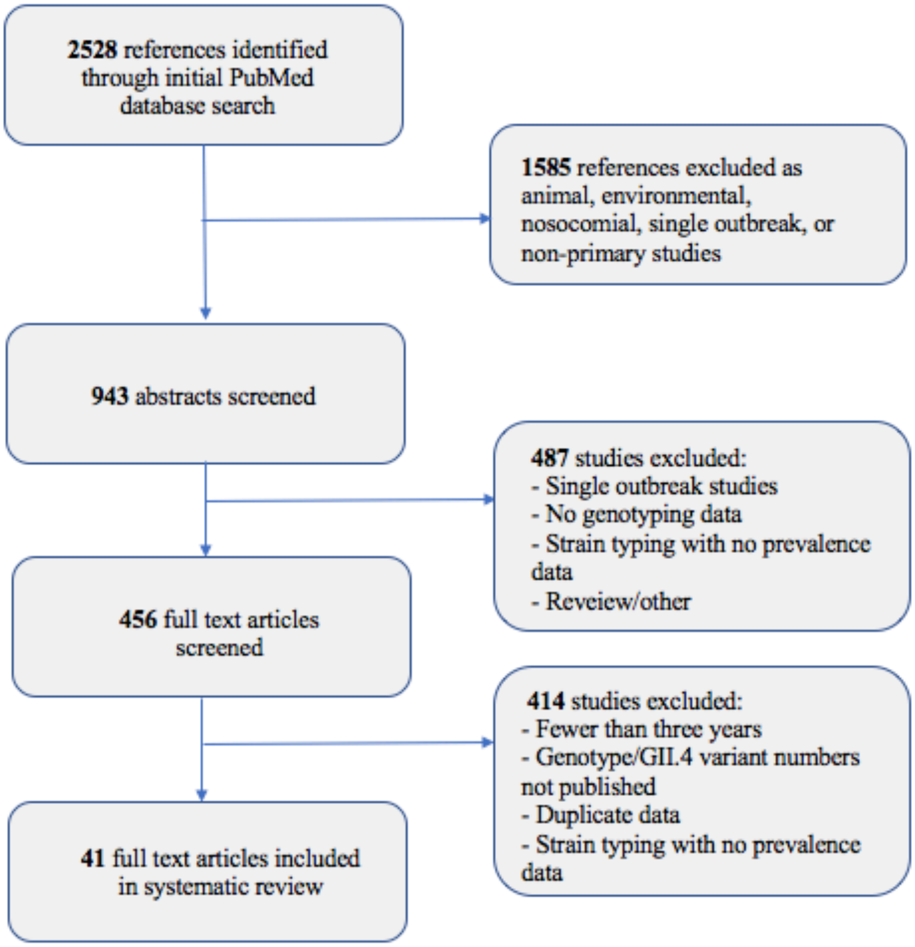
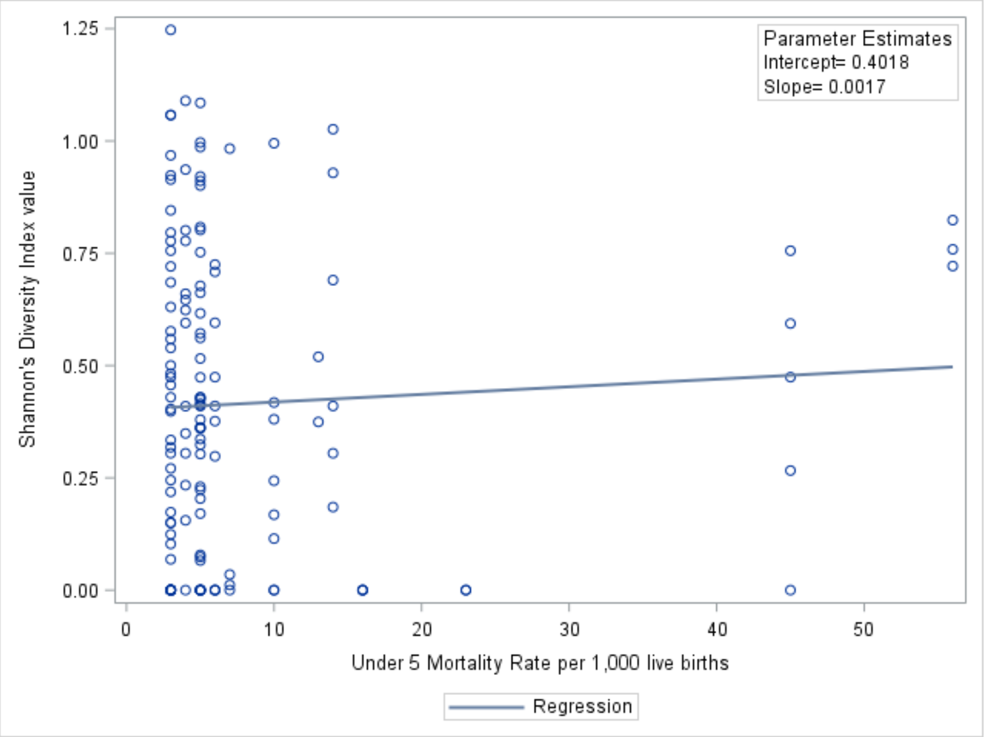

2528 references identified through initial PubMed database search
1585 references excluded as animal, environmental, nosocomial, single outbreak, or non-primary studies
943 abstracts screened
487 studies excluded:
- Single outbreak studies
- No genotyping data
- Strain typing with no prevalence data
- Reveiew/other
456 full text articles screened
414 studies excluded:
- Fewer than three years
- Genotype/GII.4 variant numbers not published
- Duplicate data
- Strain typing with no prevalence data
41 full text articles included in systematic review

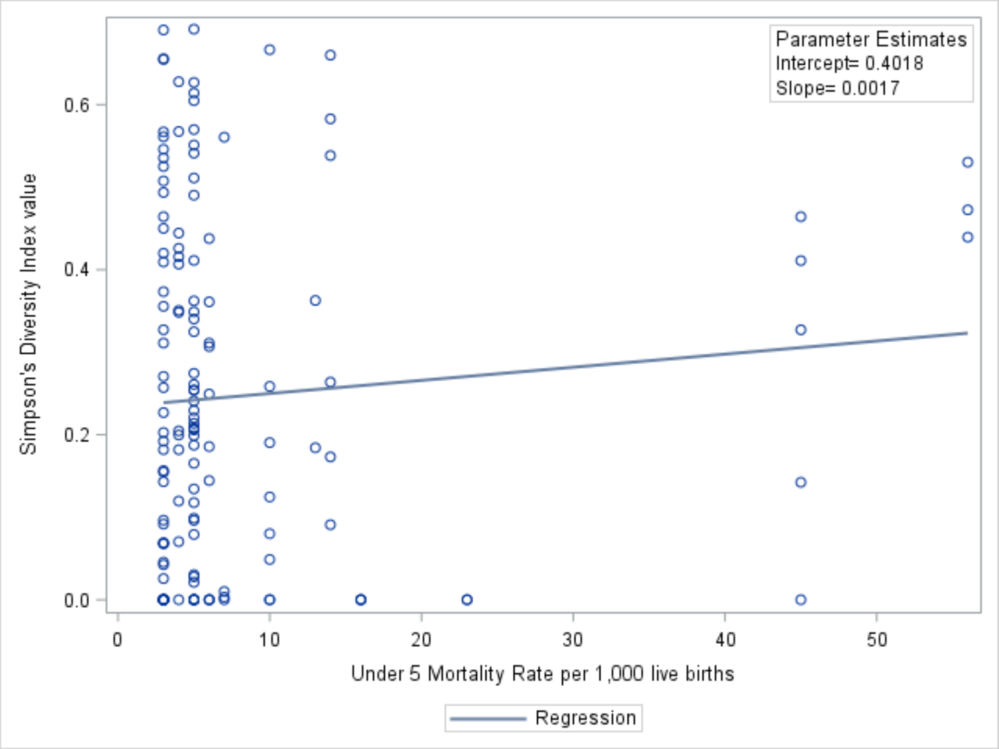
Figure 2

a) Scatter plots of I.a Under 5 mortality rate x Simpson I.b Under 5 mortality rate x Shannon II.a HDI value x Simpson II.b HDI value x Shannon IIIa. Secretor status x Simpson IIIb. Secretor status x Shannon

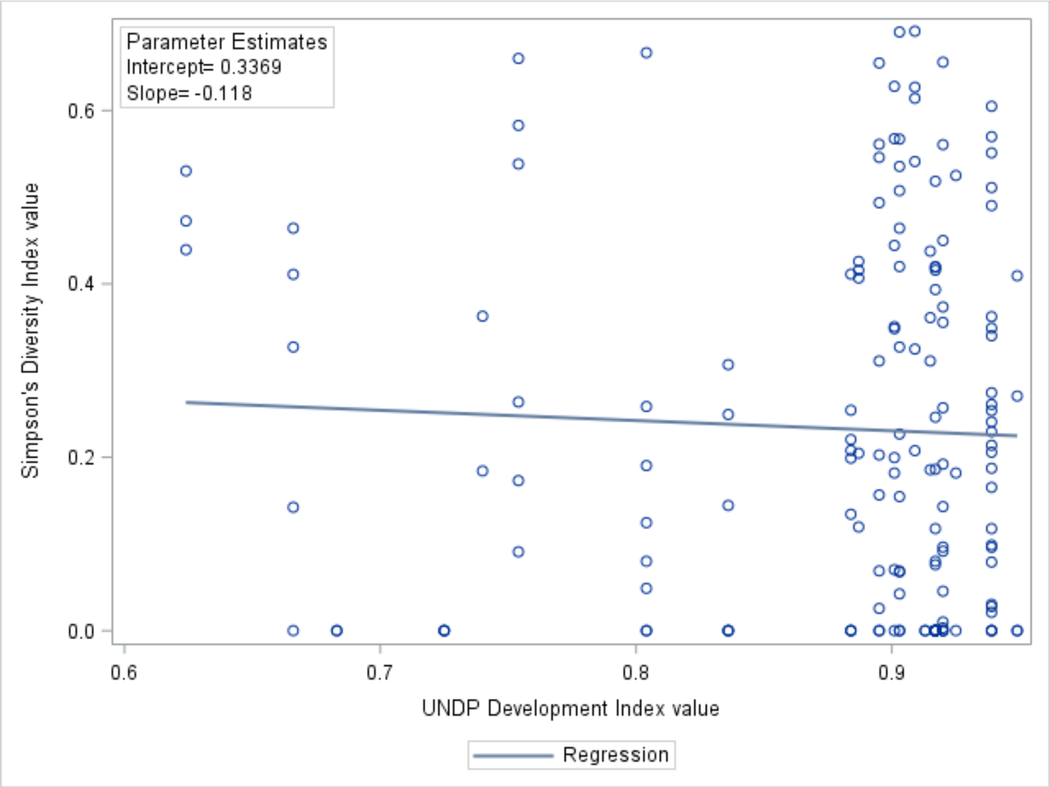
I.a


Parameter Estimates
Intercept= 0.4018
Slope= 0.0017
Shannon's Diversity Index value
1.25
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
50
Under 5 Mortality Rate per 1,000 live births
Regression

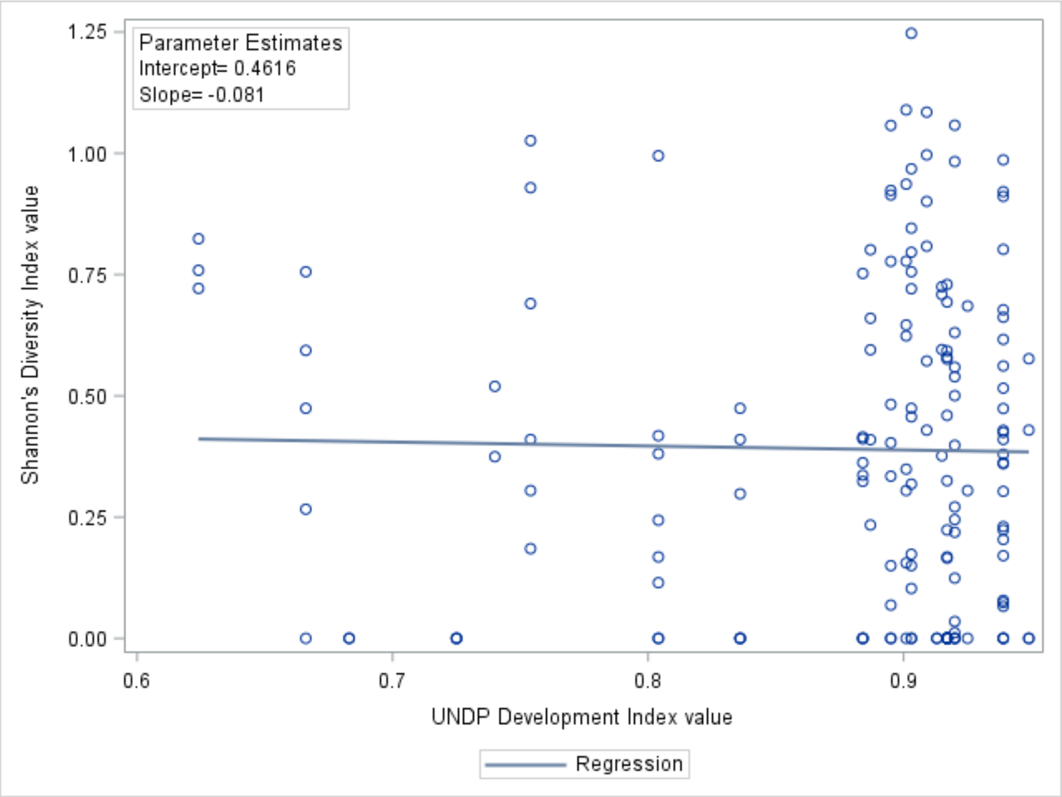
I.b


Parameter Estimates
Intercept= 0.4018
Slope= 0.0017
Simpson's Diversity Index value
0.6
0.4
0.2
0.0
0
10
20
30
40
50
Under 5 Mortality Rate per 1,000 live births
Regression

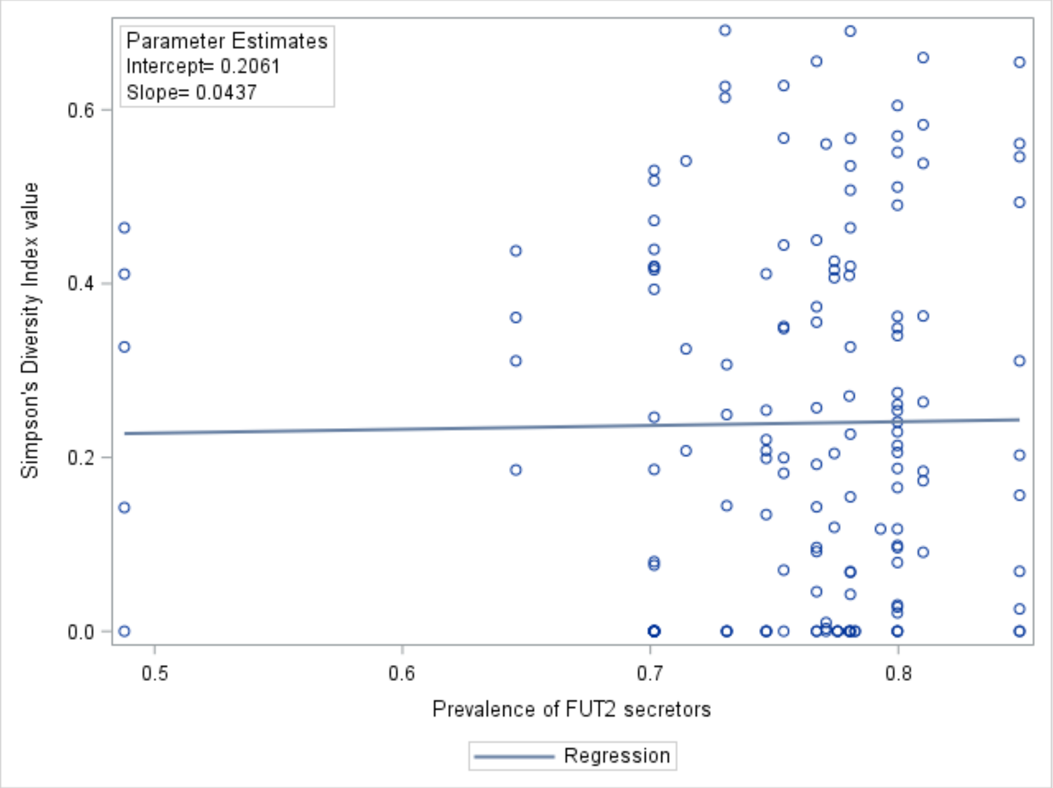
II.a


Parameter Estimates
Intercept= 0.3369
Slope= -0.118
Simpson's Diversity Index value
0.0
0.2
0.4
0.6
0.6
0.7
0.8
0.9
UNDP Development Index value
Regression

II.b


Parameter Estimates
Intercept= 0.4616
Slope= -0.081
Shannon's Diversity Index value
0.00
0.25
0.50
0.75
1.00
1.25
0.6
0.7
0.8
0.9
UNDP Development Index value
Regression

III.a


Parameter Estimates
Intercept= 0.2061
Slope= 0.0437
Simpson's Diversity Index value
0.0
0.2
0.4
0.6
0.5
0.6
0.7
0.8
Prevalence of FUT2 secretors
Regression

III.b

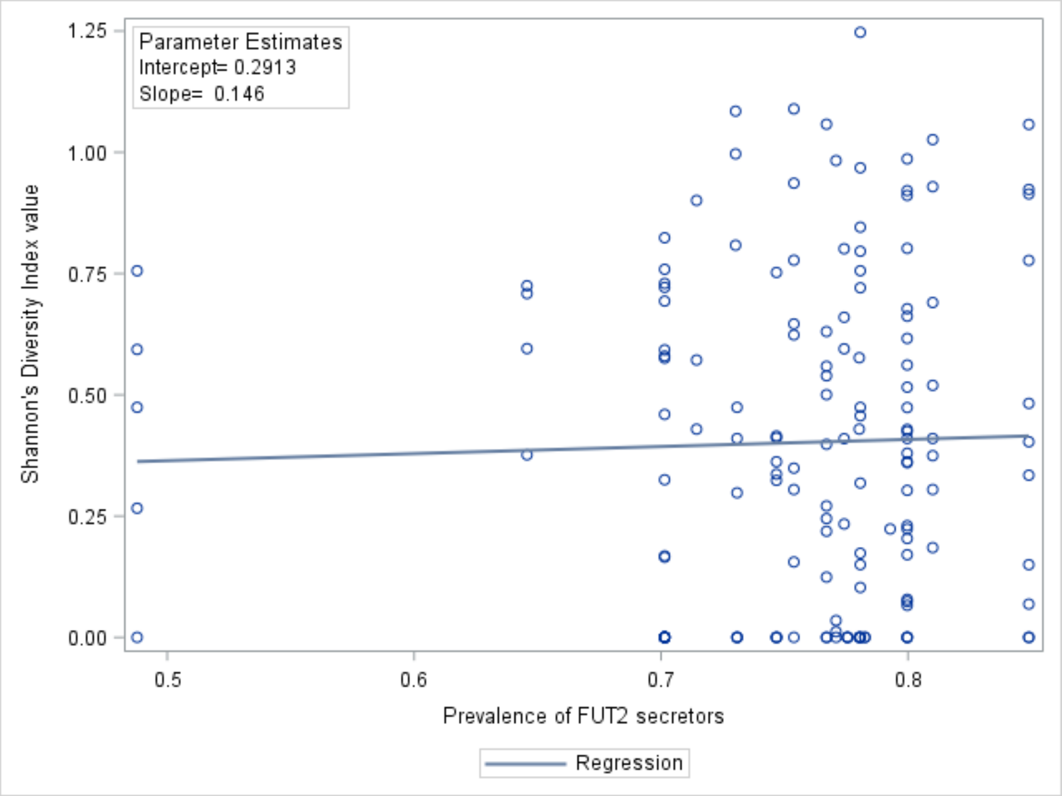
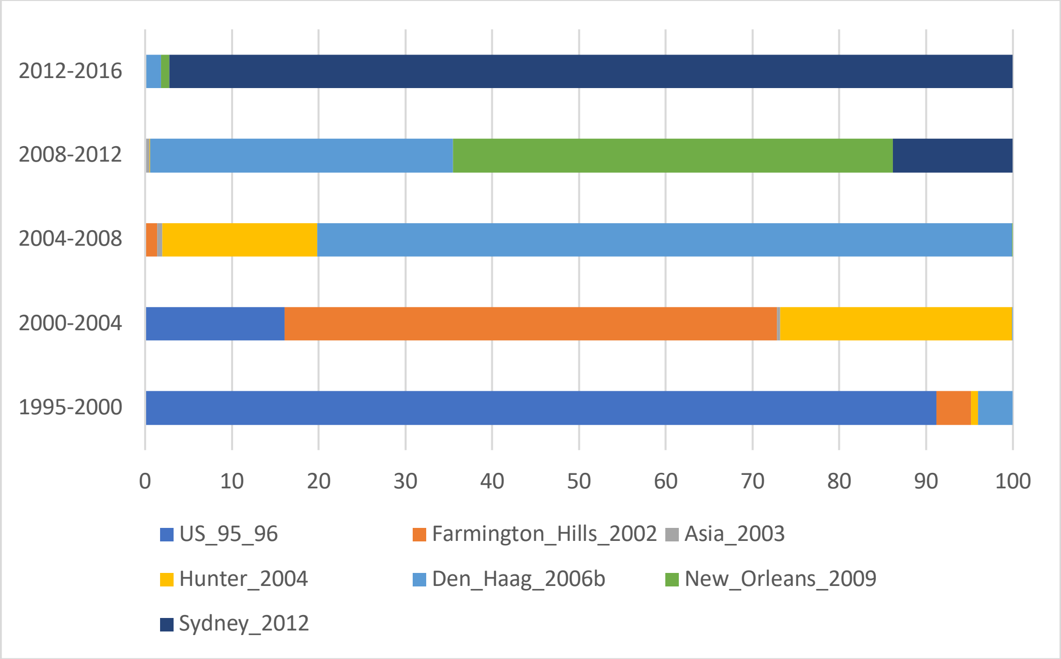

Parameter Estimates
Intercept= 0.2913
Slope= 0.146
Shannon's Diversity Index value
0.00
0.25
0.50
0.75
1.00
1.25
0.5
0.6
0.7
0.8
Prevalence of FUT2 secretors
Regression

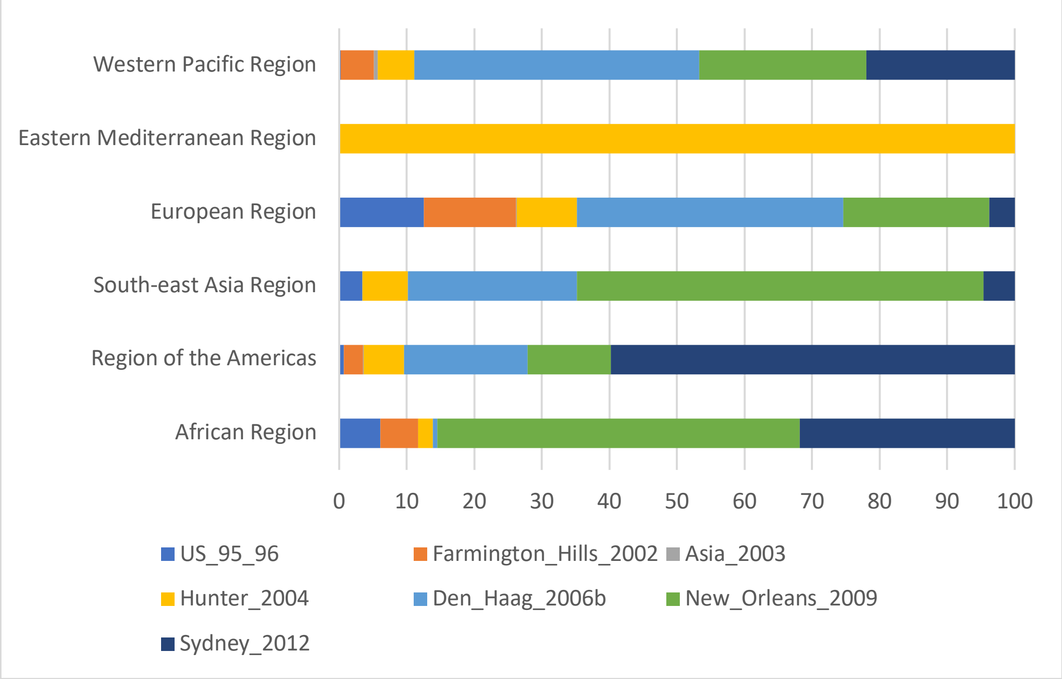
Figure 3

GII.4 variant distribution among studies by a) Time Period b) WHO Regions c) Age Group d) Reporting Unit

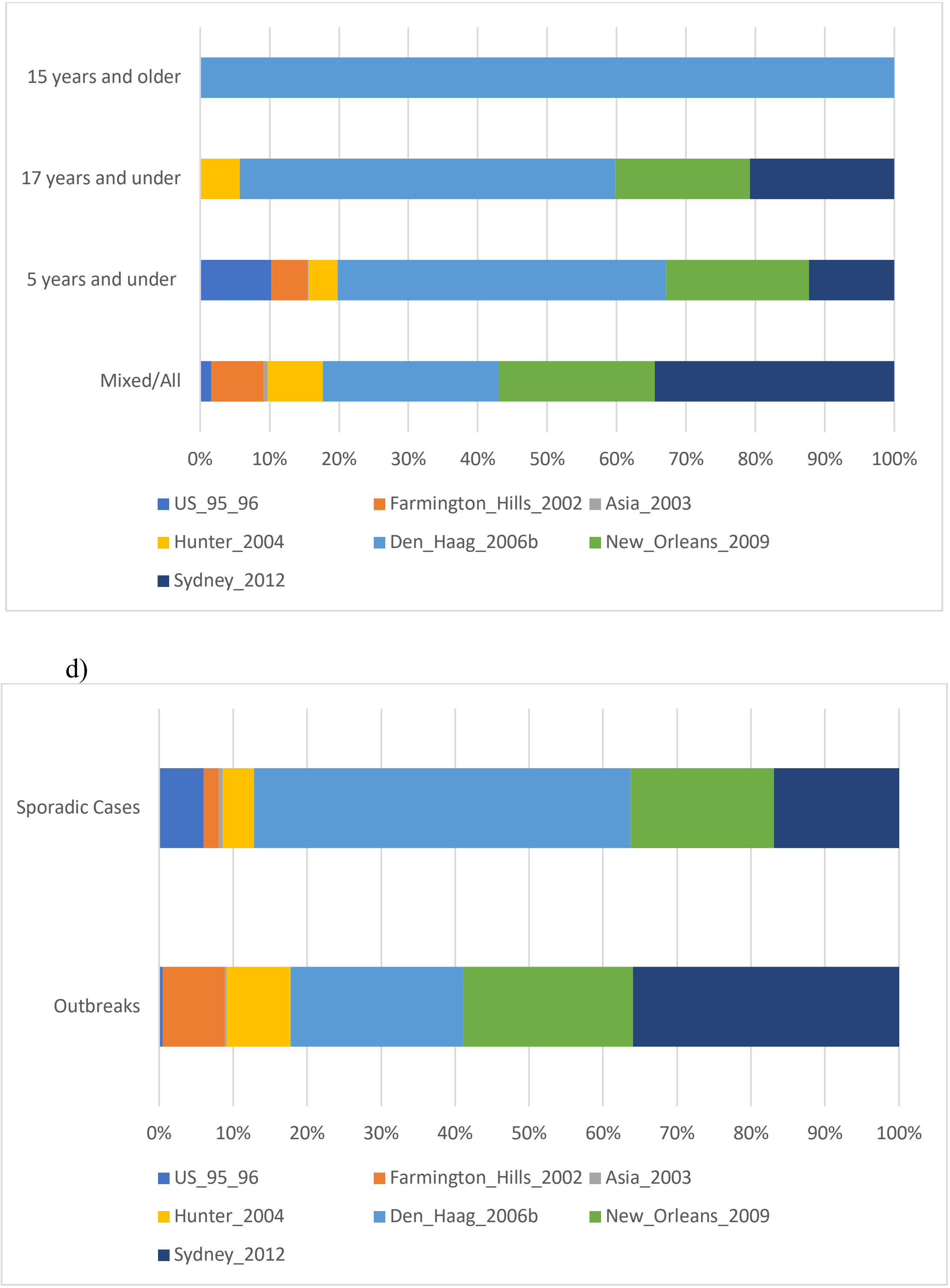
a)



b)



c)

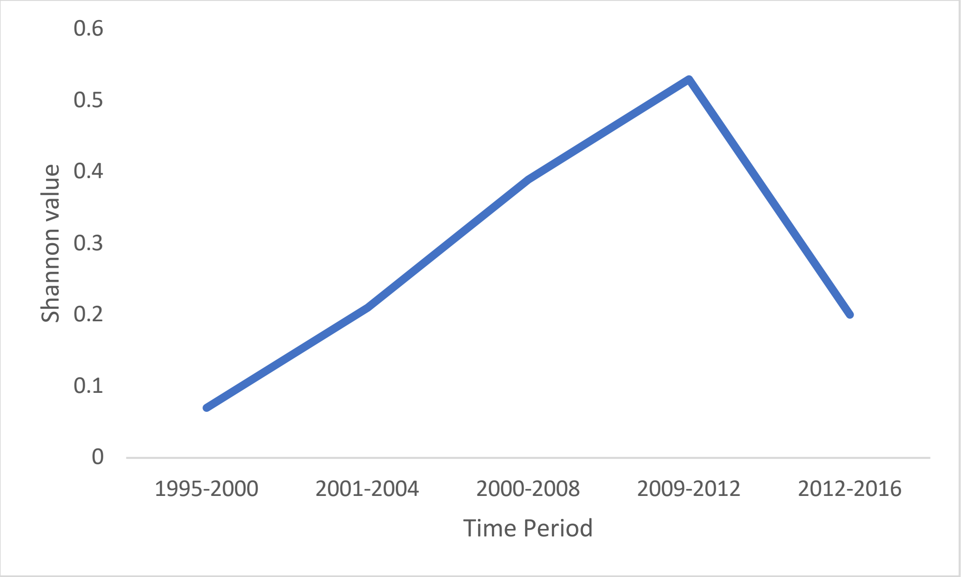

15 years and older
17 years and under
5 years and under
Mixed/All
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
US_95_96
Farmington_Hills_2002
Asia_2003
Hunter_2004
Den_Haag_2006b
New_Orleans_2009
Sydney_2012
Sporadic Cases
Outbreaks
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
US_95_96
Farmington_Hills_2002
Asia_2003
Hunter_2004
Den_Haag_2006b
New_Orleans_2009
Sydney_2012

d)

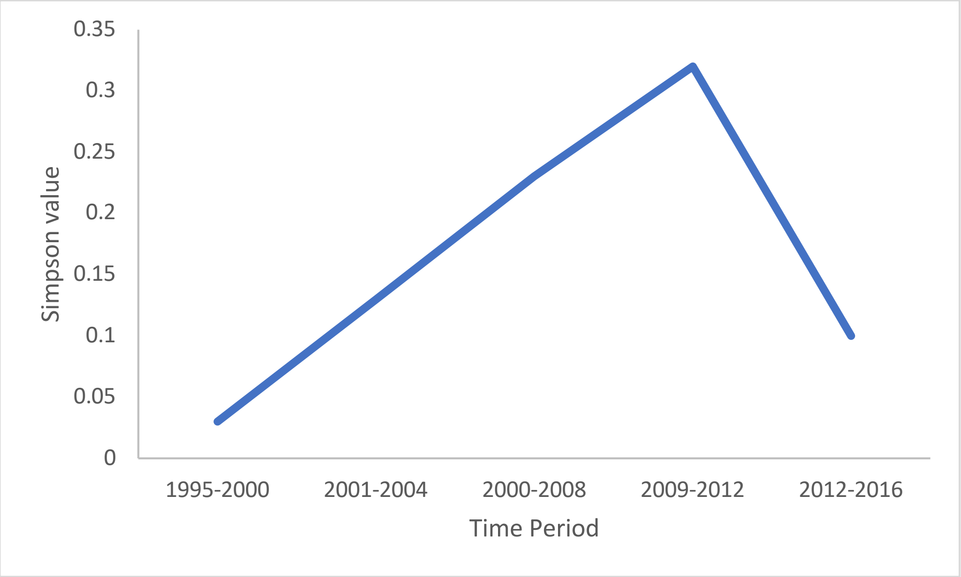
Figure 4

Trend of GII.4 variant diversity over time for both a) Shannon's and b) Simpson's diversity indices

a)



b)



Supplemental Table

Classification of GII.4 variants used in this study

Strain Name	**Pandemic Variant**
Dresden174_1997	US95_96
408/97003012/1996/FL	US95_96
Grimsby_1996	US95_96
Houston_2002	US95_96
OCS960352_1996	US95_96
OCS960352_1996 variant 1	US95_96
OCS960352_1996 variant 2	US95_96
US_95_96	US95_96
1996 variant	US95_96
Farmington_Hills_2002	Farmington_Hills_2002
2002 variant	Farmington_Hills_2002
Asia_2003	Asia_2003
04/05/JP/CHN	Asia_2003
Hunter_2004	Hunter_2004
Hunter_2004 sub-cluster A	Hunter_2004
2004 variant	Hunter_2004
Hokkaido_2004	Yerseke_2006a
Rhy1440_2005	Yerseke_2006a
Terneuzen_2006a	Yerseke_2006a
2006a variant	Yerseke_2006a
2007 variant	Yerseke_2006a
Yerseke_2006a	Yerseke_2006a
Den_Haag_2006b	Den_Haag_2006b
Den_Haag_2006b (Sublineage "O")	Den_Haag_2006b
Den_aag_2006b (Sublineage "Y")	Den_Haag_2006b
Hokkaido_2008	Den_Haag_2006b
Lincolnhouse_2006b	Den_Haag_2006b
Minerva_2006b	Den_Haag_2006b
Nijmegen115_2006b	Den_Haag_2006b
Shellharbour_2006b	Den_Haag_2006b
2006b variant	Den_Haag_2006b
Wuhan_2007	Den_Haag_2006b
New_Orleans_2009	New_Orleans 2009
Lincolnhouse_2006b variant 1	New_Orleans 2009
Lincolnhouse_2006b variant 2	New_Orleans 2009
New_Orleans_2009 sub-cluster A	New_Orleans 2009
New_Orleans_2009 sub-cluster B	New_Orleans 2009

New_Orleans_2009_unassigned subcluster	New_Orleans 2009
New_Orleans_2009 variant 1	New_Orleans 2009
2010 variant	New_Orleans 2009
Sydney_2012	Sydney_2012
Apeldoorn_2007	Non Pandemic Cluster
Auckland_2010	Non Pandemic Cluster
Bristol_1996	Non Pandemic Cluster
Brynhaven_2003	Non Pandemic Cluster
Camberwell_1994	Non Pandemic Cluster
Chiba_2005	Non Pandemic Cluster
Ehime_05_30	Non Pandemic Cluster
EmmenE006_2002	Non Pandemic Cluster
Kaiso_2003	Non Pandemic Cluster
OC07138_2007	Non Pandemic Cluster
Orange_2008	Non Pandemic Cluster
Osaka_2007	Non Pandemic Cluster
Portsmouth_2004	Non Pandemic Cluster
pre 1996 variant	Non Pandemic Cluster
04/05/AU/NL	Non Pandemic Cluster
2000 variant	Non Pandemic Cluster
2001 variant	Non Pandemic Cluster
2008a variant	Non Pandemic Cluster
414055_2004	Non Pandemic Cluster
Untypeable GII.4	Untypeable GII.4

CHAPTER 3

Summary, Public Health Implications & Possible Future Directions

Norovirus infection is widespread, causing more cases and outbreaks of AGE worldwide than any other non-bacterial agent. This comes at a huge cost both to individuals and health systems, incurring unnecessary morbidity and mortality- particularly in low-income countries- as well as large financial burdens on hospitals which require large resource expenditures to prevent and control outbreaks both within the community and hospital setting. Norovirus does not confer life-long immunity- and while host factors play a part in determining individual susceptibility, the transmission of norovirus is not inhibited by permanently immune individuals in the population. This, in large part, is due to the tendency for norovirus to evolve so rapidly and efficiently, effectively evading host immunity. Moreover, the evolution of norovirus and its increasing genetic diversity is thought to be driven by heterogeneity in the human immune system. Since the identification of the prototypical norovirus genotype, GI.1, in the early 1990s, over 40 different genotypes of norovirus have been identified. The GII.4 genotype has been the most common of these genotypes, itself diversifying into numerous variants and subvariants capable of causing pandemic outbreaks on a 2 to 3-year basis. This genetic diversity within norovirus has made it extremely difficult to target with any single control or prevention measure; strategies to prevent norovirus on a large scale remain entirely out of reach and control measures remain costly and on an individual-level basis.

Understanding the specific nature of norovirus evolution and diversity, especially in regard to its pandemic-causing strains, is crucial for the development of strategies to prevent norovirus on a global scale. The widespread benefits of such prevention methods have been demonstrated by the global reduction in rotavirus cases since vaccine introduction in 2006. Efforts to develop a norovirus vaccine are underway, but will require knowledge of mechanisms of norovirus evolution and dynamics, as well as data specific to the spatial and temporal distribution of norovirus genotypes. Additionally, information regarding which populations to target with a vaccine will be necessary for the planning of vaccine introduction. This review helps to inform both of these areas, summarizing the distribution of GII.4 variants using available data over the last two and a half decades, and providing insightful preliminary information regarding possible associations between various factors and the degree of norovirus diversity. Firstly, the observation that genotype diversity has been steadily increasing over time implies that vaccine development would likely need to be a dynamic, evolving process in order to account for the increasing diversity of norovirus genotypes. This might involve introducing new vaccines on a regular basis, similar to what is done for influenza.

will need to be developed in a way that they can account for this increasing diversity or may need to be introduced on a cyclic basis, similar to the influenza vaccine. Should this be the case, understanding norovirus evolution will be essential for predicting patterns of novel strain emergence so that vaccines can be rapidly adjusted to target them. Secondly, the knowledge that different age groups may not be equally susceptible to different norovirus genotypes may help to inform who should be the primary beneficiary of any vaccine. Though those that appear to host a greater diversity of norovirus genotypes are apt to be more susceptible to infection, they may be more difficult to target with a single vaccine that will likely not target all genotypes in current circulation.

This review also highlighted a number of areas in which future research should be done. Like many reviews of its kind, it emphasized the need for data collection in less developed countries where resources to do such studies are often the poorest, but the need is greatest. The paucity of data from these countries will make prevention efforts much more difficult to specialize to these areas where they are often most difficult to implement for a variety of other reasons. Based on the results of this study, it may also be worthwhile to direct efforts towards studying variation in norovirus infection among different age groups in more detail. Our study showed weak associations between these factors but was not able to describe them in much detail due to the nature of data collection on the age variable, which is often limited in systematic reviews and meta-analyses. Research directed specifically at this question may further help to inform vaccine development and other prevention strategies. Lastly, recent studies have indicated that future pandemic norovirus outbreaks may be more increasingly caused by recombinant types and the emerging GII.17 strain, and less by GII.4 variants. These findings may have been substantiated in this study by the drop off in GII.4 diversity in the last four or so years. Focusing research and data collection on these norovirus strains will be crucial for the continued surveillance of the virus. This review has also underscored the need to move forward in this field of research with a unified classification scheme and typing method, particularly as non-GII.4 strains gain predominance.