Analysing *Escherichia coli* O157 outbreaks in Scotland, Canada and the United States of America

Kate Grayston Snedeker

A thesis submitted for the Doctor of Philosophy Degree in the College of Medicine & Veterinary Medicine

Section of Public Health Sciences
Division of Community Health Sciences
University of Edinburgh
2008

Table of Contents

Declaration

I hereby declare,

(i) That this thesis is composed by myself

(ii) That the work presented within this thesis is my own unless otherwise stated

(iii) That this work has not been submitted for any other degree or professional qualification

List of Tables

List of Figures

Acknowledgements

Many thanks to my advisors Prof. Robin Prescott, Dr. Darren Shaw and Prof. Mark Woolhouse for all of their advice and support during the writing of this thesis. A very special thanks to Prof. Prescott for agreeing to advise me back as an MSc student, not realising that our last meeting would be nearly four years later – on his final day before retirement!

Thanks to Polly Golding, Kate Biggin, Nicola Coates-Dutton and the rest of the PhD Peers group for their friendship and support over the last four years.

Thanks to the members of IPRAVE, Epi-Group and Coffee & Chat for their helpful comments and suggestions over the course of this thesis.

Thanks to Rosa Bisset, Wilma Warwick and Sarah McAllister for their administrative support and to Sarfraz Mohammed for his computing support.

I also wish to thank a number of people for their contributions to the research in this thesis:

Mary Locking, Alison Smith-Palmer, John Cowden and Prof. Bill Reilly at Health Protection Scotland for providing the Scottish *E. coli* O157 outbreak data and for their all their help over the last three years.

Thai An Nguyen and Kathryn Blanton at the Centers for Disease Control and Prevention for providing the United States *E. coli* O157 outbreak data and their comments on drafts of Chapter 4.

Kathryn Doré and Carole Tinga at the Public Health Agency of Canada for providing the Canadian *E. coli* O157 outbreak data set, and for their help in providing additional information for and comments on drafts of Chapters 5 and 6.

John Archer at the Wisconsin Division of Public Health for providing copies of the reports on the Layton Avenue Sizzler and Mayfair Road Sizzler Outbreaks.

Lesley Allison and Mary Hanson at the Scottish *E. coli* O157 Reference Laboratory for providing a tour of the laboratory and advice on matters pertaining to phage typing.

Benjamin Bolker for providing R code to enable the construction of plots with 95% confidence intervals around the regression lines.

Abstract

Introduction *Escherichia coli* O157 is a major cause of serious infectious gastrointestinal illness in the developed world, exacting a high toll in terms of both morbidity and economic costs, with the cost of a single case as high as £3 million. Some of the highest rates of cases, approximately 20% of which are part of outbreaks, have been reported in Scotland, Canada, and the United States. Despite the efforts put into surveillance, it is estimated that up to 95% of cases go undetected. Additionally, there has been little published research on temporal trends in outbreaks or cases, the nature of primary and secondary outbreak cases or the true prevalence of cases in Scotland. Improved knowledge about long term trends, the epidemiology of outbreaks and actual case prevalence could help focus and improve surveillance efforts and reduce the number and impact of cases.

Aims and Methods The aim of this thesis is firstly to provide an introduction to the current knowledge on *E. coli* O157 and the statistical techniques to be used in the thesis, including a background *E. coli* O157 epidemiology and transmission.

Next, using data sets provided by Health Protection Scotland (Scotland), Centers for Disease Control and Prevention (United States) and the Public Health Agency of Canada (Canada), temporal trends in outbreaks and cases in each country will be analysed used linear and generalised linear models to determine whether these trends can be modelled using simple linear models, and if so, whether or not the trends are statistically significant. Where possible, trends in the three countries will then be compared using analysis of covariance to see if there are any statistically significant differences between countries. The correlation between ill and ill and positive cases will be compared across countries. The results will be discussed in relation to the issues and limitations involved in analysing date within countries and between countries with differing populations and surveillance systems.

Data on laboratory confirmed primary and secondary cases from outbreaks in Scotland, England, Wales, Japan, Canada, United States, Scandinavia and Ireland will be obtained from published papers, and the outbreaks will be described in terms

of mode of transmission (food, milk, water, animal contact/environmental or nursery), country, case numbers and median age of cases (<6, 6-16, 17-59, ≥60). Statistical procedures will be used to check for statistically significant differences in the log-transformed number of ill, confirmed, primary and (untransformed) secondary cases and compare secondary case rates between countries, modes of transmission and median age categories.

The final section will present a framework, based on the truncated negative binomial distribution, for estimating the true prevalence of *E. coli* in Scotland. The framework can demonstrate which distributions of outbreak sizes and actual event detection probabilities (sporadic cases + outbreaks) best explain the observed distribution of event sizes.

Results It was found that most outbreak and case trends between 1996 and 2004 in Scotland, 1996 and 2003 in Canada, and 1998 and 2004 in the United States could be described using simple linear models. In each of the three countries, the inability to fit simple linear models to particular trends could generally be ascribed to the effect of one or two disproportionately large outbreaks which acted as outliers and to low number of data points. In Scotland there were statistically significant decreases over time in the number of sporadic cases, the number of foodborne cases and the number of ill cases per outbreak, while in United States, the trend in the number of ill cases from outbreaks decreased statistically significantly. Lastly, in Canada, a statistically significant increase exists in the trends in the number of outbreaks, both overall and in those spread person to person and by water.

When the trends in the number of outbreaks, ill cases and outbreak size were compared between countries, there were few statistically significant differences. These differences were likely due to the substantial amount of data missing from the Canadian data set. The results from this comparison of temporal trends and the analyses of the trends in each country provide the first statistical analysis of temporal trends in outbreaks within and between countries.

Statistical analyses of the primary and secondary cases in outbreaks indicated that approximately 19% of outbreak cases are secondary. In addition there were very few statistically significant differences in secondary or primary case characteristics between countries, with the results suggesting that median age and mode of secondary transmission are more important in determining the rate of secondary cases in an outbreak.

The framework for estimation of true event and cases prevalence of *E. coli* O157 was explored using data from Scotland between 1996 and 2004. Several potentially appropriate models were presented, all suggesting that the level of under-reporting in Scotland is greater than 99%. The exploration raised a number of issues with the framework, most importantly that of differences in the definition of a sporadic cases in Scotland and in the framework. These issues were discussed, with suggestions for changes and adjustments to the framework so that it might be more appropriate for prevalence estimation.

Chapter 1 -- Prospectus and Literature Review

1.1 Prospectus

1.1.1 Introduction

In 1972, Danish researchers identified two new *Escherichia coli* (*E. coli*) antigens, O159 and O157, in porcine samples (Furowicz & Orskov 1972). A decade later, a strain of *E. coli* with one of those antigens – O157 – emerged through a series of haemorrhagic uraemic syndrome (HUS) cases and outbreaks in the United States, Canada and the United Kingdom (CDC 1983; Day et al. 1983; Johnson et al. 1983; Lior 1983; O'Brien et al. 1983; Riley et al. 1983; Wells et al. 1983) as a major contributor to severe infectious gastrointestinal illness. In 2007, twenty five years after the first reported outbreak of *E. coli* O157, infections remain a significant public health issue, especially in terms of morbidity and mortality in children (Beutin 2006; Welinder-Olsson 2005) and financial cost (Frenzen et al. 2005; Roberts & Upton 2000).

The number of cases – estimated to be as high as 73,400 per year in the United States (Mead et al. 1999), 21,600 to 103,000 in Canada (Thomas et al. 2006) and 1600 in England and Wales (Adak et al. 2002) – is high, but it is the associated mortality and morbidity that make *E. coli* O157 infections a particularly significant public health issue. In the United States and England & Wales, between 20% and 50% of persons with *E. coli* O157 are hospitalised and 0.8 to 1.2% of infected persons die (Adak et al. 2002; Chalker & Blaser 1988; van de Giessen et al. 2006; Voetsch et al. 2004b). These rates are much higher than in *Campylobacter* and non-typhoidal *Salmonella* where only up to 24% of infected persons are hospitalised and 0.015 to 0.2% die (Adak et al. 2002; Centers for Disease Control and Prevention 2006c; Chalker & Blaser 1988; Mead et al. 1999; Voetsch et al. 2004b).

There are a number of reasons that these high rates of illness and death are a particular public health concern: the considerable financial costs associated, the number of severe illnesses and deaths in young children, and the increasing evidence of long term effects from infections. A study in the United States estimated that the cost of *E. coli* O157 infections in the United States for 2003 was $405 million, with the costs for single cases with complications as high as $6.2 million (Frenzen et al. 2005). An earlier study of a 1994 outbreak in Scotland suggested that cost per case

that developed HUS was approximately £62,400, with the long term costs of the outbreak over 30 years exceeding £168, 032 per case or £11.9 million total (Roberts et al. 2000). The outbreak resulted in £600,000 of hospital costs despite being controlled within 36 hours of the time it was recognised (Roberts 2000). Much of the illness and death in outbreaks is related to HUS in children (see 1.5.2), as will be discussed below and in section 1.6.5. As approximately 15% of children under the age of 10 who are infected with *E. coli* O157 progress to HUS (Tarr et al. 2005), 3-5% of those with HUS die and a further 12-30% have severe sequelae (Nataro & Kaper 1998), a high rate of *E. coli* infection is a serious public health issue. Furthermore, a number of recent studies suggest that even infections with non-severe milder symptoms may have long term detrimental health effects (Garg et al. 2005). For instance, less than five years after the Walkerton Outbreak in Canada, the increase in risk of hypertension and reduced kidney function was significantly associated with the severity of symptoms during the outbreak (Garg et al. 2005). Increased rates of irritable bowel disease were also seen in infected persons from the Walkerton Outbreak (Marshall et al. 2006). These long term sequelae add to the financial burden of the disease in terms of medical cost and lost productivity, and thus each case prevented reduces personal and economic cost.

The continued public health significance of *E. coli* O157 infections, in particular the potential for high levels of morbidity, has been demonstrated by a series of well-publicised, large outbreaks of 50 or more cases which took place in many countries in the last three years. In 2005, cross contamination between raw and cooked meat products in a Welsh butcher shop resulted in the largest ever reported in outbreak in Wales. The outbreak involved 118 persons with confirmed infections, one of whom – a young boy – died (Outbreak Control Team - NHS Wales 2007). The following year two major multi-state outbreaks, each involving more than 70 cases, occurred in the United States. In the first outbreak, contaminated bagged spinach sickened 205 people, 103 (50%) of whom were hospitalised; at least three deaths were linked the outbreak (Centers for Disease Control and Prevention 2006h). A subsequent outbreak with probable links to food items at Taco Bell restaurants resulted in at least 71 cases and 75% of ill persons had to be hospitalised (Centers for Disease Control and Prevention 2006d). That same year, Scotland experienced its first outbreak of

sorbitol fermenting *E. coli* O157, an event of concern because this type of *E. coli* O157 is much more difficult to detect using standard laboratory methods (Incident Control Team NHS Fife 2007), and a young child died in an unrelated infection (BBC 2006). *E. coli* O157 continues to be of concern in 2007, with at least two outbreaks in Scotland in the last two months (Health Protection Scotland 2007c; Health Protection Scotland 2007d) and, at the time of this writing, an ongoing recall of more than 21 million pounds of minced beef in the United States (United States Department of Agriculture 2007b) associated with at least 40 cases of *E. coli* O157 infection (Centers for Disease Control and Prevention 2007d). That such widespread contamination could occur despite the prevention efforts by the government and the meat industry (Wachsmuth et al. 1997), suggests that minced-beef associated infections remain a risk and thus a public health issue of note. The occurrence of the two above mentioned outbreaks related to lettuces and spinach reinforces the significance of infections as a public health issue, particular as salad greens cannot be cooked or pasteurised like milk products or meat.

With *E. coli* O157 infections still a major issue, a great deal of research has been, and is being conducted. Particular foci have been the control of the bacteria in cattle (Duffy 2003; Matthews et al. 2006a; Naylor et al. 2005; Shaw et al. 2004; Sheng et al. 2006), risk factors for infection (Bryant et al. 1989; Coia et al. 1998; Locking et al. 2001; Mead et al. 1997; Parry et al. 1998; Voetsch et al. 2006; Werber et al. 2007) and mechanisms of infection (Caprioli et al. 2005; Sussman 1997; Vallance et al. 2002).

1.1.2 Overall aims

However, there are a number of important epidemiological issues that have not been addressed in the published literature. These issues – comparison of temporal trends between countries, estimating the rate of secondary cases in outbreaks and estimating true infection prevalence in Scotland – involve the analysis of outbreaks as a whole, rather than outbreak by outbreak. The overall aim of this thesis was first to model temporal trends in *E. coli* O157 cases and outbreaks, where appropriate, using simple linear models, and to compare the modelled trends between countries to gain knowledge about the possible epidemiological factors involved in *E. coli* O157

infection. Emerging from this study is the aim of creating and testing a framework to provide a more accurate estimation of true infection prevalence in Scotland. Finally, using data from published outbreak reports, the last aim is to estimate the overall proportion of outbreak cases which are the result of secondary transmission and to determine whether median age, country, modes of primary transmission or mode of secondary transmission have a statistically significant effect on the rate of secondary cases in outbreaks.

1.1.3 Chapters 2-6: Comparison of temporal trends

In the first section of this thesis, the aim is to compare temporal trends in *E. coli* O157 between countries in order to better understand the epidemiology of the bacterium with the *a priori* hypothesis that it is possible to fit simple linear models to such trends.

There has been very little statistical analysis of long term temporal trends in individual countries, including those which have had high national or regional rates of infection such as Scotland, Canada and the United States. Instead, most published analyses have been descriptive in nature and/or focused on outbreak investigations (see Chapter 7, Table 7.2 for an extensive listing of published outbreak reports). This emphasis on descriptive and outbreak reports likely results from the need of infectious disease agencies such as Health Protection Scotland to place a primary emphasis on acute issues such as the investigation and control of outbreaks, and the lack of consistent long term surveillance data sets. Yet, analyses and between-country comparisons of temporal trends are important because they have the potential to reveal possible long term trends, suggest as to whether prevention programs are having an effect on case and outbreak trends, and suggest as to whether epidemiological factors affecting *E. coli* O157 infection trends are linked to within-country factors.

With a national level outbreak data set available for Canada for the first time, and more than eight years of data now available in existing surveillance data sets from Scotland and the United States, temporal trend comparison between these countries is now possible. Inclusion of other countries, such as England & Wales and Japan in the comparison would be of interest, but data is not available.

Thus, the aim in the next four chapters is to compare temporal trends in both cases and outbreaks between Scotland, the United States and Canada, where possible, in particular relating to modes of transmission to see which trends are statistically significant. In order that the comparisons are both statistically feasible given the relatively low number of data points and potentially explainable in terms of biological mechanisms, analytic methods are limited to simple linear models. However, before comparisons can be made, the data set from each country will have to be considered and the trends modelled statistically. For instance, studies have suggested that in Scotland there has been in a shift in predominant modes of transmission from food to modes such as water and animal exposure (Strachan et al. 2006). Statistical models might be able to indicate whether this trend is statistically significant.

Chapter 2 will provide an introduction to the comparison, including a presentation of the issues involved in comparing data between countries, and an introduction to the statistical methods using in modelling of temporal trends. The subsequent three chapters will present the results from the modelling of the trends within each country (in order): Scotland, United States and Canada. The sixth chapter of the thesis will present the comparison of temporal trends between Scotland, the United States and Canada with a discussion of the trends that can be compared, particularly whether or not there are any statistically significant differences between countries. Additionally, there will be a further exploration of any issues that were anticipated or arise with regards to the comparability of data. These include factors such as large outbreaks or shifts in surveillance methods which significantly affect temporal trends, as well as the differing time periods for which data was available and the comparability of mode of transmission categories between countries.

1.1.4 Chapter 7: Secondary cases in outbreaks

The aim of the analyses in Chapter 7 is to estimate the rate of secondary spread in *E. coli* O157 and determine whether the mode of primary or secondary transmission, country and age of cases have a statistically significant effect on the rate. The research for this thesis on secondary cases developed out of an initial exploration of factors that might affect parameters in the distribution of outbreak sizes (Chapter 8).

Outbreak reports and observations in some literature reviews suggest that secondary cases, cases in which infection is not acquired from the original source of the infection, but from direct or indirect (e.g. though water) contact with another infected person, may account for as many as 40% of cases (Armstrong et al. 1996; Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a; Parry & Palmer 2005). Secondary cases are often noted in outbreaks which occur in locations such as nurseries, nursing/care homes or private households. Such locations frequently have a high proportion of persons – the young, ill and elderly – who are at a higher risk for severe outcomes. Additionally, the severity of outcomes in secondary cases may be equivalent to that for primary cases (Locking et al. 2006a). However while Scotland has published annual data on the proportion of secondary cases overall (outbreak and sporadic cases) (Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a), there has been little systematic quantification or characterisation of secondary outbreak cases in the published literature.

The study in Chapter 7 aims to fill this gap in the published literature by using data from published outbreak reports to estimate the rate of secondary cases in outbreaks and to analyse the relationships between primary and secondary cases by mode of transmission, country and age of cases. The results of the analyses will provide insight on whether factors such as age, country and mode of transmission affect the rate of secondary cases in outbreaks. Information on which, if any, factors are associated with higher rates of secondary cases could potentially be helpful in targeting programs aimed at reducing secondary transmission in outbreaks.

1.1.5 Chapter 8: New framework for the estimation of case prevalence

In Chapter 8, the aim is to develop a framework to estimate the true *E. coli* O157 infection prevalence. The need for a tool to estimate of infection prevalence that had less uncertainty than the reporting triangle-based estimations emerged out of the work on modelling of temporal trends which involved incomplete data sets. *A priori*, the hypothesis was that the actual prevalence of *E. coli* O157 infections in Scotland could be modelled based on the known distribution of detected event size and a hypothesised function of the probability of case detection.

A series of studies from the United States, Canada and England & Wales have indicated that *E. coli* O157 infections are significantly under-reported – with high estimates suggesting that as many as 98% of infections may not be reported to national surveillance agencies (Adak et al. 2002; Bender et al. 2004; Michel et al. 2000; Thomas et al. 2006). However, the current studies have relied on data from surveys which calculated the number of cases lost to surveillance at each stage between infection and reporting to the surveillance agency (see Chapter 1.9 for more detail on these stages) or on estimates of hospitalisation rates. Such stepwise estimation methods are known to have inaccuracies and provide wide ranges of estimates (Adak et al. 2002; Bender et al. 2004; Michel et al. 2000; Thomas et al. 2006). This is because the calculations may, for instance, be based only on data from one specific region within a country resulting in bias due to differences in underlying rate or physician/lab practices. Also, any error is compounded from step to step. No study estimating true prevalence has ever been published for Scotland, despite Scotland having had some of the highest rates of reported infection in the world. Yet, an accurate estimation of infection prevalence is of importance because it would allow the need for prevention and treatment resources to be more accurately calculated, and thus ensure that enough funding is allocated for treatment, prevention and research. In particular, given the historically high rates of infection in Scotland as compared to other developed countries, it would be of interest to determine whether the high rates reflect a true high prevalence or rather a higher rate of case detection.

Given these issues, the eighth chapter of this thesis proposes and explores the use of a framework for estimating *E. coli* O157 infection prevalence using the Scottish data. The framework is based on the assumption that every infection/case is part of an event, whether the event is a sporadic case or an outbreak, and while an event is either detected or not, not all of the cases in an event are necessarily detected. The use of survey data is avoided by calculating the true number of cases based on the distribution of the proportions of detected event sizes, which is constructed from a matrix based on the distribution of actual event occurrence and function of case detection. If a model can be constructed in which the distribution of the proportions

of detected event size occurrence matches that observed in Scotland, it can be used to estimate the true prevalence of infection.

Not only does it have the potential to provide a more accurate estimate of infection prevalence, this new approach can demonstrate what sized outbreaks are being missed by surveillance. This information, partnered with data from estimates based on surveys, might be able to provide a more detailed and accurate estimate of gaps in current surveillance systems and suggest foci for improving surveillance systems. In this thesis, the framework will be presented with a discussion of how the appropriate distributions/functions and parameters were selected. Models of the distribution of detected event sizes will also be constructed using the framework, and presented in order to determine whether the approach can be used to accurately estimate model prevalence.

1.1.6 Conclusion

The aim of the studies presented in these eight chapters involves first the comparison of temporal trends in Scotland, the United States and Canada (Chapters 2 -6) using simple linear models. From this research emerges need for a better estimate of infection prevalence, addressed by the development of a different framework for estimating *E. coli* O157 infection prevalence with application to the Scottish data (Chapter 8). Lastly, to look at factors which may affect the size of outbreaks, secondary case rates in outbreaks will be estimated and the significance of country, mode of primary and secondary transmission and age in secondary case rate will be analysed (Chapter 7).

1.2 Introduction to the literature review

First however, a background to *E. coli* O157 will be provided, with a particular focus on topics pertinent to the analyses in this thesis, including verotoxins, modes of transmission, methods of detection and case and outbreak epidemiology. Whilst a general background to *E. coli* O157 will be provided, the particular focus is on topics that are of relevance to and provide a background to the studies in Chapters 2 – 8. These include verotoxins, modes of transmission, methods of *E. coli* O157 case detection, case types (outbreak and sporadic), outbreak detection, outbreak and

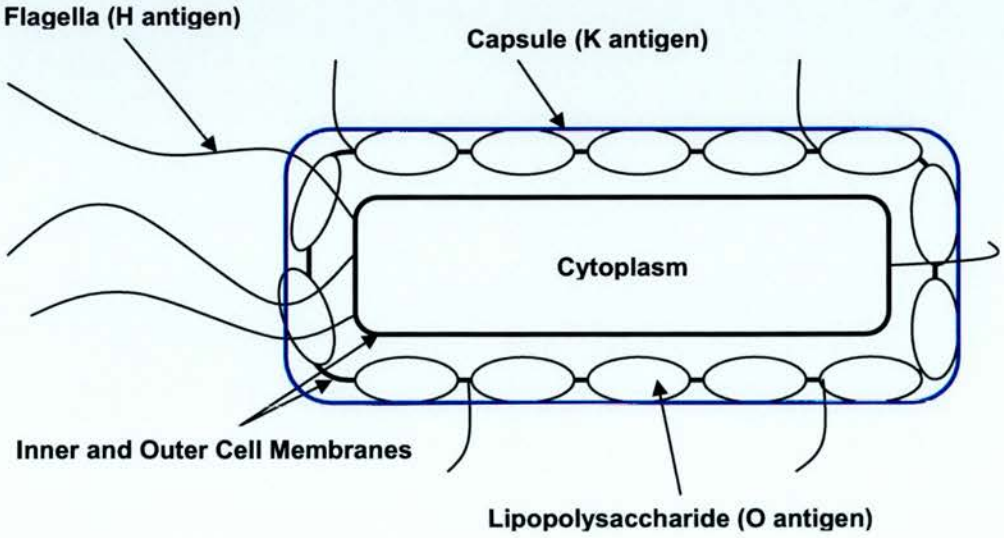
sporadic case surveillance and methods of analysing temporal trends. Information was obtained both from books and from searches of databases including Web of Science and Medline using one or more of the terms "VTEC", "STEC", "O157", "verotoxins", "outbreaks", "infectious intestinal illness" and "infectious gastrointestinal illness". Search terms for data on methods of assessing under-reporting and estimating prevalence also included "underreporting", "burden of illness", "estimation of prevalence", "prevalence", "reporting triangle" and "burden". Additional data was obtained from posters presented at the VTEC 2006 conference in Melbourne, Australia, the websites of Health Protection Scotland (HPS), the Centers for Disease Control and Prevention (CDC) and the Public Health Agency of Canada (PHAC) as well as from conversations and correspondence with epidemiologists at HPS, the CDC and the PHAC.

1.3 Molecular and microbiological background

1.3.1 *E. coli*

E. coli, a common species of gram-negative bacteria (e.g. not stained by gram stain) is an important part of the normal human and mammalian intestinal flora (Kaper et al. 2004), however not all strains are innocuous to human hosts (Nataro & Kaper 1998).

Figure 1.1: Structure of an *Escherichia coli*
The structure of an *Escherichia coli*, showing the locations of the O, K and H antigens.



The bacterium is short and rod shaped with flagella that contain H antigens. Lipopolysaccharides, which contain the O antigens, are found at the border of the outer cell membrane and the capsule (Fig. 1.1) (Murray et al. 2002; Sussman 1997).

Strains are classified both by their O (lipopolysaccharide or somatic), K (capsular) and H (flagellar) surface antigens, with serogroups defined by the O antigen and serotypes by both O and H antigens, and by the nature of their pathogenicity in humans (i.e. "pathotypes") (Naylor et al. 2005). At least 170 O antigens, 56 H antigens and 80 K antigens have been identified (Orskov & Orskov 1992). There are six classes of *E. coli* which are known to cause gastrointestinal illness in humans – enterohaemorrhagic *E. coli* (EHEC), enteroinvasive *E. coli* (EIEC), enteropathogenic *E. coli* (EPEC), enteroaggregative *E. coli* (EAEC), diffusely adherent *E. coli* (DAEC) and enterotoxigenic *E. coli* (Kaper et al. 2004). This review will focus on EHEC.

1.3.2 Enterohaemorrhagic *E. coli*

EHEC strains are defined by the production of the virulence factor Shiga toxin (Stx, also known as verocytotoxin (Vt)) (see 1.3.3) and the ability to adhere to and interact with intestinal cells via the presence of the locus of enterocyte effacement (LEE). This is a 35kb gene segment which encodes the ability of the bacteria to attach and interact with intestinal epithelial cells (Kaper et al. 2004). Infection by EHEC in humans is characterised by haemorrhagic colitis – bloody diarrhoea and HUS. Of the EHEC serogroups, including O26, O111 and O103, the most important and high profile in North America, the United Kingdom and Japan is *E. coli* O157 (Caprioli et al. 2005; Kaper et al. 2004).

1.4 *E. coli* O157

1.4.1 History

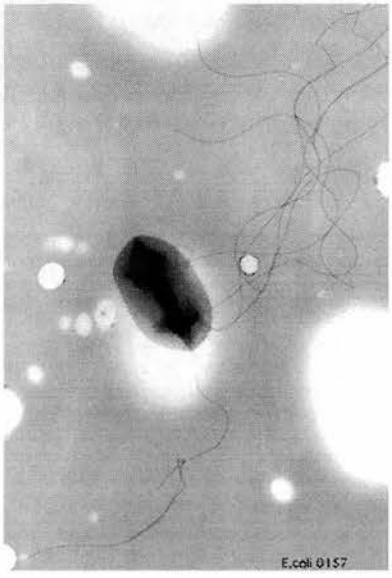
The O157 serogroup of *E. coli* was first identified in pigs in 1972 (Furowicz & Orskov 1972), but not recognized as a human pathogen for another decade (Riley et al. 1983). At that time the serotype *E. coli* O157:H7 was linked to two outbreaks of bloody diarrhoea and HUS in Oregon and Michigan attributed to consumption of undercooked minced beef (Riley et al. 1983), as well as being linked to cases or outbreaks in Canada (CDC 1983; Lior 1983) and the United Kingdom (Day et al. 1983). Simultaneously, Karmali and colleagues made the connection between HUS

and a cytotoxin produced by strains of *E. coli* (Karmali et al. 1983), and the cytotoxin, verotoxin, was thereafter identified in the serotype isolated in the initial HUS outbreaks (O'Brien et al. 1983). However, evidence from retrospectively tested specimens and serum samples has implicated *E. coli* O157 in cases of HUS as far back as 1974 in the United States (Riley et al. 1983), the United Kingdom (Day et al. 1983), Canada (Johnson et al. 1983) and the Netherlands (Chart et al. 1991). It is thought that the O157:H7 serotype evolved from *E. coli* O55:H7 strains which acquired new virulence factors – attributes that allow the bacteria to successfully invade hosts - via lateral transfer (Whittam 1998). Lateral transfer includes transduction, in which genetic material is transferred from cell to cell via phages (short for bacteriophage), viruses containing either RNA or DNA which infect bacteria (Griffiths et al. 1996). The verotoxin genes, which are the key virulence factors in *E. coli* O157, are thought to have been acquired via transduction (Whittam 1998).

1.4.2 Structure

E. coli O157 (Figures 1.1 and1.2) is an anaerobic bacteria (Coia 1998a) with O antigen type 157 (Griffin and Tauxe, 1991).

Figure 1.2: Transmission electron micrograph of *Escherichia coli* O157:H7
A transmission electron micrograph of *E. coli* O157:H7 showing the flagella. The image is taken from the Center for Disease Control and Prevention's Public Health Image Library, provided by Peggy S. Hayes and credited to Elizabeth H. White.



Within the O157 serogroup, the most important serotype is O157:H7, though the non motile (NM or H-) serotype is also included in sections of this thesis. The

chromosome for *E. coli* O157 is approximately 5.5 Mb in length (Hayashi et al. 2001), which is in the middle of the genome size spectrum for bacteria, and the bacterium normally contains pO157, a 60MDa plasmid involved in virulence and found in most O157 strains and in other EHEC strains (Nataro & Kaper 1998), and produces one or both of two verocytotoxins (Vtxs) (Mead & Griffin 1998).

1.4.3 Verotoxins

The vast majority of O157 *E. coli*, including O157:H7, are verocytotoxin/shiga-toxin-producing *E. coli* (VTEC or STEC), meaning that they produce verotoxins (Sussman, 1997). (All EHEC are VTEC, however there are VTEC which are not EHEC). These verotoxins, verotoxin 1 (Vt1) and verotoxin 2 (Vt2), so named because of their toxicity to cultured Vero cells (a lineage of African green monkey kidney epithelial cells frequently used in cell culture) (Konowalchuk et al., 1977), are often called Shiga-like toxins (SLTs, stxs) because of their similarity to the Shiga toxin of *Shigella* dysenteriae (Noël & Boedeker 1997). Vt1 differs from Shiga toxin by only a single amino acid (Feng, 1995; Coia et al., 1998), with two variants Vt1c and Vt1d (Karch et al. 2005) having been identified in recent years. Vt1c is generally associated with mild clinical symptoms, whilst there as of yet no data linking Vt1d to specific clinical outcomes (Kuczius et al. 2004). Vt2, which is 55-60% homologous with Vt1 (Coia 1998a; Kaper et al. 2004) has four variants: 2c, 2d, 2e and 2f (Ochoa and Cleary, 2003). These variants differ mostly in their b subunit, the amino acid sequences of which have from 84% to 98% homology with Vt2 (Melton-Celsa and O'Brien, 1998). Strains of E. coli O157:H7 may contain one or both of these toxins - those with neither are not considered to be VTEC.

Of the two Vt types, Vt2 is much more common in human infections, occurring in more than 85% of isolates in Scotland from 2002-2004 (Locking et al. 2003b; Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a), England & Wales from 1992-2003 the EU in 2000 (Fisher 2002), Washington State in 1987 (Ostroff et al. 1989), and Canada from 1994 - 2000 (Woodward et al. 2002). However, the proportions of strains that also produce Vt1 vary widely between countries, though the variation within countries is much less pronounced over time. In England and Wales, the percentage of isolates between 1992 and 2003 that were positive for both

Vt1 and Vt2 ranged from 19% to 28% (Table 1.1) and in Scotland fewer than 16% of isolates were Vt1+Vt2+ (Thomson-Carter et al., 1996; Locking et al., 2003).

Table 1.1: Verocytotoxin gene types of *E. coli* O157 isolates in England & Wales: 1989 – 2003
The percentage of isolates or outbreaks from England and Wales for each three-year or single year period which produced only Vt2, Vt1 and Vt2 or only Vt1. Data from (Frost et al., 1993; Thomas et al., 1996) (Willshaw et al., 2001a; Health Protection Agency, 2004). Blanks indicate where no information was provided in the source.

	Vt 2 only	Vt 1 and Vt 2	Vt 1 only
1989-91	82% of outbreaks		
1992-94	75.4%	24.1%	0.5%
1995	80.2%	19.2%	0.6%
1996	76.1%	23.3%	0.6%
1997	72.8%	26.1%	1.1%
1998	74.7%	24.6%	0.7%
1999	78%	21%	<1%
2000	72%	27%	<1%
2001	76%	23%	<1%
2002	76%	23%	
2003	71%	28%	<1%

Contrastingly in North America & Japan, nearly three quarters of isolates have both Vt1 and Vt2 (Mead and Griffin, 1998). For instance, 80% of 444 isolates tested in Canada in 1995 were positive for both VT types, and both types were found in samples from 16 of 22 (73%) outbreaks in the United States from 1982 to 1993 (Strockbine et al., 1998). Over time, as illustrated by Table 1.1, the percentages of isolates that were Vt1+Vt2+ only varied between 19% and 28% over 11 years (Frost et al., 1993; Thomas et al., 1996) (Willshaw et al., 2001a; Health Protection Agency, 2004). The presence of Vt1 alone is very rare, comprising less than 2% of cases in Canada and England & Wales (Table 1.1) and not found in recent years in Scotland (Locking et al. 2003b; Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a; Woodward et al. 2002).

Vt2 is 100 times more toxic to rabbits than Vt1 (Noël & Boedeker 1997), and the presence of Vt2 has been associated with a greater severity of disease in humans (Friedrich et al. 2002). Vt2-only strains have the highest risk of HUS, whilst those with just Vt1, the lowest (Friedrich et al. 2002; Melton-Celsa & O'Brien 1998; Ostroff et al. 1989). For instance, in England & Wales, between 1983 and 1994, 90% of isolates from *E. coli* O157 infected persons with HUS contained just Vt2, while overall only 70% of isolates were Vt2+Vt1-. In 1995-2001, the percentage of Vt2+ only rose to 96% for isolates from HUS patients and 76% for all isolates (Smith et al., 2003).

However, whilst it is clear from the studies discussed above that verotoxin type has an important role in determining pathogenicity, the data sets provided by the three health organisations Centers for Disease Control and Prevention (CDC), Public Health Agency of Canada (PHAC) and Health Protection Scotland (HPS) and used in thesis do not include data on Vt type. Thus no analysis or comparison of Vt type trends could be performed. Additionally, analysis of Vt type as a potential factor in secondary case rate was precluded by the lack of data on Vt type provided in published reports.

Vt production alone does not make an *E. coli* O157 strain capable of causing disease. It has been suggested that the 60 MDa virulence plasmid and the LEE are also vital for pathogenicity. The plasmid contains the gene for a haemolysin, which may help the bacteria to obtain iron from blood and the LEE codes for factors that play a vital role in allowing the bacteria to attach to the intestinal epithelial cells (Mead and Griffin, 1998). Other virulence factors include colonization factors such as Intimin (genes for which are in the LEE), ToxB, Saa and Efa-1/LifA, effectors such as EspP, EspF, EspH, and toxins including StcE and Ehx (Kaper et al. 2004).

1.4.4 Phage type

Phage typing was one of the earliest methods that emerged for categorizing *E. coli* O157:H7 strains and is still widely used (Strockbine et al. 1998) in countries like Scotland (Locking et al. 2006a). It most often used for preliminary linking of individual cases to identify potential outbreaks because the procedure is not time intensive. However, the availability of phage typing is limited as few labs are able to accommodate the upkeep of phage stocks. The procedure involves exposing a isolate sample to a battery of selected lytic phages and noting the lysis pattern that results (Strockbine et al., 1998). Each phage type (PT) has a unique series of lytic patterns.

The current phage-typing scheme for *E. coli* O157, used in many countries including Canada and Scotland, was set up in 1987 (Ahmed et al. 1987) and expanded in 1990 to recognize 62 distinct phage types (Khakhria et al. 1990). It has since been further expanded to approximately 82 phage types (PTs) (Strockbine et al. 1998), though the phage types of *E. coli* O157 referred to as PT 28 in Scotland and PT 21 in England &

Wales were shown to be (nearly) identical, and thus in 1998, it was agreed that the phage type should be redesignated PT 21/28 (Anonymous, 1998). This scheme is used in countries worldwide, including the U.S., Canada, Japan, Great Britain, Australia and parts of Europe (Strockbine et al., 1998), though isolates are not routinely typed in the United States where pulse field gel electrophoresis is used as the standard. Since isolates generally are not typed, phage type information is not included in the United States data set.

In the two decades since the phage-typing scheme was first introduced, there have been distinct temporal and between-country variations in phage types. Changes in phage distribution over time have been seen in countries where isolates are routinely phage-typed. For instance, typing of Canadian samples in 1990 revealed 35 distinct phage types, with PT 1, PT 4, PT 8, PT 31 and PT 14 the most common (Table 1.2) (Khakhria et al., 1990).

Table 1.2: Phage types of *E. coli* O157 in Canada
Phage types in Canada in 1982-1989 and 1995. Data is taken from (Khakhria et al., 1990) (Khakhria et al., 1997). Blank spaces are left where phage types are not mentioned.

	PT 1	PT 2	PT 4	PT 8	PT 14	PT 31
1982- 89	30.5%	4.8%	21%	13.5%	8%	8.9%
1995	5.9%	11.3%	4.2%		52%	9.8%

Five years later, the most common phage types were PT 14 (52%), PT 2 (11.3%), PT 31 (9.8%), PT 1 (5.9%) and PT 4 (4.2%) (Khakhria et al., 1997). However, because phage type information on the majority of Canadian outbreaks was not available for this thesis, temporal trends in phage types will not be examined in Chapter 5. Changes in phage types can also be seen in England & Wales and Scotland (Table 1.3).

Table 1.3: Phage types of *E. coli* O157 in England and Wales
Common phage types of *E. coli* O157 isolates in England and Wales from 1983 to 2003. Data is taken from (Thomas et al., 1996; Cheasty et al., 2000; Smith et al., 2003; Health Protection Agency, 2004). Blank spaces are left where phage types are not mentioned.

	1983-1994	1992-1994	1997-1999	1995-2001	1998	1999	2000	2001	2002	2003
PT1	7%	7%								
PT2	49%	46%	32%	32%	31%	26%	19%	21%	16%	16%
PT4	6%		7%		8%	8%	6%	3%	8%	4%
PT8		8%	17%	16%	18%	16%	21%	17%	21%	26%
PT14		4%				4%	4%		3%	3%
PT21/28		2.4%	19%	20%	16%	29%	30%	36%	33%	35%
PT32		3.6%	8%		8%	7%	10%	5%	6%	5%
PT49	22%	17%								

Between 1983 and 1994, PT 2 was the predominant phage type in England & Wales, followed by PT 49 and PT 1 (Smith et al., 2003). By 1997-99, PT 2 had decreased to 32% of the total, while PT21/28, PT 8 and PT32 were the next most common (Cheasty et al., 2000). Between 1995 and 2001, PT2 was most common.

In Scotland, the most common phage types in 1992 were PT49 and PT2, but a decade later PT 21/28, PT 8 and PT 2 were most frequently seen, with PT 2 a distant third (Table 1.4)(Locking et al., 2003). PT 49 went from being the most common phage type in 1992, to the fourth most common, behind PT 2, PT 28 and other phage types in 1994, and by 2002, was not amongst the top 8 most common phage types (Reilly, 1997; Locking et al., 2003). Temporal trends in the two most common phage types – 21/28 and 2 - will be examined in detail in Chapter 3.

Table 1.4: *E. coli* O157 phage types in Scotland
The most common phage types in Scotland from 1984 to 2003. Stars (*) indicate phage types of which mention is made, but no specific value is provided. Blank spaces are left where phage types are not mentioned. Data is taken from (Coia et al. 1996; Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a; Sharp et al. 1994a; Smith et al. 2002)

	1984-93	1994	1998	1999	2000	2001	2002	2003	2004
PT1	*								
PT2	33%		14%	9%	7%	9%	5.2%	9%	17%
PT4	*		6%			*		6%	
PT8	*		7%	11%	7%	11%	13%	6%	12%
PT14	*							0.7%	3%
PT21/28	*		59%	57%	67%	63%	64.5%	65.8%	58%
PT32	*		6%		8%	5%	3%	2.1%	4%
PT49	36%	8%				*			

Differences in phage types are also evident between countries (Cheasty et al., 2000). For instance, when human samples from six countries were typed, PT 47 was found only in Japan and PTs 49, 50, 51 and 52 only in the United Kingdom (Khakhria et al., 1990). In another instance, while PT 21/28 was very common in England & Wales and Scotland between January of 1997 and June 1999, it was not among the three most common phage types in Northern Ireland, Ireland, Denmark, Belgium, Finland, Germany or Sweden (Cheasty et al., 2000). Differences can also be seen in phage types of outbreaks: in 1997, the leading phage types in outbreaks in Canada were PT 14 (80% of 19 outbreaks), and PT 4, 31, 32, 34 (5% each), yet in Scotland, the only overlap was PT 4 (9% of 11 outbreaks), while PT 21/28, 2 and 8 were the types for the remaining 81%. As, however, data on phage types was only available

from Scotland, these differences will not be analysed in Chapters 4 (United States) and 5 (Canada) or as a part of the comparisons in Chapter 6.

Phage and Vt types are strongly linked, and though the mix of phage types may change over time, the VT profiles of particular phage types usually do not (Willshaw et al., 2001b). Data on Scottish cases indicates that in 1995, all PT49 isolates were Vt1-Vt2+ and those from PT54 were Vt1+Vt2+(Thomson-Carter et al., 1996). Seven years later in 2002, PT21/28 and PT14 isolates were all Vt2+ only, PT2 isolates were predominantly Vt1-Vt2+ and PT8 isolates were predominantly Vt1+Vt2+ (Locking et al., 2003), and 2004 all PT 21/28 isolates were also Vt2+ only, whilst almost all PT8 isolates were Vt1+Vt2+. Similar patterns are seen in England & Wales, where during the periods of both 1992-1994 and 1995-2001, PT 8 isolates were predominantly Vt1+/Vt2+, while PT2, PT21/28 and PT49 isolates were predominantly Vt1-/Vt2+ (Smith et al., 2003). The association between phage type and verotoxin profile presumably is linked to inability or ability of specific typing phages to lyse bacteria containing specific verotoxin types.

1.4.5 Testing and screening procedures

1.4.5.1 Testing -- Sorbitol fermentation test

E. coli O157 can be distinguished from other bacteria in lab specimens by several methods, of which the sorbitol fermentation test is the most widely used (Strockbine et al. 1998). The sorbitol fermentation test is based on the fact that the majority of *E. coli* O157 strains cannot ferment D-sorbitol overnight, and thus can be distinguished from other *E. coli* and bacterial strains by testing for this trait (Mead and Griffin, 1998). Samples are incubated in Sorbitol-MacConkey Agar (SMAC), and clear plaques indicate the likely presence of *E. coli* O157. The test is straightforward and inexpensive, but due to the fact that sorbitol fermenting VTEC strains have been found in Europe and Scotland (Gunzer et al., 1992; SCIEH, 2002), and a study has shown that *E. coli* O157:H7 may mutate into the sorbitol-fermenting phenotype in sorbitol producing foods,(Feng, 1995) the plain SMAC test is no longer considered to be the gold standard.

To combat this problem, modifications have been made to increase the quality the original SMAC test. Samples can be refined prior to SMAC incubation by

immunomagnetic separation with polystyrene beads coated in O157 antibodies (Strockbine et al., 1998). Also, the SMAC media can be made more specific by the addition of other components. CR-SMAC contains cefixime rhamnose (Chapman et al., 1991) and CT-SMAC contains cefixime and potassium tellurite (Zadik et al., 1993). CT-SMAC is commonly used for screening isolates in the United Kingdom, and is one method used in Italy (Smith et al., 2002). However, some strains, like the sorbitol-fermenting strain that was isolated in Scotland in 2002, are also tellurite sensitive, so even CT-SMAC agar may not be sufficient to detect the presence of some *E. coli* O157 strains (SCIEH 2002a). This may hamper the detection of cases or outbreaks of sorbitol-fermenting outbreaks, in particular where the symptoms of many patients may be relatively minor. In addition, the diagnostic techniques differ between countries (Smith et al. 2002), with potentially lower rates of infection reported in countries or regions that use less specific or sensitive isolation methods. The identity of colonies that come up positive via the SMAC test can then be further confirmed by testing for O157 antigen with O157 antiserum or latex reagent kits (Mead and Griffin, 1998).

1.4.5.2 Testing - other detection methods

E. coli O157 cases can also be identified by a number of techniques which test for the presence of Vtxs or nucleic acids (i.e. genes) associated with *E. coli* O157 (Strockbine et al. 1998). Biologic methods involve assays which can detect free verocytotoxin in faecal samples, though there may be false negative results when the tests are done in the first few days of an infection (Strockbine et al. 1998). Immunologic methods such as immunoblots or enzyme immunoassays (EIAs) detect *E. coli* O157 by using antibodies specific to vtx to identify the presence of the toxins. Finally, the presence of VT genes, O157 plasmid or eae gene nucleic acids in isolates can be detected using DNA probes or Polymerase Chain Reaction (PCR)(Strockbine et al., 1998). A number of assays, including western blots, cell culture assays and EIAs can also be used to detect antibodies to vtxs or *E. coli* O157 cell components in serum. These methods may improve the recognition of cases, especially in situations where testing is delayed and thus shedding of the bacteria has ceased. Serum sampling has in some outbreaks picked up cases that were not identified through faecal testing (Marsh et al. 1992). For instance, 165 cases in the Wishaw Outbreak

where either stool samples were negative or were not taken tested positive via serology (Cowden et al. 2001). Thus, the use of serological testing could potentially improve the recognition of secondary cases, the rate of which are analysed in Chapter 7.

1.4.5.3 Subtyping

Further identification of an *E. coli* strain can be made via phenotypic or genotypic methods. Among the phenotypic properties that can be used to divide *E. coli* O157 into further subtypes are phage type (see 1.4.4), verotoxin type, antibiotic resistance and the presence or absence of *E. coli* attaching and effacing (aea) genes (Karch et al., 1999). The major genotypic method is Pulse Field Gel Electrophoresis (PFGE). Since the number of *E coli* O157:H7 subtypes continues to increase (Willshaw et al. 1997) and exact determination of subtype is often required to link and trace both animal and human cases in outbreaks, it is now considered good practice to determine phage type, VT type and also run a PFGE to examine subtype (Willshaw et al., 1997).

PFGE, which reveals the unique DNA restriction fragment length polymorphisms of each subtype, is also frequently used in typing O157 strains because it is effective at identifying identical or closely related O157 strains (Willshaw et al., 1997). While closely related strains can vary by up to two bands, meaning that for instance, isolates within a single outbreak may not have a completely identical pattern of bands, PFGE typing has been of great utility in proving links between cases otherwise not known to be related (Bender et al. 1997; Strockbine et al. 1998). Currently, PFGE typing is widely used in countries such as Canada, the United States and Scotland, though between-country comparisons may be limited by the use of different restriction enzymes for PFGE. Verotoxin types, discussed in more detail in Section 1.4.3, can be determined by immunologic methods, the most common being the use of PCR with probes for specific VT types with or without restriction analysis (Karch et al. 1999; Kuczius et al. 2004).

1.5 Pathogenesis and clinical symptoms

1.5.1 Symptoms and treatment

The pathogenicity of *E. coli* O157 in humans results from the ability of the verotoxins, which inhibit protein synthesis, to damage the endothelial cells of the intestinal lining (U.S. Food & Drug Administration, 2003). The damage to the intestinal endothelial cells causes acute gastrointestinal disease, known as haemorrhagic colitis. Most cases are relatively mild, and as many as 15% of infections may be asymptomatic (Todd & Dundas 2001). Typical symptoms, which include watery and/or bloody diarrhoea, severe abdominal cramping and nausea, but rarely any fever, usually begin after an incubation period of up to two weeks – with an average of about 3 days - and last about a week (Coia et al., 1998; U.S. Food & Drug Administration, 2003). In as many as 90% of laboratory confirmed cases, diarrhoea becomes bloody after a few days (Karch et al. 2005). Treatment for infection is generally focused on ameliorating symptoms, monitoring for signs of HUS or other complications and maintaining hydration (Mead & Griffin 1998; Tarr 1995; Todd & Dundas 2001). The use of antibiotics is contraindicated because of the risk of killing normal intestinal flora which might help to prevent toxin absorption by the gut, and because certain antibiotics may in fact promote the expression and release of verotoxins (Zimmerhackl 2000). Studies suggest that antibiotic treatment increases the risk of poor clinical outcome and HUS (Pavia et al. 1990; Wong et al. 2000). The use of anti-motility agents is also discouraged because slowing down normal gut peristalsis slows down clearance of the bacteria from the intestine and thus increases the exposure of the endothelial cells to the verotoxins (Tarr 1995). The majority of cases (up to 85-95%) resolve, with or without professional medical care, within a week of symptom onset (Mead & Griffin 1998; Todd & Dundas 2001).

1.5.2 Complications - HUS

However, in anywhere from 3% to 25% of the cases, usually in the very young or old, the acute illness leads to HUS (Mead and Griffin, 1998; Todd and Dundas, 2001; Ochoa and Cleary, 2003) (Naylor et al. 2005). HUS, a syndrome marked by acute renal failure, thrombocytopenia and microangiopathic hemolytic anemia (O'Brien and Kaper, 1998) occurs when the verotoxins enter the bloodstream from the large intestine, where they are brought via leukocytes to endothelial cells in other

locations, most vitally the glomeruli in the kidneys. Once the verotoxins enter the endothelial cells, they trigger pathological changes which cause the characteristic symptoms of HUS (see above) by interfering with normal protein synthesis (Karmali 2004). HUS is a major cause of acute renal failure in children and the mortality rate can be as high as 20% (Griffin and Tauxe, 1991). Those who do survive may have long-term sequelae such as kidney failure, neurological damage, colonic stricture and glucose intolerance (Mead and Griffin, 1998). Infections that do not progress to HUS have not been thought to have such long term effects, but recent studies of persons infected during the Walkerton Outbreak in Canada suggest that *E. coli* O157 infection may indeed be linked to increase risks of pathologies such as high blood pressure, irritable bowel syndrome and decreased renal function (Garg et al. 2005; Marshall et al. 2006).

Among the suggested risk factors for developing HUS are being female, very young or advanced age (<5 years or >65 years) and mental retardation (Griffin and Tauxe, 1991) (Todd & Dundas 2001). It is suggested that young children are at a higher risk for HUS because larger amounts of verotoxin are released into their systems during an infection (Monnens et al. 1998), though the reasons behind this are not fully understood. In addition the risk may be linked to the non-fully developed immune status of young children (Cimolai et al. 1990). However, the most indicative factor in determining HUS risk seems to be the Vt composition of the implicated *E. coli* O157 strain. Strains expressing only VT2 have the highest risk of causing HUS (Ochoa and Cleary, 2003), while those with both verotoxins having a lower risk and those expressing just Vt1 the least risk.

1.6 Epidemiology

1.6.1 Infectious dose

Investigations of outbreaks and sporadic infections suggest that the infectious dose for *E. coli* O157:H7 is less than 100 organisms (Caprioli et al. 2005), and could be as little as 10 organisms. Tuttle and colleagues, in studies of hamburger meat implicated in a 1992-93 outbreak, deduced an infectious dose of ~700 organisms, but suggested that the actual dose was likely to be far less, as the meat was cooked before being eaten, thus killing many organisms (Tuttle et al., 1999). Other

indications of low infectious dose include the occurrence of waterborne transmission, where infection occurs despite the bacteria counted being diluted via the water. Data from an outbreak caused by contaminated dry fermented salami (Tilden et al., 1996) suggested an infectious does of less than 50 organisms. The low infectious dose is of importance because it may account for the large size of some outbreaks, and the high rates of person to person transmission within household. For instance, in large waterborne outbreaks like the Walkerton Outbreak (O'Connor 2002) (Chapter 5, see 2.4.1.3) and the Washington County Fair (Centers for Disease Control and Prevention 1999a) (Chapter 4, see 4.2.3.1), which are a significant factor in temporal trends, the low infectious dose may have made it possible for many people to be infected, despite the bacteria being transmitted through large water systems. In the Wishaw Outbreak (Cowden et al. 2001) (Chapter 3, see 2.4.1.1), another large outbreak, the fact that only a small quantity of bacteria had to pass between raw and cooked meats, and then survive reheating, may have contributed to the over 270 confirmed cases.

1.6.2 Reservoirs

Though it is a zoonotic bacterium, *E. coli* O157:H7 has not been shown to be strongly pathogenic in any species but humans. It is, however, posited that some VTEC serotypes may cause diarrhoea in young calves (Naylor et al. 2005) and a disease in greyhounds, cutaneous and renal glomerular vasculopathy, appears to be connected with *E. coli* O157:H7 (Moxley & Francis 1998).

E. coli O157:H7 is most commonly found in ruminants, particularly cattle (Renter and Sargeant, 2002) and sheep (Chapman et al., 1997). Though studies in farms and slaughterhouses have provided a wide range of estimates for the prevalence of the serotype in cattle, with a range of 0% to 36% (Naylor et al. 2005; Renter & Sargeant 2002), cattle are now thought to be the predominant source of *E. coli* O157 implicated in human disease (Naylor et al. 2005). In sheep, prevalence has been estimated at 31% (Kudva et al. 1996), and contact with ovine faecal material has been implicated in an outbreak involving at least 20 people (Howie et al. 2003). The bacteria can remain viable in bovine faeces for nearly two months and in water for up to a month, though counts decrease much more rapidly in slurry (Maule 1999), thus

contaminated faeces can present a considerable risk via direct or indirect contact (i.e. via water or food). A number of factors have been suggested to affect the shedding of the bacteria by cattle, including feed type, age, season and housing (Renter & Sargeant 2002). Additionally, it has been posited that a small minority of cattle – super shedders – may be responsible for most of the bacteria shed (Matthews et al. 2006b). Though the prevalence of the bacteria in sheep has thus far not been shown to be as high as in cattle (Naylor et al. 2005), one outbreak in Scotland was caused by direct handling of sheep faeces (Howie et al. 2003).

In addition to ruminants, *E. coli* O157 has also been isolated from deer (Sargeant et al., 1999), birds (Hancock et al., 1998), dogs (Trevena et al., 1996; Hancock et al., 1998), horses (Hancock et al., 1998), flies (Shere et al., 1998), pigeons (Shere et al., 1998), raccoons (Shere et al., 1998), opossums (Renter et al., 2003), goats (Bielaszewska et al., 1997; Milne et al., 1999), rats (Cizek et al., 1999), cats (Trevena et al., 1996), turkeys (Heuvelink et al., 2002) and pigs (Chapman et al., 1997; Mead and Griffin, 1998). Though the bacteria has been found in such a wide range of animals, evidence of animal to human transmission, direct or indirect, has been found only for cows (Synge et al., 1993; Shukla et al., 1995), deer (Rabatsky-Ehr et al., 2002), goats (Shukla et al., 1995), sheep (Licence et al., 2001; Howie et al., 2003), horses and dogs (Trevena et al., 1996). Such transmission is of great importance in human infection because of the extent of contact between humans and these animals in settings like zoos, farms and through handling and consumption of animal products such as cheese, milk, minced beef and lamb. These methods of transmission will be discussed in the next section.

1.6.3 Transmission

Research has shown that *E. coli* O157 infection can be transmitted via direct or indirect contact with infected animal or humans, faecal matter and/or by-products (Caprioli et al. 2005; Karch et al. 2005). This transmission may be foodborne, waterborne, person-to-person, environmental, via animal contact or any combination of the above. Temporal trends in mode of transmission will be explored in Chapters 2-5, compared across countries in Chapter 6 and mode of transmission will be

explored as a potential factor in the rate of secondary cases in Chapter 7. The next sections will provide an introduction to each mode of transmission category.

1.6.3.1 Foodborne transmission

Most outbreaks in the developed world have been linked to foodborne transmission of *E. coli* O157. Beef products, in particular, have been very frequently identified as a vehicle of transmission (Mead and Griffin, 1998), especially in the United States and Canada, where the vast majority of outbreaks, including the largest outbreaks, have involved minced (ground) beef (Riley et al., 1983; Wells et al., 1983; Bell et al., 1994; Feng, 1995; Todd, 2000). Beef has also been a factor in outbreaks in Scotland and England & Wales (Feng, 1995; Wall et al., 1996; Wachsmuth et al., 1997), though in recent years other modes of transmission have become increasingly important (SCIEH, 2000b; 2001; 2002; Locking et al., 2003; SCIEH, 2003).

Other bovine and meat products have also been implicated in *E. coli* O157 outbreaks, including venison (Wachsmuth et al., 1997), raw and pasteurized milk (Centers for Disease Control and Prevention 1999b; Centers for Disease Control and Prevention 2000; Centers for Disease Control and Prevention 2001a; Centers for Disease Control and Prevention 2002a; Chapman et al. 1993; Coia 1998a; Cowden 1997b; Keene et al. 1997; Upton & Coia 1994), yogurt (Coia, 1998), cheese (Cowden, 1997; Coia, 1998) and cheese curds (Centers for Disease Control and Prevention, 1999). Many infections have also been caused by cooked meats, which may be served in foods like meat pies, sandwiches, stews and salads (Griffin and Tauxe, 1991; Cowden et al., 2001).

Increasingly, however, *E. coli* O157 outbreaks are being caused by a much wider variety of food products (Coia, 1999). These include lettuce (Centers for Disease Control and Prevention, 1999; 2000), spinach (Centers for Disease Control and Prevention 2006h), alfalfa sprouts (Breuer et al. 2001), salads, coleslaw, fruit salad, grapes (Centers for Disease Control and Prevention, 2001) and fruit juices (Willshaw et al., 1997). The probable or confirmed source of infection for most of these foods is cross-contamination by faeces from cows or other animals that are introduced onto the fields via irrigation, wild animals and/or runoff (California Food Emergency Response Team 2007), or contamination during packaging (Myers et al. 2002;

Palumbo et al. 2004) and/or meal preparation (Proctor 2000a; Proctor 2000b). However, it is not always possible to determine the exact source of an infection; for instance in 2006 spinach outbreak, samples from local rivers, wild pigs known to have been in the spinach fields and cows from local farms all tested positive for the outbreak strain (California Food Emergency Response Team 2007), suggesting a number of pathways for contamination of the spinach including direct faecal contamination from the wild pigs and contamination of irrigation water by the cows or pigs.

The considerable potential for cross-contamination via faecal matter in crop fields was illustrated in a series of experiments in which *E. coli* O157 was shown to survive up to 22 days in sieved soil. Further experiments showing that the bacteria could survive up to 60 days on stainless steel, 5-7 days on anti-microbial impregnated plastic cutting boards at 22° C and at least two weeks at 4°C demonstrate the possibility for cross-contamination during food preparation. In addition, *E. coli* O157 is able survive in highly acidic conditions (Feng, 1995), the knowledge of which has emerged through discoveries of outbreaks spread by mayonnaise, apple cider (Besser et al., 1993; Mead and Griffin, 1998) and salami (Tilden et al., 1996; MacDonald et al., 2004).

1.6.3.2 Waterborne transmission

More incidences of *E. coli* O157 spread through drinking or recreational water supplies have been noted in recent years, in part an indication of the bacteria's hypothesised relatively low infective dose, as mentioned above (Feng, 1995; Mead and Griffin, 1998).

Contaminated drinking water, either from wells or municipal water supplies, has been the source of a number of outbreaks in the U.S. (Swerdlow et al., 1992; Centers for Disease Control and Prevention, 1999), Scotland (Licence et al., 2001), Canada (Health Canada, 2000), and England & Wales. Also implicated in the transmission of *E. coli* O157 infections are swimming in faecally contaminated water in paddling pools (Brewster et al. 1994; Cowden 1997b; Hildebrand et al. 1996), lakes/beaches (Ackman et al. 1997; Centers for Disease Control and Prevention 2001a; Centers for Disease Control and Prevention 2002c; Harrison & Kinra 2004; Paunio et al. 1999;

Samadpour et al. 2002b) and swimming pools (Centers for Disease Control and Prevention 1999b; Friedman et al. 1999; Verma et al. 2007). Contamination of natural water sources may be from animals (e.g. direct contact or runoff containing faecal material) or humans (faecal accidents or sewage), and pool contamination from the latter. The potential for waterborne transmission, even weeks after the initial contamination of the water is suggested by the finding that viable numbers of *E. coli* O157:H7 have been shown to survive in 18°C river water for up to 26 days (Maule, 1999).

1.6.3.3 Person to person transmission

Person to person/animal to person and environmental transmission have become increasingly important in recent years (Locking et al., 2003) (Steinmuller et al. 2006; Strachan et al. 2006). For instance, in Scotland, the percentage of outbreaks where the primary mode of transmission was to person-to-person contact went from 0% to 50% between 1998 and 2002. Trends in other countries as well as Scotland will be analyses in Chapters 2 to 5.

Person to person transmission may be the primary mode of transmission or spread an infection initially acquired from direct animal (David et al. 2004; Payne et al. 2003) or foodborne contact (Bell et al. 1994). It is more frequent in places of close contact, especially where there is potential for less than ideal sanitary conditions, in particular nursery schools/day-care centres (Centers for Disease Control and Prevention, 2002b; Health Canada, 2003) (Al-Jader et al. 1999; Allaby & Mayon-White 1995; Centers for Disease Control and Prevention 2002a; Galanis et al. 2003; Public Health Laboratory Service 2000a; Reida et al. 1994; Sugiyama et al. 2005) and residential institutions/nursing homes (Afza et al. 2006; Bisset et al. 1992; Coia 1998b; Coia 1999; Levine et al. 1991; Lior 1983) . Secondary transmission within households has also been noted and quantified, with estimates for the sporadic household transmission rate in Wales of 4 to 15%, and the presence of a person in the same household with diarrhoea very strongly associated with sporadic infection in the United States (Slutsker et al. 1998). Within households, younger children are the mostly likely both to transmit and to be infected by *E. coli* O157 (Parry & Salmon 1998), and it has been postulated that person to person transmission rates in settings

where young children are present may be elevated because young children have been shown to shed bacteria in their faeces for extended periods of time (Belongia et al. 1993; Karch et al. 1995; Shah et al. 1996; Swerdlow & Griffin 1997). Incidents in hospitals and institutions demonstrate the possibility of nosocomial transmission – transmission in a hospital setting due to direct or indirect contact between patients and items (dressings, surgical instruments etc.) used in their care (Coia et al., 1995), but may also caused by contaminated food (Health Canada, 1997; 2004).

1.6.3.4 Animal contact

Infection due to direct animal contact is a significant mode of transmission (Steinmuller et al. 2006; Trevena et al. 1999) in terms of infections in young children. The potential for direct animal to human transmission of *E. coli* O157 was first recognized in a case where the infected child had contact with animals on the family farm (Synge et al., 1993), and subsequent cases, both symptomatic and non-symptomatic, have been noted in farming families & workers (Health Canada, 1998; Trevena et al., 1999). However, the majority of cases, at least the symptomatic ones detailed in the published literature or included in the data sets provided for the studies in Chapters 2-6, have been at petting zoos or open farms. It has been suggested that this may occur because people who have frequent contact with animals build up immunity to the bacteria, leaving infrequent visitors the most likely to have severe enough symptoms to be end up being reported to a surveillance agency (Wilson et al. 1996).

The first known petting farm outbreak in Canada was in 1999 and involved contact with sheep & goats (Middlesex-London Health Unit, 1999), while the first reported outbreaks in the United States took place a year later in two different states and involved calves and young cattle (Gage et al., 2001). Petting farm linked outbreaks have also occurred in countries such as England (Shukla et al., 1995; Milne et al., 1999; Trevena et al., 1999), Denmark (Anonymous 2004) and Scotland (Strachan et al. 2006). Such outbreaks are clearly of particular concern because they frequently involve young children for whom complications are more common. (Steinmuller et al. 2006)

1.6.3.5 Environmental transmission

Environmental transmission, often via contact with animal faeces on the ground has been postulated to account for an increasing frequency of cases (Strachan et al. 2006). Recent environmentally related outbreaks in Scotland have included an outbreak at a scout camp, where scouts camped on field recently occupied by infected sheep (SCIEH, 2000a; Howie et al., 2003) and an outbreak where persons had exposure to a potentially faecally contaminated stream on a campsite (Strachan et al. 2006). Other outbreaks with environmental links have occurred in Austria (Grif et al. 2005), the United States (Centers for Disease Control and Prevention 2006e; Varma et al. 2003) and England and Wales (Chapman 2000). Recent studies suggest that environmental transmission may be a significant factor in *E. coli* O157 outbreaks in Scotland (Strachan et al. 2006) and the United States (Steinmuller et al. 2006), and contact or likely contact with animal excreta has been shown as a statistically significant risk factor for sporadic infection (Locking et al. 2001). A strongly significant association between sporadic infection (see 1.6.6.1) and contact with farming environments has also been demonstrated in England (O'Brien et al. 2001a).

1.6.4 Seasonality

E coli O157:H7 infections also show seasonality (Bach et al. 2002), with the highest rates of infection during the summer months of July-September in the northern hemisphere (Mead and Griffin, 1998). This pattern has been noted in Canada (Michel et al., 1999), Wales (Chalmers et al., 1999), England & Wales (Wall et al., 1996; Willshaw et al., 2001) and Scotland (Douglas and Kurien, 1997), and it is postulated to result from better growing conditions for bacteria in the warmer months of the year as well as higher number of people participating in activities that put them at risk for contact with *E. coli* O157 in the summer, such as barbeques, visits to farms and parks. The existence of a seasonal pattern is of importance because it affects the choice of statistical techniques and data format (yearly data points rather than monthly or daily data points) for the analysis of temporal trends (see Section 1.8) – in Chapters 3 to 6.

1.6.5 Age and sex specific rates

The highest rates of infection are in young children (Parry & Palmer 2005), with peak infection rates in most countries for children under the age of five (Centers for Disease Control and Prevention 1995b; Centers for Disease Control and Prevention 1996a; Centers for Disease Control and Prevention 1997; Centers for Disease Control and Prevention 1998; Centers for Disease Control and Prevention 2001b; Centers for Disease Control and Prevention 2002b; Centers for Disease Control and Prevention 2003b; Centers for Disease Control and Prevention 2005e; Centers for Disease Control and Prevention 2007g; Public Health Agency of Canada 2006). The differences in infection rates between age groups are likely linked to pathophysiological factors (see 1.5.2), the higher risk for children to acquire infections in the home (see 1.6.3.3) and the association of children with locations such as petting zoos and swimming pools (see 1.6.3.4). The links between age and infection rates will be examined in more detail in Chapter 7, when the median age of cases in outbreaks is analysed as a potential factor in the rate of secondary cases. There is also male-female pattern in infections, with women accounting for more than half of infections in Wales (Chalmers et al., 1999), Europe (Fisher, 2002), Canada (Waters et al., 1994; Michel et al., 1998), United States (Centers for Disease Control and Prevention, 2002; Locking et al., 2003) and Scotland. However, though more women are infected overall, in children rates of infection tend to be higher for boys: outside of Canada (Waters et al., 1994; Michel et al., 1998) and England & Wales (Thomas et al., 1996), where girls catch up around the age of 15, boys tend to make up the majority of infected cases until about the age of five. The higher rates of infection in post-pubertal women is postulated to result from the fact that women are more likely to be involved in activities that put them at risk for infection such as taking care of young children and preparing food. However, for reasons of maintaining patient confidentiality, sex data was not available for most of the studies in this thesis, so differences in infection rates between men and women will not be considered in any of the analyses.

1.6.6 Outbreaks vs. sporadic cases

E. coli O157 cases are generally defined as being either a sporadic case or part of an outbreak (Locking et al. 2003a). The definition of an outbreak varies between

countries, with some countries including Scotland and England & Wales reporting as "general outbreaks" those clusters of cases which involve "members of more than one household or institution" (Locking et al. 2004), and other countries using different definitions for outbreaks, the definition sometimes varying by surveillance system or region (Tinga et al. 2006). For instance, a recent publication on the epidemiology of *E. coli* O157 cases in Japan defines an outbreak as having more than 11 cases in a certain area and time frame (Sakuma et al. 2006) and in the United States an outbreak has been defined a cluster of at least two infections with a "common epidemiologic exposure" (Rangel et al. 2005). The implications of these differences in definitions will be discussed in Chapter 6 when some of the temporal trends are compared between countries.

1.6.6.1 Sporadic Cases

Sporadic cases make up the majority of reported *E. coli* O157 infections, comprising from 73% to 83% of annual reported confirmed cases between 2002 and 2004 in Scotland, and accounting for as many as 90% of reported annual confirmed cases in the United States (official CDC data set, Elizabeth Blanton) and 96% of reported cases in Ontario between 1990 and 1994. However, the definition of a sporadic case does vary between countries (Locking et al. 2003a)(personal communication, Thai An Nguyen, CDC). This will be discussed in detail in Chapter 6 and in Chapter 8, with regards to between-country comparison of data and defining an *E. coli* O157 event. Briefly, cases within single household clusters in Scotland are considered to be sporadic, whilst in countries like Canada and the United States these clusters are considered to be outbreaks (and thus the cases are outbreak cases). Thus the counts of sporadic cases in Scotland include cases which have epidemiological links to other cases and thus would not be defined as sporadic cases in Canada or the United States.

1.6.6.2 Outbreaks

The first reported outbreaks were in 1982, when clusters of cases in Oregon and Michigan in the United States were linked to consumption of minced beef. In the twenty five years since, there have been outbreaks across the world, including the United States (Rangel et al. 2005), Canada (Woodward et al. 2002), Scotland (Cowden 1997b), Japan (Michino et al. 1999), Denmark (Jensen et al. 2006), Sweden

(Sartz et al. 2007), Finland (Paunio et al. 1999), France (Espie et al. 2006), Germany (Conedera et al. 2007), Ireland (Anonymous 2005), Australia (McCall et al. 1996), England & Wales (Willshaw et al. 1997) and the Canary Islands (Pebody et al. 1999). The majority of reported outbreaks have been small, with the geometric mean number of ill cases in Scotland, Canada and the United States since 1996 (1998 for the U. S.) all under 9 cases (see Chapter 6 for more details). However, a number of large outbreaks – some of which are influential in the analyses conducted in Chapters 2-6 – have occurred.

With more than 6000 cases and as many as 1682 confirmed cases (National Institute of Infectious Diseases 1996), the 1996 Sakai City outbreak in Japan remains the largest reported outbreak. The cases were linked to contaminated radish sprouts served as part of school meals across the city (Michino et al. 1999). Other large outbreaks include the 2000 Walkerton Outbreak in Canada (1346 ill cases, see 2.4.1.3), caused by contamination of a public water supply (O'Connor 2002), the 1996 Wishaw Outbreak in Scotland, in which cross-contamination between cooked and uncooked meats in a butchers shop sickened over 500 people (279 laboratory confirmed, see 2.4.1.1) (Cowden et al. 2001). An outbreak caused by contaminated drinking water at the 1999 Washington County Fair, New York, United States resulted in more than 700 illnesses (Centers for Disease Control and Prevention 1999a)(see 4.2.3.1), and the 1993 multi-state outbreak in the western United States involved 501 confirmed cases (Bell et al. 1994). Outbreaks will be examined in detail in further chapters of this thesis, with Chapters 2 – 5 examining trends in single countries, Chapter 6 comparing across countries and Chapter 7 examining cases within outbreaks.

1.6.7 Primary cases vs. secondary cases

E. coli O157 cases can be primary or secondary. Primary cases are cases which result from direct contact or exposure to the initial or point source of an outbreak, whilst secondary cases are generally acquired through direct or indirect spread (e.g. through water or a contaminated surface) from another infected person (Locking et al. 2003a). Normally, the terms primary and secondary are only used in the context of outbreaks. However, in countries such as Scotland and England & Wales where cases in single household clusters are considered to be sporadic, some sporadic cases

may also be considered to be secondary (e.g. household contacts) (Locking et al. 2003a; Parry & Salmon 1998).

Definitions of secondary cases vary between countries (Locking et al. 2003a; Parry & Salmon 1998), and even between outbreaks as the number of cases defined as secondary may depend on the exact definition established in an outbreak investigation. For instance in the 1993 minced-beef outbreak in Washington, people were defined as having secondary infection if they "became ill within 10 days of a household or other close contact with another patient and had not eaten at chain A during that time" (Bell et al. 1994), whereas in a 1986 outbreak also related to minced beef the definition was more restrictive because, "a secondary case was defined as culture-confirmed *E. coli* O157:H7 illness in a household contact of a primary case which began at least 48hrs after the onset of illness in the primary case" (Ostroff et al. 1990).

This lack of a common definition may be one factor in the lack of comprehensive estimates of the magnitude of secondary spread. Papers suggest that secondary spread is common (Armstrong et al. 1996; Coia 1998a), with data from single outbreaks suggesting that secondary cases may account for up to 40% of cases and most outbreaks having a secondary rate no higher than 10% (Armstrong et al. 1996; Coia et al. 1996). Secondary spread has also been documented in households, with children the most likely to infect or be infected (Parry & Salmon 1998), and has been shown to occur via asymptomatic cases (Marsh et al. 1992). Such cases are of concern because, in particular, as discussed above, children are at a higher risk for severe complications. Yet, no comprehensive multi-country study of secondary cases rates has been published. The study of secondary case rates and characteristics in outbreaks in Chapter 7 of this thesis is intended to fill this gap in the literature.

1.6.8 Asymptomatic cases

Asymptomatic infection has been shown to occur in outbreaks, as well as in close contacts of infected persons (Armstrong et al. 1996). Detection of asymptomatic cases usually occurs in an outbreak investigation when either contacts of symptomatic cases or persons with possible exposure to the source of contamination are screened (Akashi et al. 1994). The rate of asymptomatic cases reported for single

outbreaks and in single-country population-based studies would suggest that asymptomatic cases may not be the exception. For instance, some examples of the rate of asymptomatic cases include 17% of confirmed cases in a 1990 Japanese outbreak at a nursery school (Akashi et al. 1994), up to 12.5% of confirmed cases in the Wishaw Outbreak (Cowden et al. 2001) and 16% of cases in a Minnesota (USA) nursery school outbreak (Belongia et al. 1993). However, investigational techniques may affect the number of asymptomatic cases that are reported – e.g. broad based screening like that done in some Japanese outbreaks (Akashi et al. 1994) may pick up a higher percentage of asymptomatic cases.

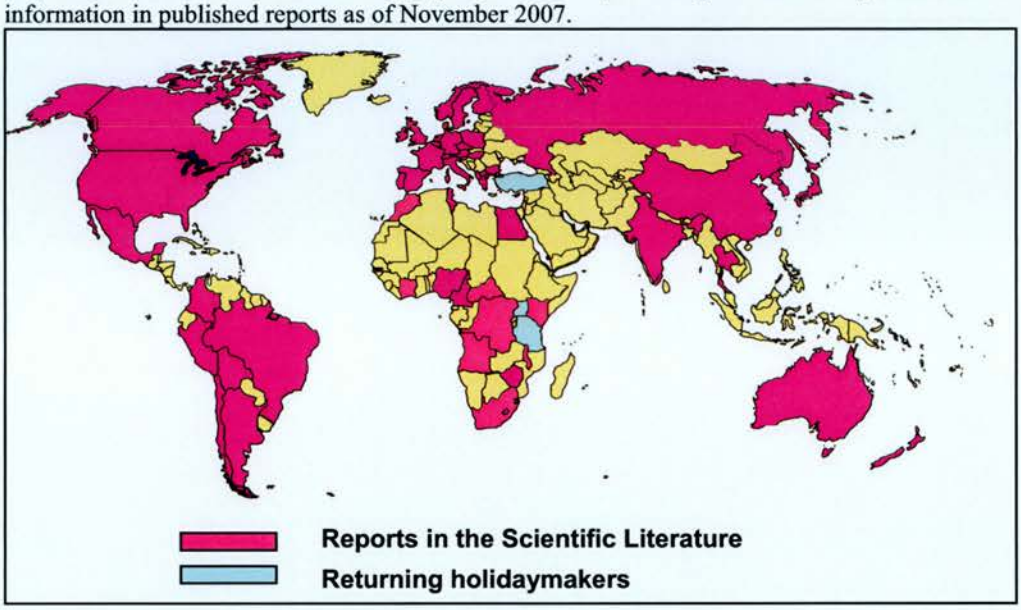
Few estimates of the overall prevalence of asymptomatic cases in any one country exist, but between 2002 and 2004 from 6.5% to 10% of reported and laboratory confirmed cases in Scotland were asymptomatic. More than half of these cases were not linked to a reported outbreak, but rather to sporadic cases within the same household (Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a). Approximately 10% of serotype O157 cases reported by EnterNet countries (EU countries, Australia and Canada) in 2005 were asymptomatic (European Commission 2007). These reported rates of asymptomatic infection of up to 17% and the high rates of under-reporting of infections estimated in countries such as England & Wales, United States and Canada (Adak et al. 2002; Hedberg et al. 1997; Michel et al. 2000) suggest at a much higher prevalence of infection than is reported. The issue of true prevalence will be explored in Chapter 8 where a framework for estimation of prevalence will be presented.

1.6.9 Geographic distribution

E. coli O157 infection has been primarily reported in the developed world, with the highest rates of reported annual infection during the last two decades having been seen in Scotland, Canada and the United States (Mead & Griffin 1998; Sakuma et al. 2006), the three countries which are covered in depth in this thesis, as well as Japan. Since 1982, infections have been reported in more than 30 countries in every continent other than Antarctica, with confirmed infections in the countries indicated in pink on Figure 1.3. However, infections may have occurred in other countries

since it has been postulated that infections are more likely to go undetected in countries without developed *E. coli* O157 surveillance systems (Ackers et al. 1998).

Figure 1.3: Countries where *E. coli* O157 infection has been reported
The countries where *E. coli* O157 infection has been reported, adapted from a figure courtesy of Prof. Bill Reilly. As not all cases are detected and reported in the published literature, infections may have occurred in countries not highlighted on this map. The figure has been updated to reflect information in published reports as of November 2007.



However, infections may have occurred in other countries since it has been postulated that infections are more likely to go undetected in countries without developed *E. coli* O157 surveillance systems (Ackers et al. 1998). Additionally, studies have indicated that non-O157 VTEC/EHEC infectious are more common in places including Australia (Goldwater & Bettelheim 1995), South America and parts of continental Europe (Bach et al. 2002; Karch et al. 2005).

In 2005, eight continental European countries contributing to the Enter-Net database (which will be discussed below) reported more than 50% of infections dues to non-O157 serotypes: Denmark, Germany, Luxembourg, Norway, Slovenia, Italy, Austria and Switzerland (European Commission 2007). As a result, these countries and regions are less likely to be included in the analyses in this thesis.

1.7 Surveillance

1.7.1 Methods of surveillance

Organized surveillance systems for *E. coli* O157 have been established in many of the countries where *E. coli* O157 cases have occurred, with some of the most

comprehensive programs in the three countries covered in this thesis: Scotland (Locking et al. 2003a), the United States (Centers for Disease Control and Prevention 2003a) and Canada (Public Health Agency of Canada 2007). Surveillance systems can be either active or passive, or involve components of both types (Centers for Disease Control and Prevention 2006f). In active surveillance, data – usually on cases – is actively sought from health or laboratory professionals, while passive surveillance systems rely on data being reported by laboratories, physicians or other public health personnel.

The reporting systems for the three above mentioned countries will be detailed in the corresponding chapters (2 – 5), but a general summary of surveillance will be provided below. In countries such as Scotland, the United States and Canada, reporting of all cases to the national infectious disease agency (e.g. HPS, CDC, PHAC etc.) is mandatory, though the regulations within political sub regions (e.g. states or provinces) may vary. For instance, reporting was mandatory at the state level (e.g. to the state public health agency) in only 48 U.S. states by 2000 (Rangel et al. 2005). In Scotland, cases are reported to HPS (Locking et al. 2003c), while in the United States all laboratory confirmed and probable cases are reported through the National Notifiable Disease Surveillance System (NNDSS) (Centers for Disease Control and Prevention 2007g) and in Canada cases are reported to the National Notifiable Diseases program (NND) (Public Health Agency of Canada 2006). This passive surveillance may also be supplemented by active surveillance programs such as FoodNet in the United States. Established in 1995, FoodNet has expanded to cover 10 surveillance areas each of which ranges from a few counties to an entire state, and overall comprise about 15% of the US population (Centers for Disease Control and Prevention 2007f). The program involves active surveillance conducted via regular contact with laboratories to gather data on confirmed cases of infectious foodborne illnesses as well as population, physician and laboratory surveys to increase the knowledge of disease prevalence. In Scotland, the enhanced surveillance program, started in 1999, includes active components including an attempted case by case follow-up of all reported cases (Locking et al. 2003a).

Surveillance for outbreaks differs between countries and is usually done a passive basis. Outbreak cases are reported via paper or internet based systems from local/regional public health offices and laboratories to the national surveillance organizations (Health Protection Scotland 2007f), with PFGE types of isolates reported to national laboratories compared via PulseNet in Canada and the United States. In Scotland there is a single reporting form used to capture reporting information, but in other countries the form and thus the type and quality of the outbreak data captured, can differ by state/province/health board (personal communication, Carole Tinga, PHAC and Thai An Nguyen, CDC). The issues of data comparability will be discussed in each country specific chapter (Chapters 2 – 5) and in detail when the trends between countries are compared (Chapter 6).

Within Scotland, outbreaks are reported to Health Protection Scotland (Health Protection Scotland 2007f), while in Canada the Foodborne, Waterborne and Zoonotic Infections Division at the Public Health Agency of Canada (PHAC) maintains a database of outbreaks (Tinga et al. 2006) and the National Microbiology Laboratory issues yearly reports on all outbreaks for which samples are submitted to the laboratory (National Laboratory for Enteric Pathogens 2006). In the United States, outbreaks are reported to the CDC via electronic or paper records (personal communication, Thai An Nguyen, CDC). Until 2006, foodborne and non-foodborne outbreaks were reported via separate forms and records maintained in separate databases (Centers for Disease Control and Prevention 2003a; Centers for Disease Control and Prevention 2004a). The fact that the forms have been updated at least five times since 1989 raises the issue of data comparability over time. Changing forms and reporting methods can affect the quality and quantity of data, and for the United States in particular, such effects are a significant issue and will be addressed in detail in Chapter 4.

As *E. coli* O157 infections have been reported in more than 20 countries (see 1.6.9), and a number of outbreaks have crossed international borders (Pebody et al. 1999), multi-national surveillance networks have been created. In terms of the countries covered in this thesis, the major network has been EnterNet, which covers 36 countries, primarily from Europe, but also from North America, Africa, Australia and

Asia (European Commission 2007). As of the 2005 report, data on *E. coli* O157 had been received from 26 countries for all or part of the period of 2000 to 2005, though some countries provide only case counts, while others provide more detailed information such as outbreak information and a breakdown of cases by factors such as serotype and phage type. The network provides an important opportunity for future comparison of national case and outbreak trends, as well as ability to coordinate approaches towards multi-country outbreaks. However, the data provided in the reports has thus far been descriptive, with no statistical analysis of trends within or between countries (European Commission 2007). EnterNet has been discontinued as of 2007, but data collection and analysis will be continued under the auspices of the new European Centers for Disease Control (ECDC). Multi-country collaboration also exists via the PulseNet groups which allow comparison of PFGE types from isolates, linking cases together within and between countries (CDC 2007).

For all passive surveillance – outbreak or sporadic – incomplete reporting is an important issue, in terms of the number of cases or outbreaks that may not be reported to national surveillance agencies. The causes of, and research on, under-reporting prior to the level of surveillance will be discussed in Section 1.9, but even when cases are identified and laboratory confirmed, the data provided to the national surveillance agency may not be complete. In the United States, not all states reported case counts to the CDC until 2002 (Centers for Disease Control and Prevention 2004b), and in Canada, not all provinces or territories have provided data each year to the NND and/or the outbreak database (Public Health Agency of Canada 2006; Sockett et al. 2006). Thus, any analysis of data from surveillance must be interpreted based on knowledge regarding omissions in the data sets.

1.7.2 Surveillance data – publication

Routine surveillance data in most countries, including the three analysed in Chapters 2-6, are published in reports or publications issued by the national surveillance agencies. The depth of detail in the reports varies by country, with all countries releasing at least a yearly listing of total laboratory isolates or case numbers. In Scotland, quarterly and yearly listings of outbreaks as well as an in depth summary of findings from enhanced surveillance are published in HPS Weekly(currently

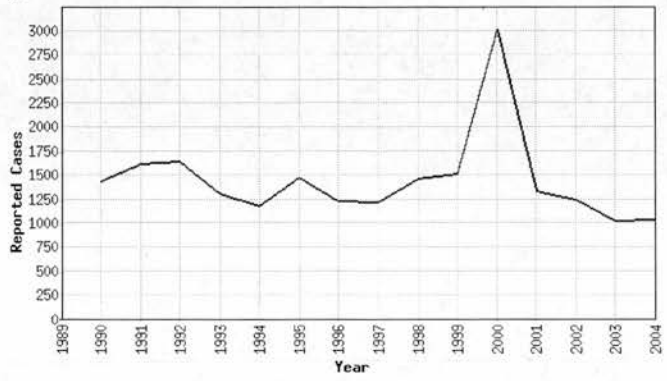
available for years 2002 to 2004) (Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a; SCIEH 1997b; SCIEH 2000a; SCIEH 2003b; SCIEH 2003d; Smith-Palmer et al. 2005). Outbreak listings include health board region, suspected mode of transmission and case number (ill and confirmed), while the enhanced surveillance reports include statistics with graphs or charts on the breakdown of cases by such factors as serotype, asymptomatic/symptomatic, outbreak/sporadic, and secondary/primary.

For the United States, counts of reported cases are issued weekly in the Morbidity & Mortality Weekly Report (MMWR) and summarized in more detail in yearly MMWR reports on notifiable diseases (Centers for Disease Control and Prevention 2007g). Information on outbreaks, other than MMWR reports on specific outbreaks, does not appear to be routinely released, but listings of all foodborne outbreaks and *E. coli* O157 outbreaks between 1998 and 2002, have been posted on the CDC website (Centers for Disease Control and Prevention 2007e). In Canada, data on case numbers and rates, by sex, age, province and year can be accessed in graphical and table format via Notifiable Disease On-Line (Public Health Agency of Canada 2006). The National Microbiology Laboratory also issues yearly reports covering isolates (sporadic and outbreak) reported to the National Enteric Surveillance Program (National Laboratory for Enteric Pathogens 2002). Monthly data on reported cases is also released in the Notifiable Disease Monthly Report (Public Health Agency of Canada 2005). Temporal data is generally presented in the form of graphs, with examples given in Figure 1.4 a-c. These tables show trends in total cases or rates. Figure 1.4a shows the total number of VTEC cases, which includes non-O157 *E. coli,* in Canada, while Figure 1.4b includes plot lines for *E. coli* O157 rates per 100,000 of laboratory confirmed *E. coli* O157 in the countries that make up the United Kingdom.

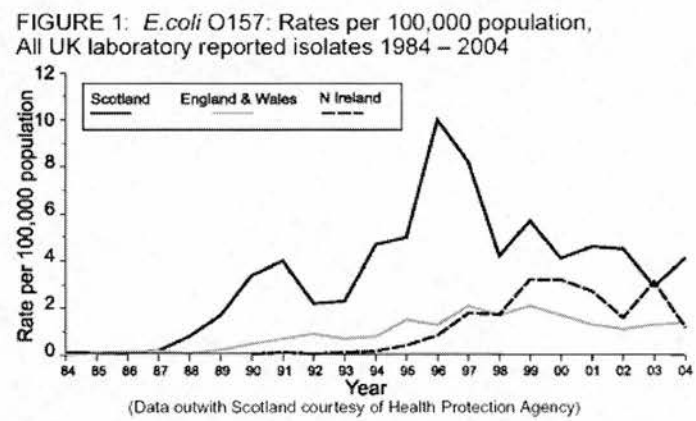
Figure 1.4c shows the number of cases reported to the two reporting systems in the United States (PHLIS and NNDSS) – NNDSS numbers are higher because they include both confirmed and probable cases (see Chapter 2 for details on these reporting systems). These figures are all purely descriptive, with no statistical analysis of trends included in the accompanying reports.

Figure 1.4 a-c: Examples of plots illustrating temporal trends in overall *E. coli* case numbers from surveillance reports in (a) Canada, (b) Scotland and (c) United States.
Examples of plots showing temporal trends in case numbers. Plot a is constructed using the Notifiable Diseases On-Line (Public Health Agency of Canada 2006), and shows the total number of VTEC cases in Canada each year. Plot B is from the 2004 enhanced surveillance report issued by HPS (Locking et al. 2006a). Plot C is from a 2002 report on Shiga-toxin producing *E. coli* in the United States (Centers for Disease Control and Prevention 2003a), with the PHLIS data including only laboratory confirmed isolates and the NNDSS data including confirmed and probable cases.

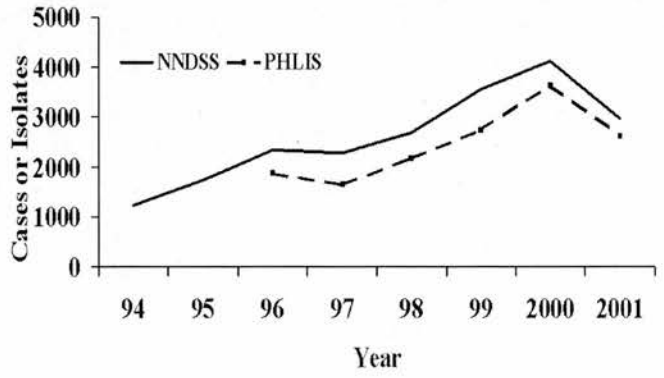
a



b



c



1.8 Methods of analysing temporal trends

Analyses of temporal trends in *E. coli* O157 have been limited to descriptive studies and analyses of seasonal trends. However, statistical analyses of trends in a wide

range of chronic and infectious conditions have been conducted. These include cancer (Agha et al. 2006; Conway et al. 2006; Maule et al. 2006) injury rates (Agnusdei et al. 1993; Brenner et al. 1999), *Campylobacter* (Miller et al. 2004b; Sandberg et al. 2006; van de Giessen et al. 2006), *Salmonella* (Guerin et al. 2005a; Guerin et al. 2005b; Khakhria et al. 1997; Marcus et al. 2004b; Menzies et al. 1994; Olsen et al. 2001; Smith-Palmer et al. 2003) and *Acinetobacter* (McDonald et al. 1999).

Temporal trends or time trend data (Maule et al. 2006) can be analysed using multi-year (Tleyjeh et al. 2005), yearly (Sandberg et al. 2006), monthly (Miller et al. 2004b; van de Giessen et al. 2006), weekly or even daily data points. Since surveillance data is commonly recorded in weekly or monthly intervals (Centers for Disease Control and Prevention 1995b; Public Health Agency of Canada 2005), these data points are often used for analyses infectious disease trends. However, while sub-yearly data permit the analysis of seasonal trends (Douglas & Kurien 1997; McDonald et al. 1999; Miller et al. 2004b), the effects of seasonality must be identified and adjusted for in order for overall trends to be analysed. One method for analysing both seasonal and yearly trends is to include terms for both year and month (or day) in a model (Guerin et al. 2005b).

The most common method for analysing temporal trends is Poisson regression (Agha et al. 2006; EURODIAB ACE Study Group 2000; Kamper-Jorgensen et al. 2006; Sandberg et al. 2006; Tleyjeh et al. 2005), with the time (year, month etc) as a continuous, categorical or offset variable. However, simple linear regression (Van der Heyden et al. 2000), Bayesian methods (Maule et al. 2006), time series analysis (Kuhn et al. 1994), Chi-square tests of trend (Joe & Kaplan 2002; Marcus et al. 2004b) and least squares regression (Agnusdei et al. 1993; Olsen et al. 2001) have also been used in some studies. Trends have been compared using analysis of covariance (Joe & Kaplan 2002) or by the inclusion of an interaction term in a regression (EURODIAB ACE Study Group 2000). Poisson regression appears to be appropriate for short time periods, with analyses done on data sets with as few as six data points. Such applicability is important when considering the data sets for Chapters 2-6, which contain only 8 or 9 data points.

These methods, in particular those shown to be applicable for other infectious diseases and for shorter time periods, will be discussed in more depth in Chapter 2 (section 2.3) when the process for selecting the appropriate statistical method(s) for analysing *E. coli* O157 trends in Chapters 3-6 are presented.

1.9 Methods of assessing under-reporting

In this section, an introduction to the techniques used to analyse under-reporting of diseases and conditions, and the under-reporting of infectious diseases is provided. A brief introduction to the under-reporting of *E. coli* O157 will be given, with a more in depth analysis provided in the introduction to Chapter 8. Existing methods of estimating of prevalence will also be mentioned as background for the framework presented in Chapter 8.

1.9.1 Under-reporting in infectious diseases

Underreporting of counts or cases occurs in a wide range of circumstances in infectious diseases. For instance, the rate of human *Schistomiasis mansoni* infections is underestimated by more than 50% by single faecal egg counts (de Vlas & Gryseels 1992). As many as 25% of Legionellosis cases in Italy were not reported to the national register (Rota et al. 2007). In Canada, the United States and United Kingdom, a number of studies have suggested that a large percentage of foodborne illnesses are never reported (Angulo et al. 1998; Cumberland et al. 2003; Handysides 1999; Imhoff et al. 2004; MacDougall et al. 2007; Voetsch et al. 2004b; Wheeler et al. 1999). One study in the United States indicated that only 21% of persons with acute diarrhoeal illnesses sought medical care, and of those persons just 16% were asked for a stool sample (Imhoff et al. 2004). The rates of medical care uptake were approximately 10% in England & Wales (Cumberland et al. 2003) and British Columbia, Canada (MacDougall et al. 2007).

In Canada the National Studies of Acute Gastrointestinal Illness (NSAGI), which involved population, laboratory and reporting practice studies (MacDougall et al. 2007), have provided a great deal of data to assist in estimation of illness prevalence. Based on this data and other information, it has been estimated that *Campylobacter*, which has the lowest rate of cases with bloody diarrhoea, has the lowest rate of

reporting at 2% to 4%, followed by *Salmonella* (3 to 8%) and *E. coli* O157 (2 to 10%) (Thomas et al. 2006). The Infectious Intestinal Disease (IID) study in England and Wales was conducted to determine foodborne illness incidence rates in the community and in persons who report to general practice with symptoms of infectious intestinal illnesses (Wheeler et al. 1999). It included both community and general practice level surveys (Cumberland et al. 2003). Data from this study, together with additional information on hospital stays and deaths, was used to estimate the burden of foodborne diseases in 1992, 1995 and 2000. Again, the reporting rate of *Campylobacter* (~10%) and *Salmonella* (~26%) was estimated to be much lower than that of *E. coli* O157 (~50%) (Adak et al. 2002). The lowest rates of reporting were found for viruses and bacteria such as *Aeromonas* spp. Similar results were seen in CDC study based on data from a number of passive and active surveillance systems in the United States (Mead et al. 1999). Where data was not available, the major factor taken into consideration in estimating under-reporting was the general severity of symptoms. *Salmonella* and *Campylobacter* were assumed to have a higher rate of under-reporting because infected persons were less likely to have bloody diarrhoea than those infected with *E. coli* O157. This factor was hinted at in the correlation between rate of bloody diarrhoea and reporting rate found in the Canadian study mentioned above, and is likely to have influenced the results of the IID study in England & Wales. Though *E. coli* O157 does not have the highest rate of under-reporting or case numbers, it has higher rate of hospitalisation and case fatality that most other infectious intestinal organisms. Thus, despite the lower case counts, it remains an important public health issue.

1.9.2 Using reporting triangles to illustrate under-reporting

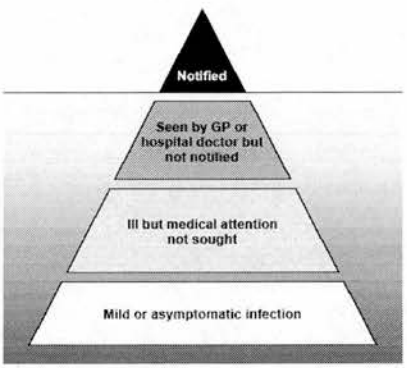
Cases of infectious intestinal illness are only picked up by surveillance programs if the infected persons seek medical care, samples are taken, tested and confirmed by a laboratory, and the positive results are reported to the appropriate surveillance agency. This incremental process is often captured in 'reporting triangles', also referred to commonly as "prevalence of illness pyramids" (Angulo et al. 1998; Griffin 1998) or "under-reporting pyramids" (MacDougall et al. 2007), which are used to illustrate the steps required for a case to be reported. By providing a clear

depiction of the process, reporting triangles can suggest where cases might be 'lost' to the reporting system. They also provide a basis for more accurately estimating both under-ascertainment and true prevalence of infections. In Figure 1.5 a, b and c, examples of reporting triangles for the surveillance of all infectious intestinal diseases in England and Wales (Handysides 1999) and in the United States (FoodNet Working Group 1997) are presented, demonstrating the attrition of cases between infection and reporting to the national surveillance body.

Figure 1.5 a-c: Examples of reporting triangles
Examples of reporting trials from (a) England and Wales, (b) United States and (c) Canada

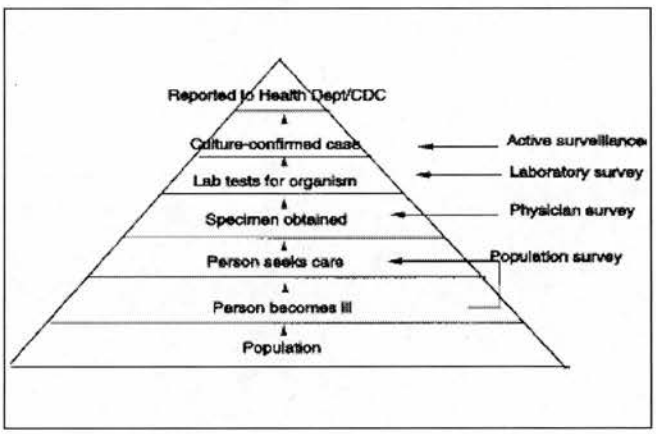
a

Figure 2 Notified cases represent only a proportion of the morbidity caused by food poisoning



Wall et al (1996)

b



Angulo et al (1998)

c

Fig. 1. Under-reporting pyramid for gastrointestinal illness in British Columbia. Overall under-reporting was characterized by estimating the proportion of cases that moved up through each of eight sequential tiers of reporting, conditional on reaching the previous tier.

MacDougall et al (2007)

In the first triangle, the sections represent the proportions of the population for which the case does not make it to the next level - i.e. the bottom section are those persons who have mild or no symptoms, and so thus would not ever be seen by a doctor. In Figure 1.5b, the sections of the triangle indicate the steps between illness and reporting, as well as indicating which type of surveys would measure the percentage of cases that progressed to each step. The final triangle is the most detailed, indicating the actual ratios of cases (as compared to every one reported case) that progress to each step.

Reporting triangles for *E. coli* O157 have been constructed, and will be discussed in Chapter 8. These triangles, however, present only one method of looking at prevalence.

1.9.3 Estimating prevalence

In addition to the studies detailed in Sections 1.9.1 and 1.9.2, which have estimated true prevalence of infections using data primary from surveys and records, there have been other studies which have used a variety of methods to estimate prevalence. A recent study estimated the prevalence of Human Papillomavirus in the United States based on data from studies of specific population groups (Weller & Stanberry 2007). Community studies with additional modelling for uncertainty were used in the estimation of prevalence for a wide range of conditions in South Africa (Bradshaw et al. 2007; Groenewald et al. 2007; Joubert et al. 2007; Norman et al. 2007; Schneider et al. 2007a; Schneider et al. 2007b). Italian researchers estimated Legionellosis prevalence using capture-recapture methods, which are based on counts of data from two different points (Rota et al. 2007) in times. Black and Craig used a Bayesian approach to estimate prevalence of *Strongyloides* infection based on results from two diagnostic tests (Black & Craig 2002). For calculating *Schistosoma mansoni* egg counts in faecal samples, de Vlas and Gryseels employed a stochastic model which incorporated such factors as worm distribution and variability of egg counts (de Vlas & Gryseels 1992). However, of most interest, given the framework proposed in Chapter 8, is the model presented for the estimation of bovine spongiform encephalopathy in cattle herds (Prattley et al. 2007). The model is based on surveillance and demographic data, but one main assumption is that the number of

cattle which are infected are binomial distributed. The framework in Chapter 8 also involves binomial distribution. However, whilst these studies all estimate prevalence for single cases, no study could be found which estimated prevalence based on counts of events (cases or outbreaks). Thus the study in Chapter 8 provides a potentially unique and different framework for estimating prevalence.

1.10 Conclusion

In this chapter, an introduction to *E. coli* and *E. coli* O157 has been presented, focusing on the history, microbiology, epidemiology and surveillance of the *E. coli* O157. This information is intended to provide a background to the studies in the next 7 chapters of this thesis. Temporal trends in *E. coli* O157 cases and outbreaks will be compared between Scotland, the United States and Canada, but before a comparison can be done, the trends in each country must be modelled individually. The first country for which trends will be modelled is Scotland.

Chapter 2 -- Comparing temporal trends in *E. coli* O157 between countries

2.1 Introduction

As discussed in the prospectus, the central aim of this thesis is to compare temporal trends in *E. coli* O157 outbreaks and cases between countries. Linear trends over time in such diseases such as cancer (Venzon & Moolgavkar 1984), gonorrhoea (Van der Heyden et al. 2000) and childhood diabetes (EURODIAB ACE Study Group 2000) have been compared, but to-date, there has been little direct comparison, descriptive or statistical, of trends in *E. coli* O157 between countries. In 1994, as the result of an exchange between Scotland and the Canadian province of Alberta, Waters and colleagues published a broad descriptive overview of *E. coli* O157 in Scotland and Alberta from 1987 to 1991 (Waters et al. 1994). Cases were described in terms of variables such as rate of infection by year, month and age, and suspected source of infection, but comparisons were made on a purely descriptive basis with the authors highlighting a number of similarities between the two countries. Since then there has been little published work on comparisons, though the establishment of the EnterNet *Salmonella* and *E. coli* O157 surveillance network, primarily involving European Union nations, has allowed for the creation of an international database of cases. Though no statistical comparisons have yet been published based on data from the database, the yearly reports (European Commission 2006; European Commission 2007) have included tables of case numbers and incidence rates from the contributing countries.

The dearth of detailed comparative trend analyses between countries in trends may be due, in part, to the great caution over comparing data collected via different surveillance systems, an issue addressed by Waters and colleagues (Waters et al. 1994). With the onset of improved and/or enhanced surveillance programs across Scotland, Canada and United States in the last decade, and the potential for a comparative analysis to provide insight into *E. coli* O157 epidemiology, it seems appropriate for a comparison to now be attempted.

However, prior to a comparison of temporal trends between Scotland, Canada and the United States being attempted, background on the countries and data sets must be provided, the issues involving such a comparison be addressed, and the trends in each country be individually modelled. In this chapter, a background to the

comparison of temporal trends in *E. coli* O157 between countries will be presented. Comparative information on the three countries included in the analyses will be included in order to provide context to the comparisons. Next, methods for statistical analysis of trends will be introduced, followed by a description of the specific methods to be used in this thesis. The issues regarding the choice of method(s) will also be discussed. The final section will offer an introduction to the data sets which will be used in the temporal trend analyses, the results of which are to be presented in Chapters 3 (Scotland), 4 (United States), 5 (Canada) and 6 (Comparison).

2.2 Comparing trends between countries

2.2.1 Selection of countries

As discussed above, the first focus of the thesis is on comparing temporal trends between countries. In order for the trend analyses and comparisons to be statistically feasible, the data sets to be analysed must come from countries with an established surveillance program. Also, the data available from the surveillance must involve enough cases/outbreaks and a long enough time period to provide sufficient statistical power. A number of countries potentially fit this profile including Scotland, the United States, Canada, England & Wales and Japan. However, for this thesis, it was only possible to obtain official data sets from Scotland, the United States and Canada, so the individual and comparison analyses will be limited to these three countries.

2.2.2 Comparing Scotland, the United States and Canada

In order to provide context for the comparison, a number of issues will be examined, including the spatial distribution of cases within each country, the definitions used within each country and the control measures used.

2.2.2.1 Spatial distribution of cases

As illustrated by the statistics in Table 2.1, the three countries being compared in this chapter – Scotland, the United States and Canada – vary widely in size, population, population density and land usage. These are factors which could potentially affect the spatial distribution in each country. Figure 2.1 would suggest that at least for population density, the patterns between the countries vary. There is more of an

east-west variation in the United States and more of a north-south variation in Canada and Scotland.

Table 2.1: Comparison of demographic statistics between countries
The table provides a comparison of demographic statistics between Scotland, the United States and Canada. A rough estimate of cattle density per hectare (p/ha) was calculated by dividing the number of cattle by the hectares of land farmed minus the hectares of land in crops.

Comparison Between Countries			
	Scotland	**United States**	**Canada**
Population (2003)	5,057,400	290, 796, 023	31, 985,000
Land size (km sq)	78, 722	9,826,630	9, 984, 670
Population density	65 person per km^2	29.4	3.1
Land farmed (hectares)	6,100,000	379,708, 062	67,500,000
Land in crops (hectares)	650,000	175,700,320	36,395,150
Cattle population	1,914,000	95,438,000	16,250,000
Cattle density (p/ha)	0.35	0.47	0.52

Data sources: (Scottish Executive 2007a; Statistics Canada 2006; U.S.Census Bureau 2006; United States Department of Agriculture 2007a)

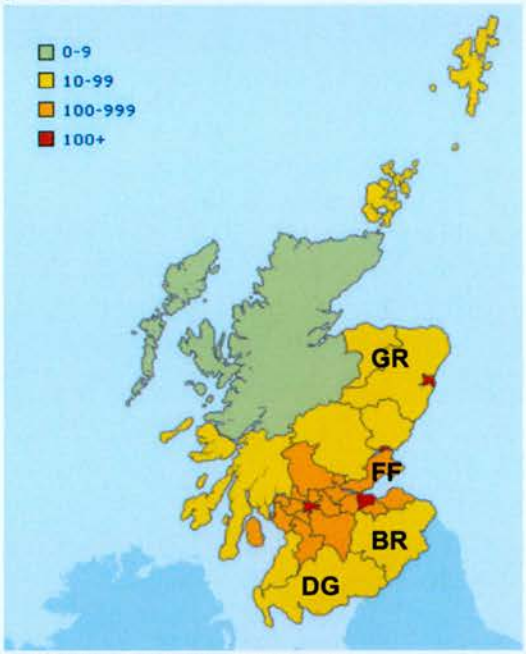
Within Scotland, the rate of sporadic cases increases from west to east and south to north, with the highest rates in the north east and the lowest in the central-west (Innocent et al. 2006) (Fig. 2.1a). Closely linked is the rural/urban divide, with higher rates in the rural areas, as demonstrated by Figure 2.1a which shows that the health board districts that have had some of the highest rates of infection tend to have low population densities.

Overall case rates also closely align with these findings, the highest rates most often in the north-east health board of Grampian (GR), and in the Borders (BR), Fife (FF) and Dumfries & Galloway (DG) (Coia et al. 1993; Coia et al. 1995; Coia et al. 1996; Cowden 1997b; Douglas & Kurien 1997; Sharp et al. 1994a; Thomson-Carter et al. 1996). Differences between rural and urban rates were also noticed in Ontario, Canada where regions associated with higher case rates were frequently of mixed-agricultural use (Michel et al. 1999). Case rates in the United States have historically been highest in the states bordering Canada, with the highest numbers of outbreaks between 1982 and 2002 in Minnesota, Washington, New York, California and Oregon (Rangel et al. 2005) (Figure 2.1b). However, as shown in Figure 2.1b, states with high rates of infection do not always have low population densities. Overall, consistently higher rates of infection have been seen in the central provinces of Alberta, Manitoba and Saskatchewan (Prince Edward Island has very high rates, but

these are likely to be influenced, as with the island health board regions in Scotland, by the very low population) (Fig. 2.1c), with infection rate and population density not appearing to be connected on the provincial level (Sockett et al. 2006).

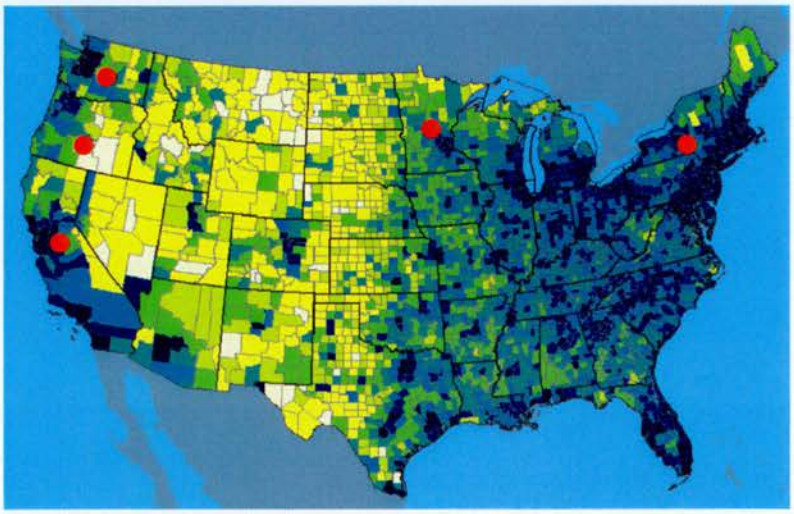
Figure 2.1 a-c: Population density of regions in (a) Scotland, (b) United States and (c) Canada
The population densities by region in the three countries being compared. For Scotland (a) the measure of density is persons per square kilometre, for the United States and the approximate locations of the health board regions with the highest rates of *E. coli* O157 infection are indicated. For the United States (b), the map shows population density in square miles, with darker colours indicating higher density and red dots indicating the states with the highest outbreak rates. For Canada, the provinces which have had the highest rates of infection are indicated by blue dots.

a



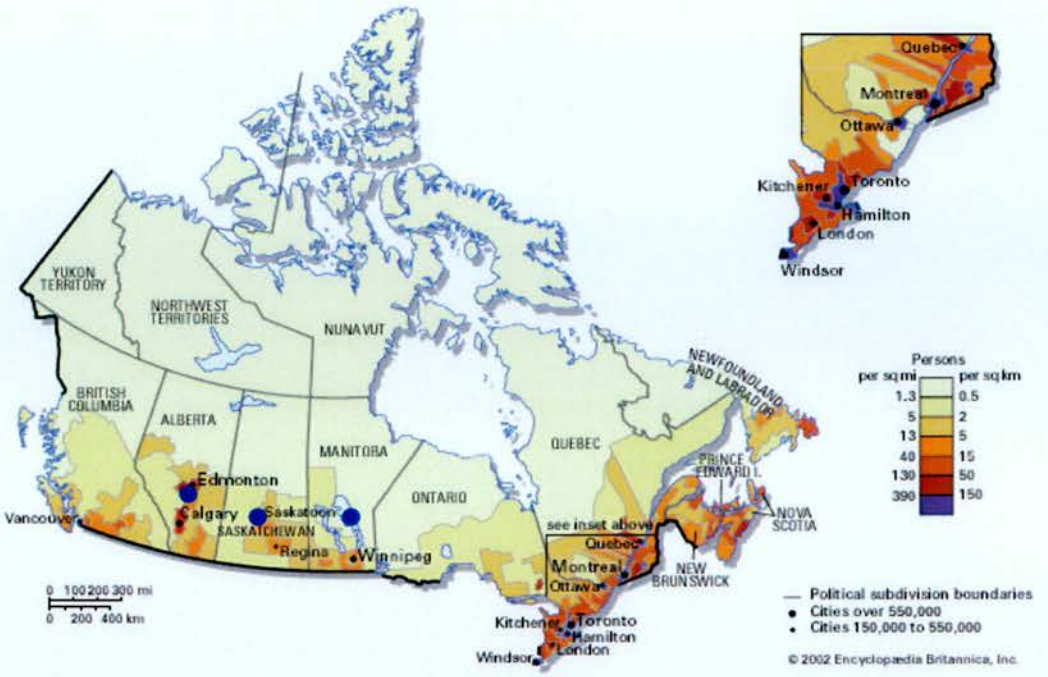
Adapted from www.scotland.org

b



Adapted from www.NationalAtlas.gov

c



Adapted from http://concise.britannica.com

2.2.2.2 Definitions

Case and outbreak definitions for each country will be discussed in the relevant chapters; however it is important to highlight the general similarities and differences between countries in these definitions (see Table 6.1).

When examining the overall number of reported (total confirmed) cases – both sporadic and outbreak-related – the definitions for the cases which are recorded in the data sets vary between countries. In Scotland, the total case count includes only cases which are confirmed by faecal isolate. The Canadian data set also is limited to laboratory confirmed cases (though not explicitly limited to confirmation by faecal isolate), but includes both O157 and non-O157 VTEC cases. Contrastingly, whilst in the United States data set only includes O157 cases, probable cases are also counted in the total. Probable cases may be laboratory confirmed, but can also be clinically compatible cases that are epidemiologically linked to other probable or confirmed cases. These clinically compatible cases are mostly non-laboratory confirmed outbreak cases; cases that in Scotland would be included in the count of the number

of ill cases in an outbreak, but not in the count of the number of ill and positive cases in an outbreak.

For outbreaks, the definitions used to compile the national data sets in Canada and the United States prescribe that outbreaks must involve at least two linked cases. In addition, for a cluster to be an outbreak in the United States, the illness must result from a common source, whilst in Canada the standard is that the cluster must involve two or more clinically or laboratory confirmed cases. Two linked cases are also required in Scotland, however to be defined as an outbreak, the cases must involve persons who are not in the same household. These 'household outbreaks', based on the information provided, are considered as outbreaks on a national level in Canada and the United States. It should be noted though, that in at least one Canadian province, the provincial database may not explicitly identify household outbreaks (Pearl et al. 2006). However, it is not known whether the outbreak data provided by the PHAC was obtained from this data set.

Both Canada (federally) and Scotland have very similar definitions for ill (also referred to as clinical, but can include asymptomatic cases) outbreak cases and ill & positive (also referred to as confirmed) outbreak cases. The latter must be laboratory confirmed whilst the former must only have compatible symptoms (excepting asymptomatic cases) and be linked to the outbreak. It should be noted, though, that the definitions used to collect data within each Canadian province or territory vary, so whilst there is a single definition for the federal data set, the numbers may not directly reflect this definition.

In the United States, the definitions have evolved along with system used to collect outbreak data. The system was at first paper-based and separated between foodborne and non-foodborne outbreaks, then computerised in 1998. However, the foodborne and non-foodborne aspects were not combined until after the time period covered in the analyses. The definition for a confirmed case is essentially as for the other two countries, in that the case must be lab confirmed. Ill cases also include those that are probable, as defined for total cases, which is again, roughly compatible with the definitions in Canada and Scotland. However, it is not clear whether secondary cases were consistently included with the count of either laboratory confirmed cases or ill

cases, as such cases were reported separately on a completed outbreak form provided for this thesis. Additionally, two data sets were merged for the analysis of temporal trends in the United States, only one of which had been edited and updated to comply with current definitions.

2.2.2.3 Control Measures

Across the three countries, a wide range of measures for the prevention and control of *E. coli* O157 infection have been taken. Consideration of these measures is of importance because of their potential influence on temporal trends.

Many of the control measures which have been undertaken in Scotland were directly or indirectly in response to the Wishaw Outbreak. As discussed in section 3.2.3.3, an enhanced surveillance program, which involves long term follow-up of confirmed cases, was initiated in Scotland in mid-1999. The program has the potential to causes changes in the number and type (outbreak or sporadic) of cases reported via surveillance. Recommendations for public control measures since the Wishaw Outbreak have come from reports produced by groups such as the Pennington Group (Pennington 1998) and a task force convened by the Food Standards Agency and the Scottish Executive Health Department with a remit that included a considerations of "what future measures would help protect public health" (O'Brien et al. 2001b). The recommendations which have been put into practice focus not only on the meat industry, but also increasingly on reducing environmental contact. Food safety measures, including the implementation of HACCP (Health Analysis and Critical Control Point), have been introduced for restaurants and other businesses that handle food items. More recently, there has been an effort to educate people about the need for proper hygiene measures at places such as open farms, petting zoos and when camping (Scottish Executive 2007b). It will be of interest to see if the implementation of these measures has had any affect on the trends in foodborne cases/outbreaks.

Additional measures targeting environmental transmission in Scotland include guidelines suggesting a minimum time period between use of a field for livestock and for camping (O'Brien & Adak 2002). A program of increased education about ways to reduce the risk of animal to human transmission has also been implemented

in the United States (Centers for Disease Control and Prevention 2007b). Given the large number of cases in the United States associated with minced beef, control measures in the U. S. have also focused on the beef industry. These measures have included a monitoring program and HACCP (Wachsmuth et al. 1997). On the consumer level, programs have encouraged safe food preparation techniques with a particular emphasis on ensuring thorough cooking of minced beef (Centers for Disease Control and Prevention 2005b). The creation of the FoodNet and PulseNet networks have provided improved surveillance and linkage of cases, while research in both Canada and the United States has proceeded on methods, such as cattle vaccines, to decrease the carriage of the bacteria in cattle (Tauxe 1998). Since such programs are aimed at decreasing the number of cases and outbreaks, one of focus of the analysis of temporal trends will be to determine whether there has been a decrease in case/outbreak numbers.

Control measures in Canada have also included education – via brochures and websites – of the public about food safety measures including proper food preparation and cooking techniques (Canadian Food Inspection Agency 2007). Day care centre policies have been assessed, and campaigns have been launched to inform consumers about outbreaks and general prevention methods. In addition, there have been changes to commercial food processing practices (Sockett et al. 2006). Since the Walkerton Outbreak, there has also been a focus on ensuring safety of the water supply (Todd 2000). Collaborative research such as the IPRAVE study, which looked at links between infection in cattle and humans, have been part of control measures in all three countries. Many of these measures have been enacted too recently for effects to be linked to statistically significant temporal trends. However, the effects will be considered in the interpretation of temporal trend models, as differing control measures could play a greater role in any differences in trends between countries than dissimilar underlying epidemiological factors.

2.2.3 Temporal trends by country

Prior to comparing temporal trends between Scotland, the United States and Canada, the trends in each of these countries must be analysed separately. The trends which can be appropriate and adequately modelled within each of the individually countries

can then be compared. The next section of this chapter will introduce and discuss the statistical methods used in the analyses of temporal trends.

2.3 Descriptive and statistical analysis of temporal trends

In the next four chapters, temporal trends in *E. coli* O157 outbreaks and cases will be analysed in three separate countries and then compared. This section will begin with a brief introduction to the statistical analysis of temporal trends, followed by a description and discussion of the methods selected for the analyses in Chapters 3 to 6. Finally the issues of multiple testing, power and uncertainty, in the context of the trend analysis, will be discussed.

2.3.1 Selection of variables

The variables from each country will be discussed in detail in the relevant chapter, but this section is intended to provide a brief overview as to the trend selected to be modelled and compared.

The selection of variables is driven both by an interest in determining the statistical significance of trends suggested in the literature, as well as by the data made available (see Section 2.4.2 on issues of data availability). Trends in overall cases, sporadic case, outbreaks and outbreak cases are of interest in order to provide a complete picture of any changes that have been occurring in terms of the epidemiology of *E. coli* O157. For instance, is there a general decrease in cases, or only in either outbreak cases or sporadic cases? The other main focus is on outbreak size and mode of transmission in outbreaks, since it has been suggested, at least in Scotland, that there has been a shift in the dominant mode(s) of transmission.

2.3.2 Methods for analysing temporal trends – non-selected methods

As introduced in Section 1.8, there a number of methods for analysing temporal trends (time trends, time series trends). These include Poisson regression (Agha et al. 2006; EURODIAB ACE Study Group 2000; Kamper-Jorgensen et al. 2006; Sandberg et al. 2006; Tleyjeh et al. 2005), with the time (year, month etc) as a continuous, categorical or offset variable, simple linear regression (Van der Heyden et al. 2000), Bayesian methods (Maule et al. 2006), time series analysis (Kuhn et al.

1994), Chi-square tests of trend (Joe & Kaplan 2002; Marcus et al. 2004b) and non-linear regression methods (Anthony et al. 2006).

- **Non-linear methods**

A number of non-linear methods can be applied, such as regression analysis with a quadratic or cubic term (Anthony et al. 2006) and the Bayesian Generalised Linear Mixed Models (GLMM) methods (Maule et al. 2006).

Polynomial functions, which include the independent variable (x) raised to various powers, can be used to model non-linear curves (Crawley 2002). A model with a quadratic term (x^2) can be used to model humped curves, whilst a cubic term (x^3) can be added to model curves that have double changes in curve shape. A quadratic model has been used to model the temporal trend in the (contribution of the) familial component of high systolic pressure (Province et al. 1989). Trends in seabird populations have also been modelled using Poisson regressions with polynomial terms (Fewster et al. 2000).

Other non-linear shapes can be described by functions such as the Rickert curve ($y = a \cdot x \cdot e^{-bx}$), which is used in fisheries research (Crawley 2002). Generalised Additive Models (GAMs), which are related to GLMs, provide smoothed curves, and can accommodate a wide range of non-linear curve shapes, have also been using in temporal trend modelling (Fewster et al. 2000; Rodriguez et al. 2007).

Bayesian modelling methods include Bayesian non-parametric GLMMs (Maule et al. 2006). Bayesian analyses are differentiated from other regression analyses by the fact that the parameter estimates are obtained from the distribution of the model parameters (Brown & Prescott 1999). The advantages of the Bayesian approach of relevance to temporal trend analyses include the ability to handle overdispersion and correlated data points (Maule et al. 2006). Additionally, non-linear models can account for changes in the direction of a trend (Maule et al. 2006).

However, non-linear methods were not considered appropriate for the analyses of temporal trends in *E. coli* O157 for several reasons. Firstly and most importantly, the *a priori* hypothesis for the analyses of temporal trends is that temporal trends in *E. coli* O157 could be modelled using simple linear techniques. This hypothesis has

been established because one of the aims of the study is to create models which can be explained via biological mechanisms, rather than simply fitting a trend with as many parameters as needed to find the closest fit to the data points. The concern with using highly parameterised models, such a polynomial and other non-linear models is that with data from such a short time period which has low counts and a great deal of variation, it may not be possible to determine true non-linear trends. Starting with linear models also is in line with the principle of Occam's razor, which in modelling means that "models should have as few parameters as possible..." and "linear models should be preferred to non-linear models..." (Crawley 2002).

- **Time series analysis**

Time series analysis, as discussed by Kuhn and colleagues, is a form of regression analysis in which the relationship between data collected at regular time periods is examined (Kuhn et al. 1994). In contrast to regular analysis, the points on the regression line (dependent variables) are predicted from "values of the outcome at previous points in time..." (Kuhn et al. 1994). Time series analyses are autoregressive, in that they account for correlation between time points. However, because of the different prediction method and the fact that they are used to look for long term patterns, time-series regressions are generally used where there are least 30 time points (Dan Haydon, personal communication). Thus, since the data sets for the temporal trend analyses only include 7 to 9 data points, time series analyses are not appropriate.

- **Chi-squared tests for trend**

Chi-square tests for trend use the Chi-test to look for trends in the proportion of population with a specific trait (Kirkwood & Sterne 2003). Such tests are not used frequently for temporal trend analyses, in particular for infectious diseases, because data from surveillance only provides information on the numbers of persons with a disease, not on the number of people without the infection. For this reason, chi-squared test for trend is not an appropriate method for the analyses in Chapters 3-6.

2.3.3 Methods for analysing temporal trends – selected methods

• Poisson regression (GLMs) and linear regression – introduction

Regression using GLMs with Poisson error structure will be detailed in Section 2.3.4.2, but briefly, the aim of such regression is to "fit to observed data a regression equation that accurately models the expected value of the dependent variable Y, E(Y), as a function λ of a set of X independent variables… and β regression parameters" (Kuhn et al. 1994). In terms of modelling temporal trends in *E. coli* O157, GLM regression with Poisson error structure models the expected number of cases or outbreaks as function of year and the parameters in the regression model. Linear regression has a similar aim, but is used when the dependent variable is continuous or when the count data are log-transformed and are thus no longer in the form of integer counts.

• Poisson regression – reasons for selection

GLMs using Poisson and binomial error structures, and linear regression have been selected for the analyses for a number of specific reasons. Firstly and most importantly, as mentioned above, the a priori hypothesis for the analyses of temporal trends is that temporal trends in *E. coli* O157 can be modelled using simple linear techniques. The reason for this hypothesis is that a major focus of the trend analyses is to create models which can be explained via biological mechanisms, rather than simply fitting a trend with as many parameters as needed to find the closest fit to the data points. Limiting the number of parameters also follows the principle of Occam's razor, which in modelling means that "models should have as few parameters as possible…" and "linear models should be preferred to non-linear models…" (Crawley 2002).

Secondly, Poisson regression is widely to model trends in count data (Crawley 2007) of relatively rare events (Kuhn et al. 1994) and most of the *E. coli* O157 data is in terms of counts, and infection is relatively rare. Other error structures e.g. Weibull and Gamma distributions and transformations such as the inverse will not be used because of the need to avoid over-parameterisation (Weibull and Gamma distributions have two parameters, as compared to the single parameter of Poisson model). Also, both Weibull and Gamma distributions are generally applicable when

the data is the form of continuous measurements (Crawley 2007), whilst the measurements in this thesis – numbers of outbreaks and cases – are counts.

The negative binomial error structure has also been considered because one of the tenets of modelling biological processes, such as *E. coli* O157 infections, is that "individuals of any one kind (or species) are seldom, if ever, randomly dispersed in space" (Cassie 1962; Miller et al. 2004a), but instead they are clustered or aggregated. In terms of *E. coli* O157, infections are aggregated into events (i.e. outbreaks) and the size distribution of these events, which underlies all *E. coli* O157 surveillance, is in itself aggregated, as there are a few large events and many small events or single sporadic cases (Locking et al. 2006a; O'Brien & Adak 2002; Willshaw et al. 2001).

The negative binomial error structure, in which one of the two parameters, k, is an inverse measure of aggregation (Fisher 1941), is used to describe populations that are aggregated or overdispersed (Bliss & Fisher 1953) such as macroparasites in wildlife host populations (Shaw et al. 1998). Thus it would seem appropriate to use models with negative binomial error structures to describe the *E. coli* O157 trends in this thesis. However, because many of the trends analysed had non-zero minimums - for instance, there must be at least two ill cases in an outbreak – the use of truncated negative binomial models would be required. Such models are considered to be outside the scope of the analyses in this chapter, in which the intent is to use simple linear models. The use of truncated negative binomials will, however, be addressed in Chapter 8 when prevalence is modelled.

In addition, Poisson regressions have been frequently used to model short term temporal trends with as few as six time points (e.g. (EURODIAB ACE Study Group 2000; Tleyjeh et al. 2005; Van der Heyden et al. 2000)). This is of particular relevance because the data sets being used contain only from seven to nine time points. Since, as mentioned above, time series analyses usually require a large number of data points in order that long term patterns might be recognized, they are not appropriate. The non-linear analyses mentioned above, in particular the GAM analyses, also tend to include longer data sets.

Finally, linear Poisson regressions are used because of the need to create models with as much statistical power to detect trend as possible, despite the low number of data points. With only 7 to 9 data points, the inclusion of more than a slope and an intercept in the model would result in very few degrees of freedom in the regression and thus low power to detect any significant changes in trend. Using non-linear methods would increase the number of parameters, and with only seven time points in some analyses, having three or more parameters would result in fewer than four degrees of freedom. Such a model would only be able to demonstrate statistical significance for very large changes in value over time.

Where the data is the form of proportions, a binomial error structure can be used with a GLM, and where logarithmic transformation is needed to normalise data, linear models can be used. Such log transformation can linearise data so that Poisson regression can be used in line with the aims of this thesis.

There are however, some implications of the choice to use linear regression methods. Firstly, such methods (with the exception of using a log transformation to linearise a curve), cannot detect a non-linear trend, so such trends may not be recognised (Maule et al. 2006). Additionally, However, GLM regression with Poisson error structure is not always appropriate for temporal trend data because the key assumption for regression is independence of observations (Zeger et al. 2006) and data points in a time series may be correlated. (Here the observations are not considered to be correlated). When interpreting a model fit, the possibility must be considered that even if a linear model can be fitted to the data, that the trend may not be truly linear. Thus a linear model may over-simply the nature of a trend. As a result, it is important that all data be plotted and examined prior to modelling so that clearly non-linear trends are excluded from linear analysis.

The next section will detail the specific methods used for the analyses in Chapters 3 to 6.

2.3.4 Statistical methods for analysing temporal trends in Chapters 3 to 6

For this thesis, temporal trends in both count and continuous data were analysed using GLMs and linear models as specified below. Methods specific to a particular country will be presented in the relevant chapter.

2.3.4.1 Generalised linear models

While the assumption in a linear regression is that the residual error structure is normal, the ability of GLMs to be adapted for other error structures permits them to be used for analysing non-normal data with categorical, count or continuous outcome variable(s) (Crawley 2007). There are two main components to a GLM model, the error structure and the link function. The error structure, for instance Poisson or binomial, specifies the error structure of the data being analysed. The link function is based on the error structure (though it can be varied in certain circumstances), and is essentially a transformation of the outcome variable prior to modelling. When the model is fitted, the predicted value is thus obtained from the linear predictor by back-transformation.

The transformation by the link function allows non-linear equations to be linearised so they can be analysed by GLM models. For example Crawley (Crawley 2002) gives the example of the equation $y=\exp(a+bx)$ which is not linear. It can however be linearised by log transforming both sides of the equation so it is now $\ln y = a+bx$. This transformation can be accomplished with GLMs using the log link function.

In GLMs the values for the parameters, indicated as a, b,... are obtained by maximizing the log likelihood via an "iterative, weighted least squares" method (Brown & Prescott 1999). Using least squares to determine the best fit parameters is based upon fitting a line for which the sum of the squares for the difference between the y value of each data point and the corresponding y value on the line are minimized. These differences, the residuals are weighted by 1/variance matrix (Brown & Prescott 1999).

The outcome variable data for this study is either count, continuous or binomial and the type of models used for each type will be discussed below.

2.3.4.2 Count data

For count data, the initial model used is a GLM with Poisson error structure. The Poisson error structure is a discrete distribution, and is thus used in instances where the data is integers, such as counts. It is particularly appropriate for count data which by definition includes only non-negative integer values – one cannot have -3 *E. coli* O157 outbreaks, for example - where the link function is log, and thus restricts outcome predictions/Poisson parameter estimates to values greater than zero. The equation for the Poisson GLM model is given below:

$$\text{Log}(y) = a + bx + \text{error from Poisson} \quad \text{or } y = \exp(a + bx) + \text{error from Poisson}$$

In R (R Development Core Team 2007) this is coded as below where y indicates the y variable

```
Model <- glm(y~x, family = Poisson)
```

GLM with Poisson error structure is a test for trend which assumes that the increase in log y is the same for each unit increase in x, which in the data set used in this chapter is one year.

Each model is checked for appropriateness by inspection of residual plots and for dispersion by calculating the ratio of residual deviance to residual degrees of freedom. Where the ratio of residual deviance to residual degrees of freedom is markedly greater or less than 1, meaning that the variance is not equal to the mean as assumed in Poisson regression, there is over- or under-dispersion (Crawley 2007). In these instances a quasipoisson model is used. The quasipoisson model is identical to the Poisson model except that the over or under-dispersion is accounted for by not setting the dispersion parameter, which is measure of the dispersion, at one, but at a value that is the ratio of the residual deviance to residual degrees of freedom. The standard errors of the parameters in the quasipoisson models are then adjusted by a factor of the square root of the dispersion parameter. This adjustment provides a more conservative estimation of statistical significance in overdispersion and a more liberal estimation in underdispersion, correcting for the fact that the distribution is not strictly Poisson. The greater the over- or under-dispersion, the greater the adjustment.

Two residual plots are examined for each quasipoisson or Poisson model. The first, the Residuals vs. Fitted plot shows the residuals from the model plotted against the predicted values. There should be no pattern in these residuals for linear models and an expanding pattern (variance proportional to the mean) for Poisson models. If there is a pattern indicating heteroscedasticity (variance increasing out of proportion with the mean), transformation or a different error structure is used normalize the pattern. The Normal Quantile-Quantile (Q-Q) plot in which the theoretical quantiles were plotted against the Studentised deviance residuals is used to check for normality of errors. If the data points deviate too much from the x=y line, transformation (after consideration of the Residuals vs. Fitted Plot) is attempted to normalize the data.

Transformations and error structures are limited to those with associations to plausible biological mechanisms. When no transformation or error structure results in acceptable residual plots, the nature of the deviation from normal error plots is described and a plain plot of the variable against time was provided. Since most regressions only involve 9 or 10 data points, deviation of the points from the normal line is accepted and in cases where the deviation was border-line, discussion of the possible influence on the regression results is included. If the residuals indicate that a transformation is needed, log10 transformations are used as appropriate. Once the data is transformed, a linear model is used, as described below for continuous data.

For the modelling of variables in which the data is transformed, for example Ill and Positive Cases per Outbreak, linear models are used, with the equation:

$Y = a+bx$ (R code: Model<-lm(y~x ...)

Model appropriateness is determined as described above.

2.3.4.3 Binomial/Proportion data

Variables that involve proportions, for instance the proportion of outbreaks that are foodborne, are modelled using GLMs with binomial error structure. In binomial GLMs, the logit link is used, the logit transformation of p being ln(p/q) where p is the proportion of successes or occurrences and q the proportion of failures or non-occurrences. Thus the equation is:

$$y \text{ (i.e. p)} = e^{(a+bx)} / 1 + e^{(a+bx)}$$

The models are run in R, using the cbind function which allows separate variables for the number of occurrences (a) and number of failures (b) to be combined to create p and q.

Model<- glm(cbind(a, b) ~ x, family=binomial…)

The values for a and b – for instance the number of foodborne outbreaks and the number of outbreaks that are not foodborne - are entered into cbind as the number of successes (foodborne outbreaks) and failures (total outbreaks – foodborne outbreaks) and these two values are used to calculate p and q, and subsequently the linear predictor (logit), ln(p/q).

In the proportion variables, the denominators (a+b) in the proportion (a/a+b) differ between years, and thus the amount of data (counts) represented by each proportion varies. For instance 0.4 could be 4/10, 2/5 or 6/15 and where more data (counts) are used in calculating a proportion – i.e. a larger denominator (a+b) - there is more statistical information and thus should be more influence in the regression. In order to distribute influence appropriately, the regressions must be weighted by the denominator (a+b). Since both a and b are entered, rather than just a proportion, the program (R, v 2.5.1) can automatically calculate the denominator and weight the regression.

A quasibinomial model is used when there is evidence of overdispersion as described above and the outcome variable could be expressed in terms of successes (a) and failures (b).

2.3.4.4 Comparison between modes of transmission - ANCOVA

It is important to avoid performing too many statistical tests because over-testing can result in false-positives in tests of significance (Kirkwood & Sterne 2003). Thus where variables are compared between modes of transmission, analysis of covariance (ANCOVA) – which combines regression and analysis of variance (Crawley 2005) - is used to determine whether there is an overall statistically significant difference in trends over time between modes of transmission. A statistically significant difference in temporal trends between modes of transmission in the temporal trend is

considered to exist when there is a significant interaction between the time and mode of transmission explanatory variables in the ANCOVA analysis. (The variables for which ANCOVA is used are number of outbreaks, proportion of total outbreaks, number of ill/ill and positive cases in outbreaks and proportion of ill/ill and positive cases.)

In ANCOVA there is at least one continuous and one categorical explanatory variable, in this study, year and mode of transmission respectively. A regression model is fitted for each level of the categorical variable with slopes and intercepts estimated for each level. A separate data set was used for the ANCOVA models, which was constructed with a separate entry for the number or proportion of outbreaks or ill cases for each mode of transmission for each year between 1998 and 2004. Both year and mode of transmission were entered as explanatory variables along with the interaction between the two variables, using the following log-linear model (the variable *number of ill cases* is used as an example below):

Number of Ill Cases = $\exp(a_i+b_i \times \text{Year})$ where a_i and b_i are separate parameters for each mode of transmission

R code: Model <- glm(y~ Year*Transmission, quasipoisson, data=....), where * indicates a test with interaction

If the p value for the F-test for interaction was less than 0.05, contrasts were used to look for statistically significant differences in temporal trends between outbreaks and models were constructed for each mode of transmission as described above.

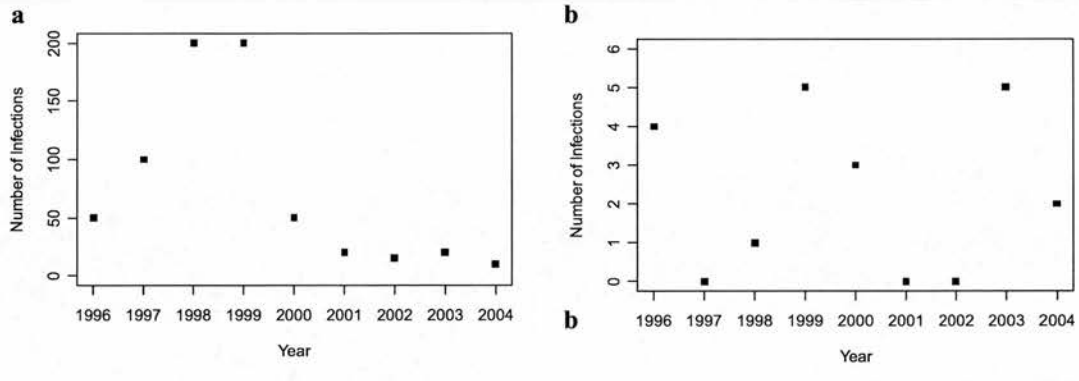
2.3.4.5 Procedure for modelling

Checking the appropriateness of the data

Prior to formal modelling, the data for each potential variable is plotted against year to check for obvious non-linearity (for example, as in Fig. 2.2a) or data that is too varied to be appropriate for formal modelling over a short time period (Figure 2.2b). Additionally, epidemiological factors that may make preclude statistical modelling are considered. For instance, if a specific mode of transmission was not considered in investigations prior to a certain year, then modelling of the trend in that mode would not be attempted. Potential outliers are also considered. However, as the time periods involved are relatively short, the presence of a single potential outlier does

not necessarily exclude data from being modelled. There is a potential for outliers, such as very large outbreaks (see section 2.4.1 for discussion of large outbreaks such as Wishaw and Walkerton). In the case of such potential outliers, data may be modelled with and without the outbreaks to assess the influence of the data point. Thus, only where it would be appear that the data is of sufficient quality for modelling, in terms of linearity, variance and the effect of surveillance changes, is the trend modelled.

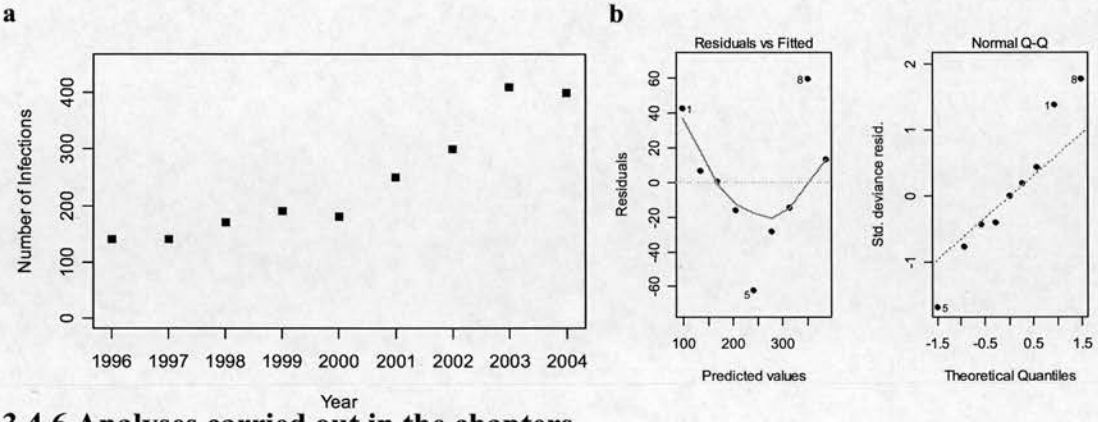
Figure 2.2 a-b: Examples of data that are not appropriate for modelling
Two examples of data plots for data that are most likely not appropriate for modelling as per the methods for Chapters 3 – 6. In (a) the trend is clearly not linear, and in (b) there are very low counts and a very large variation in the counts between years.



Checking for the appropriateness of the model

If the data is appropriate for modelling in terms of linearity, variation and epidemiological/surveillance factors, modelling is attempted using GLM regression on raw data or linear regression on log transformed data. Statistical appropriateness of the models is determined based on residual plots. An example of residual plots showing data that is appropriate to model is given in Fig. 2.3. In the residual vs. fitted plot, no clear pattern in the data points should be seen, while in the Normal Q-Q plot, the line of points should be close to the dotted (normal line). However, given the low of number of data points, some non-normality in the residual plots is considered to be acceptable.

Figure 2.3 a-b: Example of residual plots for count data
A plot (a) of sample count data and (b) the Residuals vs. Fitted Values and Normal Q-Q plots for the GLM model (with quasipoisson error structure) for the data. The residual plots suggest that the model is appropriate.



2.3.4.6 Analyses carried out in the chapters

If no statistically appropriate model can be constructed using biologically plausible limits and/or transformations, the untransformed data is plotted and no further analyses are performed. It is important to note that the determination of whether a model is appropriate is based on the residual plots and visual inspection of the data, and does not imply that models have clinical significance. Clinical significance can only be assessed if more information on clinical goals is available. Thus a model that meets the statistical requirements may not be clinically relevant.

When a statistically appropriate model can be found, the data is plotted with the variable on the y axis and years on the x-axis. In the plots, the best fit line is indicated by a solid line and the 95% confidence intervals for this line are indicated by dashed lines. The predicted best fit line, based upon the model for the known data points, is indicated by a broken line, the predicted 95% CI lines are indicated by dotted lines.

If more than one model appears to be statistically appropriate based on residual plots, models are selected to allow best comparison between data sets with and without large outbreak cases (see section 2.4.1 for details). F-tests are used for all models, except models with Poisson errors for which z values are considered. The level of significance is considered to be 0.05 for all tests.

2.3.5 Issues with the analyses

2.3.5.1 Multiple testing

Since there are a large number of different variables being modelled for each country and then compared, one of the potential issues that must be anticipated is over-testing (multiple testing). The concern with conducting too many tests of significance is that with a significance level of 0.05, 5 out of every 100 (or 5%) of associations will be statistically significant by chance (Kirkwood & Sterne 2003). In terms of the analyses performed, where for instance, approximately 46 trends were analysed for Scotland, this would mean that two or three statistically significant associations between time and a variable would be expected even if none of the trends were statistically significant.

In order to minimise and account the effects of over-testing, a number of steps will be taken. To reduce the number of tests being performed, data will be plotted and visually examined to avoid modelling of data with extreme variation or non-linear characteristics (see Section 2.3.4.5). Also, as mentioned in Section 2.3.4.4, modelling of trends in individual modes of transmission will only be done where there is a statistically significant overall difference between the modes or where there is an a priori hypothesis about a particular mode (i.e. foodborne). Several additional steps are taken post-modelling, the first of which is that the F-test for significance is used in order to provide a more conservative estimate of significance. It is possible to use post-hoc adjustments (e.g. Bonferroni) to account for multiple testing (Petrie & Sabin 2000), however this approach will not be used because of the risk, with potentially lower power to detect statistically significant trends, of not recognizing trends that are actually statistically significant. Furthermore, plots are presented with 95% confidence intervals and exact (or the highest) p-value is given so that the validity of any statistical significance can be assessed. For instance, where 95% confidence intervals are very wide and there is a p-value of 0.045, the result may be called in to question. Finally, the issue of statistically spurious results will be addressed when discussing relevant results.

2.3.5.2 Power

The second statistical issue that must be considered is that of statistical power. Since the time periods being analysed are relatively short (seven to nine years), there is concern that even if trends (or differences in trends) are present, there may not be enough power to detect them. In the context of the trend analyses, there are four main factors that affect power: the number of data points, the variability between these data points, the size of the change in trend that one wants to detect and the significance level. The latter is fixed at the standard level of 5% significance for all analyses in this thesis. The first two factors are determined by the characteristics of the data, and the power to detect statistically significant trends declines as the number of total data points decreases and/or the variability of the data points (around the regression line) increases. The size of the changes (i.e. the percent increase or decrease from year to year, which is also the change in the slope of the regression line) that the analyses need to have the power to detect, in terms of *E. coli* O157, relate to what changes are considered clinically and/or epidemiologically significant. Data on clinical significance (vs. statistical significance) do not appear to be presented in the published literature, with the exception of the target set in 2000 for the overall infection rate in the United States in 2010 (Centers for Disease Control and Prevention 2007a) which is a rate of one case per 100,000 persons.

Thus, in order to estimate the power of the analyses performed in the next few chapters to detect clinically significant trends, a number of power calculations will be presented. The power calculations, run using the TRENDS software package (Gerrodette 1993a), will be done for year to year changes of 10%, 20% and 30% for data involving large numbers of counts (total cases) and 25% and 50% for small counts (number of person to person outbreaks), using data from each country. TRENDS can calculate the power of an linear trend analysis given four factors: the number of data points, the rate of change from year to year (regression line) slope, the coefficient of variation of the data and the significance level (Gerrodette 1993b). The software is designed for linear trends, but since linear trend analysis of log-transformed data is similar to Poisson linear regression (where a log link is used), the software is considered sufficient to provide an estimate of power for Poisson regression. In addition, the package provides an exponential option, which in

assuming lognormal distribution and providing the option for variance to be proportional to $1/\sqrt{\text{mean}}$ (Gerrodette 1987), which is appropriate for a Poisson regression. The inclusion of a value for the coefficient of variation in the calculation also makes it possible to adjust for overdispersion, an important factor since some overdispersion is a potential issue in the analyses (adjusted for by using the quasi function).

The assumptions in such analyses are that the counts are not related, a basic assumption also of Poisson regression. This may not always be true for *E. coli* O157, but the deviation from independence of counts is not considered to be large and thus not likely to have a significant impact. Additionally, for linear power analysis, the change in counts must be additive, whilst for exponential, it must be exponential. As the trends being analysed in the following chapters may not be strictly linear (or strictly exponential), there may be a small amount of additional uncertainty in the estimates of power. Another assumption of the power analyses is that none of the assumptions of the regression analyses are being violated (see Section 2.3.4). Therefore if one or more of these assumptions is violated, for instance if there is overdispersion, the regression may not be as powerful as estimated. However, as mentioned above, overdispersion is adjusted for by the inclusion of the co-efficient of variation as a variable in the power analysis. Thus, the power as indicated by TRENDS should be considered as the maximum possible power, but actual power may be less.

Power calculations indicate that for all three countries, for the total number of confirmed cases (all outbreak and sporadic cases), whether or not the large outbreak cases are included, there is >99% power to detect decreases of 10%, 20% or 30% per year. For example, in Scotland there is >99% power to detect trends where there is a decrease in the number of total confirmed cases of as little as 3%. This would suggest that in the analyses with large numbers of counts, there is enough power to detect trends that might be of clinical significance. While, as stated before, the definition of clinical significance is not known, the goal set in the United States has been for a 50% drop in case rates over ten years, approximately 7% (of the original value) per year. Thus detecting a decrease of 3% would appear to be sufficient. However, when the data for number of foodborne outbreaks is used as an example of

a model with low counts and higher variability, the power to detect trends decreasing (or increasing) at a rate of 50% per year ranges from 92% to >99% in all three countries. In Canada and the United States, there is only approximately 50% power to detect trends that change (increase or decrease) from year to year at a rate of 25%.

As a result, results from analyses where there are low counts, such as the number of outbreaks, number of outbreaks for each mode of transmission, number of cases for modes of transmission such as animal/environmental, and (in Scotland) the number of outbreaks and cases associated with outbreaks that are of PT 2 or PT 21/28, should be considered with caution, as there may not be enough power to detect trends with statistical significance. In these instances, the results may in many instances suggest a non-statistically significant trend even where a clinically or statistically trend is actually present.

2.3.5.3 Sources of Uncertainty

In any statistical analysis, there will be uncertainty based on factors in the data and in the analysis. When analysing temporal trends, both within and between countries, there are a number of sources of uncertainty which affect the interpretation of the modelling. These sources and their effect will be presented here, with more detailed discussions in the relevant chapters.

The main sources of uncertainty in the modelling of temporal trends are missing data and 'artefacts' in the data such as very large outbreaks and changes in surveillance system. Whilst it is unlikely that the Scottish data sets include every case or outbreak in the country (see Chapter 8 for a discussion of under-reporting), missing data is primarily an issue for the analysis of trends in the United States and Canada. As increasing the number of data points generally enhances the precision of a model, missing data would generally be assumed to decrease the precision of modelling, as reflected by larger 95% confidence intervals around the best fit line. However, if inclusion of missing data would in fact increase the variance of points around the mean, the result of missing data could in fact be overly narrow 95% confidence intervals. As will be discussed in Chapter 2.4.1.2, potential causes of missing data in the United States data could be the parallel reporting systems for foodborne and non-foodborne outbreaks, in particular waterborne outbreaks, as well as non-reporting by

some states to the NNDSS in the early years of the data set. In Canada, the issue of missing data is much more acute because it is known that outbreak data is missing from a many provinces/territories for part or all of the time period being analysed. As will be discussed in Chapter 5.5.1, missing data would appear to be much more an influence in the first half of the data set. Thus, trend models may not reflect true trends, and so both trend lines and confidence intervals could potentially be inaccurate. The methods used to assess the possible affect of such uncertainty are addressed in section 5.5 and discussed in terms of the results of the analyses in section 5.5.1

Another factor which introduces uncertainty into the analyses is changes in surveillance. The onset of enhanced surveillance in Scotland (see Section 2.4.1.1) in 1999 has the potential to cause changes in trends, for instance due to an increase in cases picked up surveillance. If there is a change in trend caused by an artefact such as reporting or definition changes, modelling the trend over the complete time period using linear analysis is not appropriate. Two separate trends with low variability around two regression lines, for instance, might be incorrectly interpreted as one longer trend with much more variability (and thus wider 95% confidence intervals). Thus, for the Scottish data, these two trends were analysed separately to determine whether they were statistically significant (see Section 3.2.5.1). For the United States, where the time period must be truncated (see section 4.2.1.2) because of significant changes in surveillance methods, there is an increase in the size of confidence intervals due to the decrease in the number of data points being analysed.

A final source of uncertainty is mis-categorisation of outbreaks with regards to mode of transmission. Determination of mode of transmission is made based only on information provided in the data sets or from direct correspondence with epidemiologists at the relevant agency. However, since mode of transmission categories have differed between and within countries as well as over time, there are likely to be some mis-categorisations. For instance, many outbreaks in the Scotland which are listed as other than environmental may have environmental factors (Strachan et al. 2006). The effect of such uncertainty depends on variability of the data points, if they reflected the true modes of transmission, around the regression

line. If the true variability was higher than that suggested by the available data, mis-categorisation might result in a better fitting model. On the other hand, if the true variability was lower than that of the data set, mis-categorisation would result in a poorer fitting model.

2.4 Data sets

The individual country data sets, including the definitions and country-specific issues will be discussed in detail within the relevant country analysis chapters. In this section, an introduction to the surveillance measures in each country will presented, followed by a discussion of some of the general aspects regarding the obtaining data will be presented in this section.

2.4.1 Surveillance

2.4.1.1 *E. coli* O157 surveillance in Scotland

General overview

In Scotland, national level *E. coli* O157 surveillance programs are coordinated by HPS (formerly SCIEH). Incidents of infection have been voluntarily reported to HPS and its predecessors on a weekly basis since 1989 by physicians, hospitals and laboratories. Data on confirmed cases is collected from the Scottish *E. coli* O157 Reference Laboratory (SERL), which confirms the presence and serotype of *E. coli* in samples sent from diagnostic laboratories; other diagnostic laboratories and NHS public health teams. In Scotland, all isolates submitted to SERL are also phage-typed (NHS Lothians 2005). Overall totals for outbreak and sporadic cases are ascertained by the number of confirmed isolates reported to HPS. Outbreak data is usually recorded and reported via *E. coli* O157 specific outbreak report forms, filled out by the appropriate public health team(s), with information updated as appropriate by public health consultants and epidemiologists involved with the outbreak investigation. HPS may become directly involved in an outbreak investigation if an outbreak involves more than one health board region and/or the outbreak investigation team requests assistance.

Wishaw Outbreak

The largest outbreak to-date in Scotland took place in Lanarkshire between November 1996 and January 1997, with confirmed cases in two other health board regions. The outbreak, which resulted in at least 17 deaths, is termed the Wishaw Outbreak for the purposes of this study, though it also referred to as the Central Scotland Outbreak (Cowden et al. 2001). Over the course of the outbreak there were 512 ill cases and 279 ill and positive cases reported, more than seven times the number of ill cases and nearly 14 times the number of ill and positive cases in any other outbreak in Scotland between 1996 and 2004. The potential effect of the outbreak on temporal trends and how the outbreak will be handled in the analyses will be discussed in Section 3.2.3.2.

Major changes

Major changes to the *E. coli* O157 surveillance system have occurred in 1996 and in 1999. A new surveillance system was enacted in 1996, with major changes implemented in order to create a more active outbreak surveillance system that would provide information on all general outbreaks. The major change involved a switch from the World Health Organization reporting form, which included many open ended questions and was returned only for foodborne outbreaks, to a new Scotland-unique form. The Scottish form is specific to *E. coli* O157, has more focused questions and is completed for all outbreaks involving members of more than one household, regardless of mode of transmission. There was also a shift from passive surveillance, where Consultants in Public Health Medicine (CsPHM) were responsible for reporting outbreaks and returning forms, to a more active surveillance where Health Protection Scotland (then Scottish Centre for Infection and Environmental Health) tracks, logs and send out forms for all possible outbreaks. This ensures that all parties are equally informed about potential incidents (Cowden 1997a; SCIEH 1997b).

It would be expected that the changes created in 1996 have resulted in a greater number of outbreaks being reported because outbreaks are much less likely to be missed by both the CsPHMs and the HPS. In addition, more non-foodborne outbreaks are likely to be reported since the forms have to be returned for all outbreaks. Finally, with forms tailored to Scotland, and with fewer opened ended

questions, the information collected is likely to be more uniform in terminology and definition, and thus more comparable.

The enhanced surveillance program, which began in April 1,1999, is a comprehensive and integrated active surveillance program in which HPS receives, collates and cross-validates data from public health teams, the Scottish *E. coli* Reference Lab (SERL), Obsurv (the HPS infectious intestinal disease surveillance program), ENSHURE (the HPS HUS surveillance program) and other diagnostic laboratories. The information is kept in an integrated database, and questionnaires are sent out to case patients and involved medical practitioners. These questionnaires are used both to get more detailed information on cases in the short term and for long term case follow-up (Locking et al. 2003a)(HPS, unpublished data).

It is possible that changes to data collection forms and protocols as a result of the onset of the enhanced surveillance could potentially have affected long term trends. Most importantly, more in-depth investigation of cases may have resulted in more accurate assessments of outbreak mode of transmission, especially in terms of highlighting less obvious causes which would have gone under-reported before. Additionally, data from questionnaires and other surveillance sources (see above) may have helped to identify cases that were not picked up in the initial investigation and/or to link together prior to un-related cases. All of these changes could impact both the slope and magnitude of trends and/or create the false pretence of trends where none actually exists. Analyses will be conducted to determine whether the enhanced surveillance program has had a statistically significant effect on temporal trends. These analyses, which involve regressions of the data points before and after the onset of enhanced surveillance, are discussed in Section 3.2.5.1.

2.4.1.2 *E. coli* O157 surveillance in the United States

Case numbers -- NNDSS and PHLIS reporting systems

Beginning in 1994, *E. coli* O157 cases in the United States were reported to the CDC through the Notifiable Diseases Surveillance System (NNDSS) and the Public Health Laboratory Information System (PHLIS). Totals for both systems were reported in the yearly notifiable disease summaries printed in the CDC's Mortality and Morbidity Weekly Report (MMWR) between 1995 and 2001, but starting in 2002,

totals from the PHLIS system were no longer published. Beginning in 2003, selected states began the transition to the current National Electronic Disease Surveillance System (NEDSS).

Data from the NNDSS includes both probable and confirmed cases of infection; a confirmed case is considered to be a laboratory confirmed case, which is a case where either *E. coli* O157:H7 is isolated from a specimen or Shiga toxin producing *E. coli* O157:NM is isolated from a clinical specimen (Centers for Disease Control and Prevention 2006a). Between 1996 and 1999 inclusive, a probable case was defined a case "with isolation of *E. coli* O157 from a clinical specimen, pending confirmation of H7 or Shiga toxin OR a clinically compatible case that is epidemiologically linked to a confirmed or probable case" (Centers for Disease Control and Prevention 2006b). After 2000 the probable case definition was amended to include cases where there was "identification of Shiga toxin in a specimen from a clinically compatible case OR definitive evidence of an elevated antibody titre to a known EHEC serotype [for this study, O157:H7 or :NM] from a clinically compatible case. This amendment to the definition may have caused an increase in the number of cases defined as "ill cases".

Outbreak data – CDC data sets

Outbreak data are captured by different systems, the exact structures of which have changed since 1996 as procedures and forms have been updated on the local, state and federal level. Outbreaks are reported to the CDC from state and local health departments either directly by phone or fax or when appropriate, through the routine foodborne disease outbreak surveillance system. Starting in 1998, an enhanced surveillance program was put in place for foodborne outbreaks, encompassing both enhanced communication between various levels of health departments and an updated form (see section 4.2.3.2) (Lynch et al. 2006). In 2001, the reporting system became web based. (Lynch et al. 2006) The reporting form most commonly used for reports of *E. coli* O157 outbreaks, Form 52.13 (Investigation of a Foodborne Outbreak), was updated in 1996, 1999, 2000 and 2004, with major changes in 1999 as a part of the new enhanced surveillance system. The potential effect of these changes on the analyses, which include significant shifts in the total numbers of

outbreaks, and shifts in the types of modes of transmission reported, will be discussed in Section 4.2.3.2.

2.4.1.3 *E. coli* O157 surveillance in Canada

PHAC outbreak data set

For a number of years, the Health Products and Food Branch of Health Canada released annual summaries of enteric disease outbreaks; the 1994-1995 report being the last to list individual outbreaks (Todd et al. 2000). The National Microbiology Laboratory has also released annual summaries of major outbreaks and clusters based on isolates tested and data assembled by the laboratory (National Laboratory for Enteric Pathogens 2002). The current pan-Canadian enteric outbreak data set was assembled between 2003 and 2005 by the Foodborne, Waterborne and Zoonotic Infections Division of the Public Health Agency of Canada (PHAC), based on written, oral and electronic reports obtained from the individual provinces and territories and then compiled and cleaned, for the production of the "Provincial/Territorial Enteric Outbreaks in Canada, 1996 – 2003" report. For the data set, an outbreak was defined as "involving two or more clinically or laboratory confirmed cases".

Outbreak mode of transmission, stated as being "within the context of the geographic area or setting experiencing the outbreak" was classified as either foodborne, waterborne, agriculture, person-to-person, other or unknown, or no value was provided. No details were available in terms of the original collection of data by local, regional or provincial public health officials. In addition, it is of importance that the data on mode of transmission was not confirmed, with virtually no information on the strength of evidence connecting the outbreak to the reported mode of transmission. Further, the data set is known to be incomplete as data was not received from all provinces/territories for each year: of 12 provinces and territories (13 with the addition of the territory of Nunavut in 1999), only four provided data for the entire eight-year period, and complete data was available for no more than 8 provinces/territories in any one year. The specific data variables collected by each reporting province or territory differed, the consequences of which will be discussed in Section 5.5.

Total confirmed cases data set

Notification of all VTEC cases, including *E. coli* O157 cases, initially classified as "haemorrhagic colitis", has been mandatory since 1990, with the province of Alberta requiring reporting six year earlier (Waters et al. 1994; Wilson et al. 1997). Only reports on the total number of confirmed VTEC cases for each province/territory are released by the Nationally Notifiable Diseases program of the PHAC. For national surveillance purposes a confirmed case of VTEC infection is considered to be an infection with "laboratory confirmation of infection with/without symptoms" where laboratory confirmation consists of "isolation of verotoxin producing *Escherichia coli* or other toxigenic strains from an appropriate clinical specimen." (Health Canada 2000a). As such, this number includes VTEC cases in which the serotype was not O157. These non-O157 cases represented a small percentage of the total, accounting for less than 5% of VTEC cases reported between 1983 and 2000 (Woodward et al. 2002), so should not have a statistically significant effect on trends. For 2000, data from Ontario also includes non laboratory confirmed "epidemiologically linked cases as part of their VTEC reporting protocol reflecting the Walkerton outbreak" (Sockett et al. 2006). See Section 5.2.1.2 for details on how these cases are handled in the analyses.

Walkerton Outbreak

The largest reported *E. coli* O157 outbreak in Canadian history took place in Walkerton, Ontario during May and June of 2000. Termed the "Walkerton Outbreak" for the purposes of this thesis, the outbreak occurred when heavy rains washed run-off from nearby cattle farms into wells which provided the municipal water supply. Due to the fact that proper procedures for chlorination had not been followed by the managers of the water supply (O'Connor 2002), chlorine levels were insufficient to kill the contaminating bacteria. The result was an extensive waterborne outbreak in which 1346 persons were infected by *E. coli* O157 and/or *Campylobacter*, 1304 directly from the water. Of these persons, 174 were confirmed with *E. coli* O157 infection, 27 progressed to HUS and seven died. A smaller number of persons were infected with *Salmonella, Yersinia, Aeromonas, Giardia* or *Cryptosporidium*.

2.4.2 Obtaining data

2.4.2.1 Introduction

Official outbreak data sets for each country have been obtained from the respective public health or infectious disease surveillance agencies (Health Protection Scotland, Centers for Disease Control and Surveillance, Public Health Agency of Canada). Details on the sources and data are provided in the relevant country specific chapter (3, 4 or 5).

Data on non-outbreak cases is taken from publicly released in reports or on the websites of the above agencies because single case data sets were not available due to concerns of preserving the anonymity of patients. In order for such data to be released, it must be anonymised which requires extra time on the part of the providing agency. Since there has been considerable delay in obtaining outbreak data sets (A final outbreak data set from the CDC was not received until late 2006, the data set from PHAC in mid 2006), a decision has been taken not to request sporadic case data.

2.4.2.2 Time variable selection and issues

Outbreak and sporadic case data from all three countries has been provided by year, with the additional information on date or month for some outbreaks in the Canadian and United States data sets. Weekly or monthly data points are the most frequently used in analyses of infectious diseases, primarily due to the fact that surveillance data is commonly recorded in such intervals. Additionally, increased numbers of data points allow for a higher level of detail in trend analyses, and with weekly or monthly data, seasonal trends can be analysed as in McDonald et al (McDonald et al. 1999). It would be ideal to have an exact starting date – the date of first infection or date of first infection reported – for each outbreak or case.

However, for several reasons, data cannot be analysed on a monthly, weekly or daily level. The first issue is in regards to the issue of patient confidentiality. In order for data on the month or data of sporadic cases to be released, the data set would have to be further processed prior to release by the specific agency in order to strip patient-identifier material from the existing data set. This additional processing was not possible because of time constraints on HPS/CDC/PHAC staff and in getting data for

the analyses in this thesis (as a note, a request to HPS for more detailed information on sporadic data for analyses in Chapter 7 was not carried out). Similarly, monthly data on outbreaks from Scotland has not been obtained because with as few as three outbreaks per year and significant publicity around many outbreaks, there are concerns that identifying the month of an outbreak might lead to identification of the individuals involved.

Also, while in the United States data sets, outbreaks were recorded by month of first illness onset, and in the Canadian data sets, date of report and/or recognition and/or first onset were provided, the lack of standard definition for first onset between and within the two countries made it impossible to determine whether the definitions were equivalent. Additionally, since no monthly data is available for Scotland, modelling of yearly trends from each country is the most practical in terms of facilitating comparison between countries.

Finally, the use of annual data points avoids the issue of dealing with seasonal trends within each year, the study of which require different analytical methods in order for seasonal effects to be identified and separated from yearly trends. However even if seasonal trend analyses are performed, temporal trends can still be separately analysed with the data summarised by year (McDonald et al. 1999) or with separate independent variables for yearly and monthly data points (Guerin et al. 2005b). Again, however, this would require monthly or daily data from all three countries as well as consistent definitions for time variables. Even if monthly data were available from Scotland (in particular when analysing outbreak trends), the ability of analyses to detect a significant trend might be significantly affected by the very low counts for many months.

Chapter 3 -- *E. coli* O157 in Scotland: 1996 – 2004

3.1 Introduction

Though *E. coli* O157 was first identified as a human pathogen in the course of investigating two HUS outbreaks in the United States in 1982 (Riley et al. 1983), the first Scottish isolates were not reported until 1984 (Coia et al. 1995). The first reported outbreak, a cluster of four cases in Dalkeith, took place five years later. Since 1989, there have been approximately 104 general outbreaks reported in Scotland, with the definition of a general outbreak being an incident "in which two or more linked cases experience the same illness, or when the observed number of cases unaccountably exceeds the expected number" and which affects "members of more than one household or residents of an institution" (Smith-Palmer & Cowden 2004). The largest outbreak to-date, involving 512 illnesses, 279 of which were confirmed, was the 1996 Wishaw Outbreak. Overall there were approximately 3487 laboratory isolates of *E. coli* O157 reported between 1984 and 2003.

There has been discussion about the trends in the numbers of these outbreaks and isolates, (Locking et al. 2003a; SCIEH 2000c; Sharp et al. 1994a; Tarr et al. 1990), as well as about shifts over time in the dominant phage types (Reilly 1997) and the general epidemiology of reported cases (Locking et al. 2003c). The studies of these trends have, however, so far been primarily descriptive in nature (Douglas & Kurien 1997; Locking et al. 2006b; MacDonald et al. 1996; Sharp et al. 1994a), with graphs of monthly/yearly case numbers provided in periodic Health Protection Scotland (HPS) reports, and trends most frequently mentioned in the context of shifts from year to year (Douglas & Kurien 1997; SCIEH 2002b; SCIEH 2002c). Statistical models in trend analyses have only been applied for the study of seasonality (Douglas & Kurien 1997) and in the comparison of isolate numbers or proportions between two years (Locking et al. 2006a; SCIEH 2003a).

The lack of statistical analysis in relation to temporal trends is unfortunate because this type of analysis appears to have the potential to contribute a great deal towards the knowledge of the *E. coli* O157 in Scotland, and allow the trends to be compared with those in other countries. This study aims to fill this gap in the literature by describing and modelling, where epidemiologically appropriate and using simple linear models, the temporal trends in Scottish *E. coli* O157 outbreaks and case

numbers. Where the trends can be appropriately modelled, the statistical significance of trend will be assessed. If a trend cannot be modelled, the possible reasons will be discussed. Of particular interest are trends in the predominant modes of transmission and phage types, the role and/or effect, if any, of the Wishaw Outbreak in *E. coli* O157 trends (see Sections 2.4.1.1, 3.2.3.2), and the effect of the onset of the enhanced surveillance program (see Sections 2.4.1.1, 3.2.3.3) on the trends.

3.2 Materials and methods

3.2.1 Data sets

3.2.1.1 Time period

Outbreaks from the years prior to 1996 are not included in the analyses because there were major changes in the surveillance system, as mentioned in Section 2.4.1.1, which took effect on January 1, 1996 (Anonymous 1997). The potential effects of these changes on the data make it inappropriate to compare data from 1996 to the present with earlier data.

3.2.1.2 Outbreaks

The data set for Scottish *E. coli* O157 outbreaks in 1996 – 2003 has been obtained courtesy of HPS. Data for 2004 was obtained from the 2004 Enhanced Surveillance Report (Locking et al. 2006a) and the 2004 Annual Report of General Outbreaks of Infectious Intestinal Diseases in Scotland (Smith-Palmer et al. 2005), and data for 2005 as publicly released by HPS and HPA (Health Protection Agency 2006b).

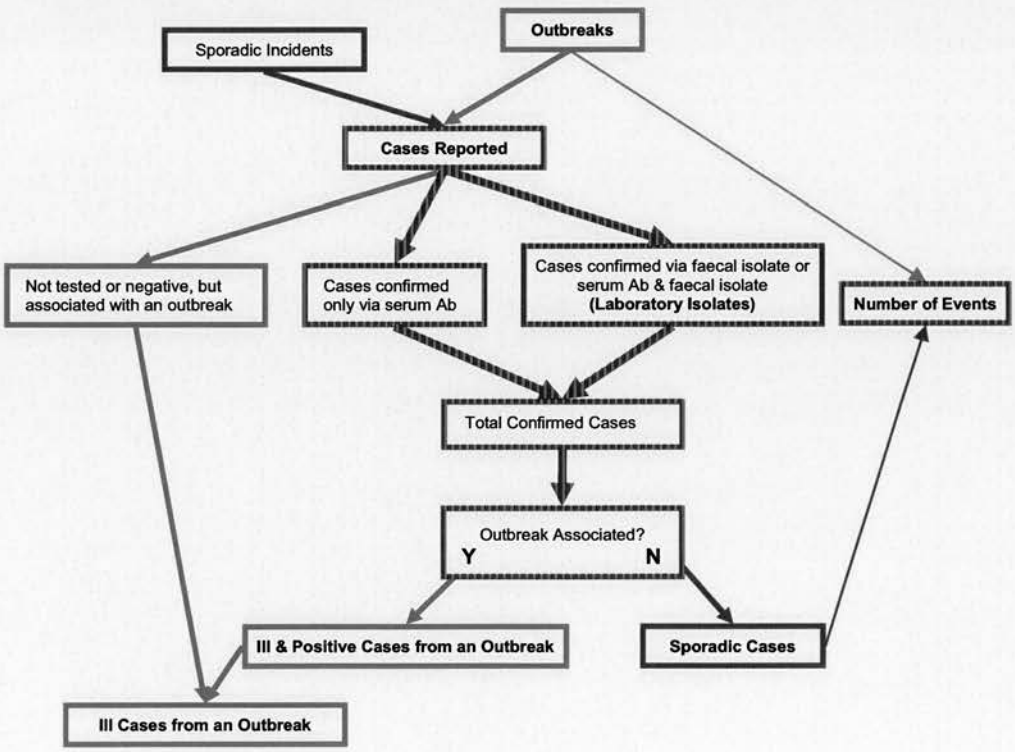
3.2.1.3 Total confirmed cases

Data on total isolates was obtained from the HPS website (www.hps.scot.nhs.uk). With the exception of the alterations made regarding the Wishaw Outbreak cases (see 3.2.3.2), no changes were made to the data or definitions in the above data sets. As such, the definitions mentioned below are applicable to both the HPS data sets and this thesis unless specified otherwise. The definitions vary from those in other countries, an issue which will be discussed in Chapter 6 where trends are compared between countries.

3.2.2 Definitions

In Scotland, the term *case* is generally used to define all epidemiologically plausible instances of *E. coli* O157 infection that have been reported to HPS, whether or not the infection has been confirmed by a laboratory. In the laboratory, confirmation of *E. coli* O157 is made either by isolation of the bacteria from faecal sample or by detection of antibodies in blood serum (see Section 1.4.5). 'Laboratory isolates' refers only to confirmations made from faecal samples (see Figure 3.1), and thus this number does not include those confirmed through blood serum samples. However, the number of cases laboratory confirmed only via serum antibodies is very small, accounting for no more than 3.3% of cases between 2002 and 2004 (Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a).

Figure 3.1: Diagram of *E. coli* O157 variables
Diagram of the *E. coli* O157 variables used in the study. Red lines indicate variables that involve outbreaks, blue lines indicate variables that involve sporadic cases and red & blue striped lines indicate variables that involve both outbreaks and sporadic cases. Variables for which trends are analysed in this chapter are indicated in bold.



Thus, the total number of confirmed *E. coli* O157 cases is the sum of the cases confirmed by the two different methods – faecal isolates and serum antibodies (see

Figure 3.1). This number can then be broken down into the number of sporadic cases – cases not epidemiologically linked to any cases other than those within the same household – and ill and positive cases from outbreaks, which are confirmed outbreak cases. The number of ill cases from outbreaks is the number of confirmed cases added to the number of cases in which the person was ill and epidemiologically linked to the outbreak, but the presence of *E. coli* O157 was not confirmed (see Figure 3.1).

This includes all cases that have been laboratory confirmed as having *E. coli* O157, including those in which the person did not display any symptoms. It is important, however, to note that just because a case is not defined as ill and positive does not mean that the person did not have *E. coli* O157. It is not always possible to obtain faecal samples from all 'ill' persons, and if there is delay before a sample is taken, the person may no longer be excreting the bacteria; Tarr et al reported that the percent of stools from which *E. coli* O157 was recovered declined from 90+% to under 33% after six days of illness (Tarr et al. 1990).

Events are defined as single incidents of infection, whether that be an outbreak or a single sporadic case. This number is calculated by adding together the number of outbreaks and the number of sporadic cases (see Figure 3.1). The count for the number of events, though, is not precise because the exact number of sporadic cases has not been released by HPS for this study. As a result, the number of sporadic cases must be approximated by subtracting the number of ill and positive outbreak cases from the total number of *E. coli* O157 isolates. However, the number of total isolates does not include those infections confirmed by serum antibodies, so this method of approximation underestimates the number of sporadic cases, and thus the number of events. On the other hand, because HPS defines an outbreak as an incident that involves members of more than one household, cases within one household that are related to each other, and according to the definition used in this chapter should be considered as one event, are actually counted as multiple events. Thus, calculating the number of events as the sum of the number of sporadic cases and the number of outbreaks is likely to provide an overestimate of the true value. As

such the number of events, as calculated for this chapter, may not be completely accurate.

In this thesis, putative outbreak mode of transmission is classified as either:

- foodborne
- waterborne
- person to person
- environmental
- animal contact
- multiple including foodborne
- multiple excluding foodborne, or
- unknown

As specified in the HPS 'General outbreaks of infectious intestinal disease' Form (Health Protection Scotland 2004), outbreaks are classified as foodborne, waterborne, person-to-person, environmental or animal contact if the infection was "mainly" spread though the listed mode of transmission. Outbreaks for which food was involved, but not the main factor, are classified as multiple modes including foodborne, and where transmission was the result of multiple non-foodborne factors with none as the over-riding cause, the outbreak was defined as multiple modes excluding foodborne. It should be noted, however, that the selection of the mode of transmission category was made by the CsPHM or other investigator who was filling out the form. Thus, the strength of evidence for the selected mode of transmission is not always known.

For the purposes of this study, to encompass all outbreaks where food played a role, the multiple including foodborne and foodborne classifications are also combined into a final category, which includes all outbreaks with a *foodborne element.* Environmental and animal contact outbreaks were combined to create '*animal/environmental*'. In addition, the outbreak listed as 'environmental or unknown' is classified as *unknown* because no further information about the circumstances of the outbreak was known. The classifications used for the Scottish data will be used as basis for establishing classifications that are comparable across countries. This issue of mode of transmission classification will be addressed again in Chapter 6, when the Scottish data is compared to that from Canada and the United States.

3.2.3 Variables – selection and issues

3.2.3.1 Variables

Table 3.1: Variables used in the analyses of Scottish data, 1996 - 2004
The variables analysed in the analyses of the Scottish data between 1996 and 2004

Variable	Definition
Total isolates	Number of *E. coli* O157 cases confirmed via faecal isolate
Sporadic cases	Number of confirmed cases not linked to other cases outside the same home
Total events	Number of infection incidents = number of outbreaks + number of sporadics
Number of outbreaks	Number of general outbreaks (not within the same household)
Number of ill and positive cases from outbreaks	Number of microbiologically confirmed cases from outbreaks
Number of ill cases from outbreaks	Total number of cases from outbreaks, including those ill, but not confirmed
Number of ill cases per outbreak	Total number of cases per outbreak
Number of ill and positive cases per outbreak	Total number of confirmed cases per outbreak
Number of outbreaks spread via: a) Food b) Multiple including food (MIF) c) Foodborne Element (Food + MIF) d) Water e) Environmental or Animal Contact/Exposure f) Person to person Number of Outbreaks, Phage Type (PT): a) PT 2 b) PT 21/28	Number of outbreaks in which the mode of transmission was a) Food b) Multiple including food (MIF) c) Foodborne Element (Food + MIF) d) Water e) Environmental or Animal Contact/Exposure f) Person to person Number of outbreaks in which the isolates were Phage Type a) PT 2 b) PT 21/28
Number of ill and positive cases from ___ outbreaks	Number of confirmed cases from outbreaks where mode of transmission/PT was a – f/ a- b
Proportion of outbreaks spread via mode of transmission a – f or phage type a or b	Proportion of outbreaks with known mode/phage type, in which mode of transmission was a – f or the PT was a or b
Proportion of ill and positive cases from outbreaks in which the mode of transmission was a – f or phage type was a or b	Proportion of confirmed cases from outbreaks with known mode/phage type in which the mode of transmission was a – f or the PT was a or b
Number of ill and positive cases per __ outbreak	Number of confirmed cases per outbreak from outbreaks which the mode/PT was a – f/a-b
Above variables, but… Without Wishaw Outbreak With Split Wishaw	 With the cases (ill or ill & positive) from the Wishaw Outbreak excluded With the Wishaw Outbreak split into five outbreaks as per Methods & Materials

When investigating temporal tends in *E. coli* O157 cases and outbreaks in Scotland, there are many variables which can be modelled. For this study, variables have been chosen both to provide a coherent analysis of *E. coli* O157 trends over the time period, and to address epidemiological aspects of *E. coli* O157 that have been of interest in recent years (Table 2.1). Consideration is also given to selecting variables that would be appropriate for comparison between countries (see Chapter 6).

Though the primary focus of the chapter is on outbreaks, it is important to place outbreaks into the wider context of *E. coli* O157 infection in Scotland. Thus the first two variables to be analysed describe the overall numbers of infection: the *number of total laboratory isolates* and the *number of total events* (see above for definition and discussion of these variables). The *number of sporadic cases*, calculated by subtracting the number of ill and positive cases in outbreaks from the number of total laboratory isolates, is also included as a way to compare trends involving outbreak cases to trends involving non-outbreak cases.

The first outbreak related variable is the *number of outbreak*s, which provides an overview of trends in outbreak numbers. When looking at overall trends, it is then important to look at the *total number of cases from outbreaks* (ill and ill and positive) as these numbers illustrate the real extent of outbreaks, as for instance a year with three outbreaks each with 70 ill cases is clearly different from a year in which there were three outbreaks, but each with less than five ill cases. The final step is to look at the *number of ill and positive cases per outbreak* in order to see if there have been any trends in the size of outbreaks.

Far from being a homogenous group, outbreaks are comprised of incidents spread by a wide range of modes including food, water and person to person contact (Mead & Griffin 1998). It follows then, that any trends observed overall may be a composite of many different trends in outbreaks spread by various modes. The trends in each mode of transmission will thus be examined. The individual modes of transmission to be investigated are foodborne, foodborne element, multiple modes including foodborne, waterborne, person to person and environmental.

It is equally as important to consider the trends in the modes in the context of their relationship to each other because analyses of the numerical trends do not reveal the relative dominance of transmission modes. Thus the next section in the analyses will examine trends in the proportion of the outbreaks in which cases were spread by each mode of transmission in order to determine whether there was a statistically significant shift in the dominance of one or a few modes of transmission. These results may help to explain whether any increases or decreases in the number of outbreaks represent a change in the dominance (or not) of a particular mode of transmission.

In addition to examining the trends in the number and proportion of outbreaks spread by each mode of transmission, the trends in the *number of ill and positive cases* from each mode of transmission will also be statistically modelled. The number of ill and positive is important because it indicates not just the number of outbreaks, but how many people are being affected by each mode of transmission. It is possible that very different models may be fitted for the number of people in outbreaks than for the number outbreaks. For instance, a mode of transmission that was implicated in just a few large outbreaks may affect more people than a mode implicated in many small outbreaks.

When examining the different trends in the modes of transmission and their relationship to the overall trend, the last major step is to look at the number of *ill cases in outbreaks by mode of transmission*. For this variable, the focus is on foodborne outbreaks because of the noted shift away from foodborne transmission over the course of the time period covered in this chapter.

In the last section, the trends in the two dominant phage types – PT2 and PT 21/28 - will be modelled. Reports have indicated that there have been increasing trends in the number of PT 21/28 cases and decreasing trends in the number of PT 2 case (Locking et al. 2003c). To determine whether there has been a statistically significant change in the numbers or relative importance of PT 2 and PT 21/28, the phage types will be examined in terms of the number of outbreaks, number of ill and positive cases from outbreaks, proportion of total ill and positive cases from

outbreaks that are PT 2 or PT 21/28, proportion of total outbreaks and number of ill and positive cases per outbreak.

3.2.3.2 Issues – Wishaw Outbreak

For the purposes of this chapter, the Wishaw Outbreak raises concerns for two reasons. Firstly, it involves an extremely large number of cases in comparison with any other outbreak in the data set. In addition, the outbreak took place in the first year of the data set, a position from which it could potentially have statistically significant influence on trends in outbreak size. Due to these concerns, in this chapter the data will be analysed with and without the Wishaw Outbreak for variables other than outbreak counts. Outbreak counts will not be analysed with the Wishaw Outbreak omitted, because the Wishaw Outbreak is still one outbreak, regardless of the number of cases involved.

Additionally, in some other countries, clusters or cohorts of infections in different locations caused by a single source are sometimes considered to be separate outbreaks (Banatvala et al. 1996; Ferguson et al. 2005; Michino et al. 1999). In Scotland, this is not true and all 279 confirmed cases in the large Wishaw Outbreak are considered to be part of one outbreak. However, within the outbreak there are a number of clearly defined cohorts (Cowden et al. 2001).

Thus, a possible hypothesis is that Wishaw is not one extremely large outbreak, but a combination of several smaller outbreaks with one initial source. This hypothesis is of interest because subsequent analyses of the Scottish data (in this chapter) have suggested that the Wishaw Outbreak is clearly different from the other outbreaks as case trends often cannot be appropriately modelled when Wishaw is included. In order to determine whether considering Wishaw as a group of outbreaks would have a statistically significant effect on trend modelling, for three variables (the number of outbreaks, the number of ill and positive cases per outbreak and the number of ill and positive cases per foodborne outbreak) the data will also be analysed with the Wishaw Outbreak split into five separate outbreaks as suggested by the information provided in Figure 2 (see below) from Cowden et al (Cowden et al. 2001). Seven sporadic cases with no link to suspect premises and 15 secondary cases could not be ascribed to a particular cohort and thus were omitted from this analysis.

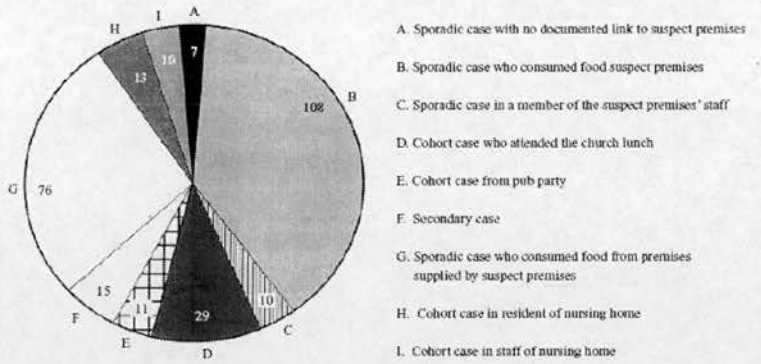


Fig. 3. Exposure of 279 confirmed cases.

The five outbreaks used for the analyses in this study are as follows

A - 23 cases – patients and staff at nursing home

B - 11 cases – persons at pub party

C - 29 cases – persons at church lunch

D - 76 cases – persons who consumed food bought at premises supplied by suspect premises

E - 118 cases – persons who consumed food bought at the suspect premises

For most variables, the Wishaw Outbreak cases are all considered to have taken place in 1996. A few cases did take place in early January 1997, and for the number of total isolates, these cases are counted as part of the total for 1997. However, since the epidemic curve of symptomatic cases presented by Cowden et al (Cowden et al. 2001) indicates that there was only one case in 1997, this difference is not considered to be an issue in terms of the trend analyses.

3.2.3.3 Issues - enhanced surveillance

Another important factor that must be considered in the analysis of any trends in *E. coli* O157 cases and outbreaks in the 1996-2004 time period is the onset of enhanced surveillance. To assess whether enhanced surveillance may have influenced the trends - or lack thereof – between 1996 and 2004, all variables will be assessed for changes in trend slope and magnitude before and after the onset of enhanced surveillance (see Section 3.2.3.3). Since enhanced surveillance became active in April 1999 (Locking et al. 2003b), and so not all data for 1999 was collected under the new system, for the purposes of this chapter, data from 1996 to 1999 will be

considered to be pre-enhanced surveillance. However, if there is any indication of a gradual effect between from 1999 and 2001, when the enhanced surveillance system was coming into effect, this will be presented and discussed as appropriate.

3.2.4 Descriptive statistics

The data from 1996 to 2004 was described in terms of the variables mentioned in Table 2.1. Specifically, the total number of ill and ill and positive cases and the geometric mean number of ill and ill and positive cases per outbreak were calculated, and the data broken down by the number of outbreaks in terms of year, mode of transmission and phage type.

3.2.5 Statistical Analyses – methods particular to Scotland

3.2.5.1 Models to assess the role of enhanced surveillance

For the Scottish data, enhanced surveillance was started in April of 1999. Since the system was not in place for all of 1999, data from 1996 – 1999 is considered to be prior to enhanced surveillance. In order to determine whether this change in outbreak handling affected the data, plots of the variable values over time are visually assessed for noticeable differences in value or slope between 1996-1999 and 2000- 2004).

For the variables in which an apparent difference was noted, regression with a binomial and linear explanatory variable and interaction between the two is used to determine whether there were statistically significant differences in outcome variable means and slopes between the pre and post enhanced surveillance time periods:

$Y = a + b_1x_1 + b_2x_2 + b_3x_3$ where x_1= Year

x_2= Year.b

0 if pre surveillance (1996-9)

1 if post surveillance (2000-4)

$x_3 = x_1 * x_2$ (interaction)

R Code: Model<- lm(y~Year.b * Year …) or

Model<- glm(y~Year.b * Year, family = …)

The model and error structure used for the regression depend on the type and structure of the appropriate model, as already determined. Statistical significance of the binomial variable indicates a significant difference in the mean values, and statistical significance in the interaction term indicates significant difference in the slopes before and after the onset of enhanced surveillance. Where the pre and post surveillance model is of interest, the trend is plotted with separate best fit lines for 1996 – 1999 and 2000 – 2004.

3.2.5.2 Analysis of phage type data

The binomial model (see section 2.3.4.3) is used with a simple binomial response variable instead of the cbind function in order to model the occurrence of PT 21/28 outbreaks versus PT 2 outbreaks over time. The response variable values are "1" when the outbreak is PT21/28, "0" when the outbreak is PT 2 and NA when the outbreak is any other PT. For this variable, the outbreak in which there are three phage types – 4, 8 and 21/28, is considered to be PT 21/28.

3.2.5.3 Modelling procedure

All variables derived from data regarding the number of cases are also modelled with and without the cases from the Wishaw Outbreak, as detailed in Section 3.2.3.2.

In the plots, the best fit line is indicated by a solid line and the 95% confidence intervals for this line are indicated by dashed lines. For the variables where the 2005 value was known – total laboratory isolates and total outbreaks – a value for the data point for 2005 is predicted based on the model for the data from 1996 to 2004 in order to determine whether the models have potential use in predicting future data points. The predicted best fit line, based upon the model for the known data points, is indicated by a broken line, the predicted 95% CI lines from 2004 to 2005 are indicated by dotted lines and the 2005 data point by a bright green dot. For the plots, darker blue dots indicate complete years prior to the onset of enhanced surveillance.

3.3 Results – Descriptive

Between 1996 and 2004 there were 2458 *E. coli* O157 laboratory isolates from outbreaks and sporadic cases reported to Health Protection Scotland by laboratories in Scotland. These isolates represent 1923 *E. coli* O157 events, including 71 outbreaks reported to Health Protection Scotland for which completed forms were returned. The number of outbreaks per year ranged from 3 to 14 (Figure 3.2a). During the period of 1996 – 2004, there were 1048 ill cases from outbreaks (range 22 to 517 per year) and 606 ill and positive cases from outbreaks (range 14 to 284 per year; Figure 3.2b). Outbreaks have a geometric mean of 5.23 ill cases (95% CI = 4.14, 6.61) and 3.95 ill and positive cases (95% CI = 3.26, 4.77).

Figure 3.2 a-b: (a) Number of outbreaks per year and (b) Number of laboratory isolates, ill cases from outbreaks and ill & positive cases from outbreaks, 1996 – 2004
Plots showing (a) the number of outbreaks and (b) the number of laboratory isolates, ill cases from outbreaks, and ill & positive cases from outbreaks per year, 1996 – 2004. In (b), the number of laboratory isolates is indicated by a solid line, the number of ill cases from outbreaks as a blue-dotted line and the number of ill and positive cases from outbreaks as a red-dashed line.

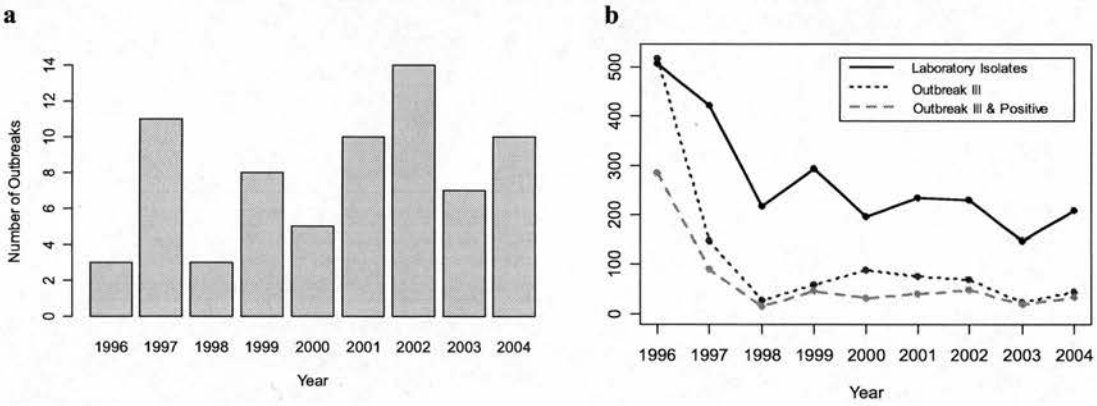
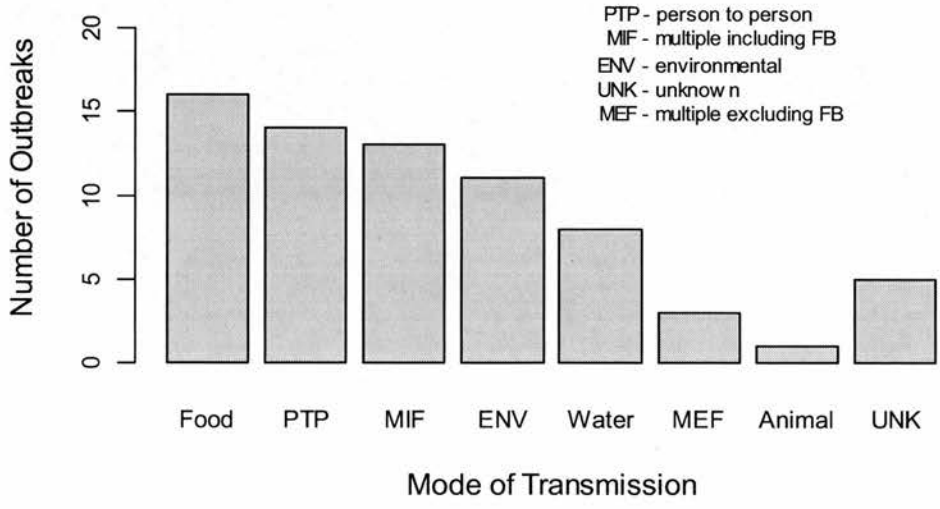


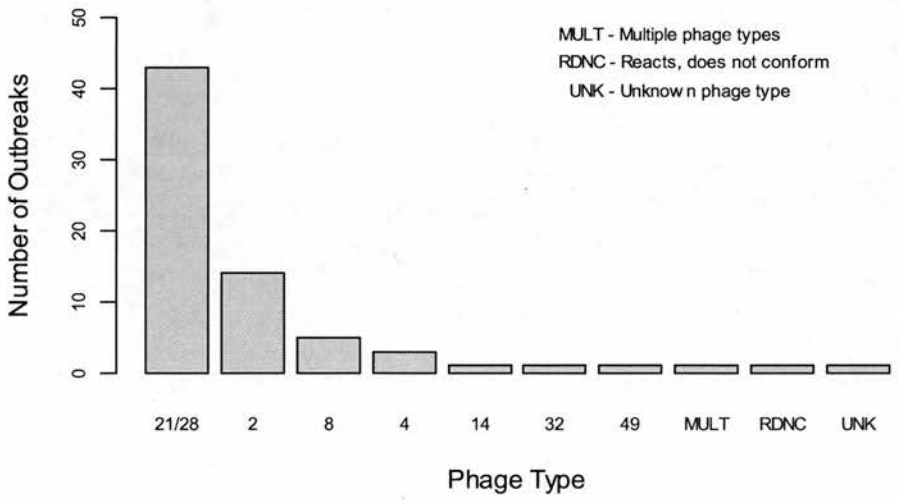
Figure 3.3: Number of outbreaks by mode of transmission (1996 – 2004)
The number of outbreaks by mode of transmission, as defined by HPS.



A putative mode of transmission is listed for sixty-six outbreaks, the putative mode of transmission for four outbreaks is unknown and one outbreak is listed as 'environmental or unknown'. The highest numbers of outbreaks - 16 – were transmitted mainly by food (FB), while 14 were transmitted mainly person to person and 13 by multiple modes including food (Figure 3.3).

Between 1996 and 2004, seven phage types were associated with outbreaks: PT 2, 4, 8, 14, 21/28, 32 and 49. In addition, isolates from one outbreak were classified as RDNC, meaning that they reacted to the phages used in the typing procedure, but the resulting pattern of reactions to the different phages did not conform to the pattern from any known phage type. Phage types 21/28 and 2 were the most common in outbreaks (Figure 3.4), accounting for 62.0% and 19.7% of outbreaks and 38.3% and 53.1% of ill and positive cases, respectively.

Figure 3.4: Number of outbreaks by phage type (1996 – 2004)
The number of outbreaks in Scotland, by phage type. The outbreak with multiple phage types had cases that were phage types 4, 8 and 21/28. In one outbreak, the isolates were not phage typed (unknown) and in another outbreak the patterns of the isolate reactions to the phages did not match any known phage type (RDNC).



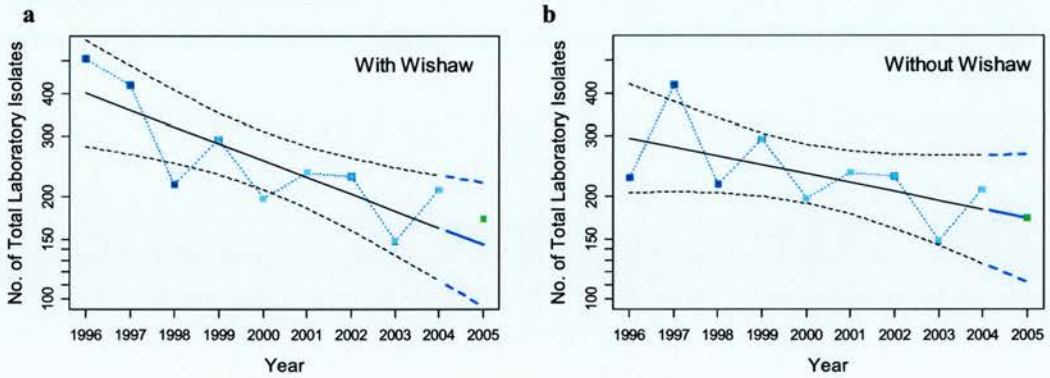
3.4 Results – Statistical analyses

3.4.1 Overall variables

3.4.1.1 Total lab isolates from sporadic and outbreak *E. coli* O157 cases in Scotland, by year

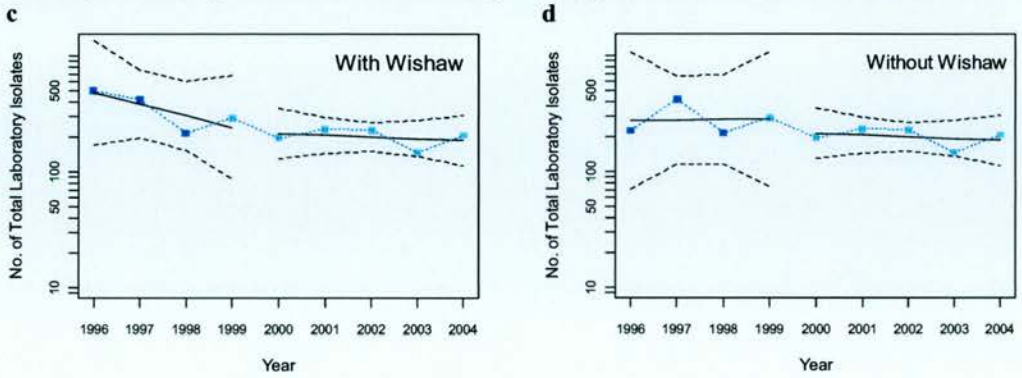
To examine temporal trends in the number of total yearly *E. coli* O157 laboratory isolates, a linear model with a log transformation of the outcome variable is used. There is a statistically significant decreasing trend in the log10 transformed number of total *E. coli* O157 isolates between 1996 – and 2004 (Fig. 3.5a; $F_{1,7}$=12.8, p=0.009). This trend is not statistically significant between 1996 and 2004 when the Wishaw Outbreak isolates are omitted ($F_{1,7}$=3.4, p=0.108; Fig. 3.5b), but is just statistically significant between 1996 and 2005 ($F_{1,8}$= 5.4, p=0.0495).

Figure 3.5 a-b: Number of total laboratory isolates, by year (1996 - 2004)
The number of total laboratory isolates per year between 1996 and 2004 (a) with the Wishaw Outbreak and (b) without the Wishaw Outbreak. In this and all subsequent plots, solid black lines are the best fine line for the regression, with the dashed black lines the 95% confidence intervals. Solid and dashed blue lines are the predicted best fit and 95% confidence interval lines, based on the model for the known data. Years in which enhanced surveillance was in place for some or all months are indicated in light blue, those prior in darker blue. When known, actual data points for future years are indicated in green.



When both a linear and binary year variable with interaction are included in the model, there is no statistically significant difference in the number of isolates or the slopes between 1996-1999 and 2000-2004, whether or not the Wishaw Outbreak cases are included (F<2.2, p>0.199; Fig. 3.5c-d).

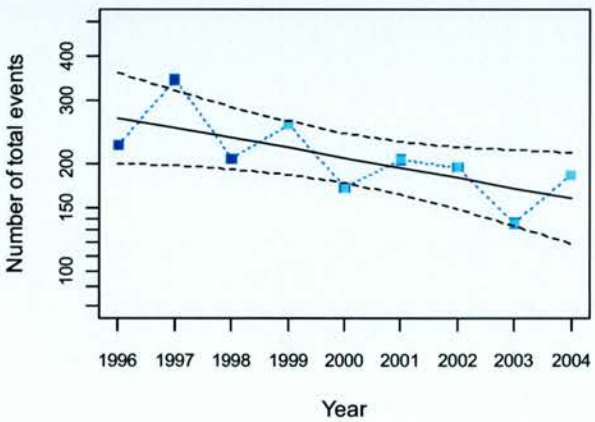
Figure 3.5 c-d: Number of total laboratory isolates, trends before (1996-1999) & after (2000-2004) the onset of enhanced surveillance
The number of total laboratory isolates with the Wishaw isolates (c) and without the Wishaw isolates (d) with the best fit and 95% confidence interval lines for trends before and after the onset of enhanced surveillance. The linear regression models used to fit the lines include the regular year variable, a binomial year variable and a term representing the interaction between the two.



3.4.1.2 Number of *E. coli* O157 events in Scotland, by year

The trend in the log-transformed number of total events, defined as the number of sporadic cases plus the number of outbreaks, can be modelled using a linear model, and there is a statistically significant decline in trend between 1996 and 2004 ($F_{1,7}$=6.2, p<0.042; Fig. 3.6). There is no statistically significant difference in the mean numbers of total events or slopes of the trend in total events before and after the onset of enhanced surveillance (F<0.8, p>0.412).

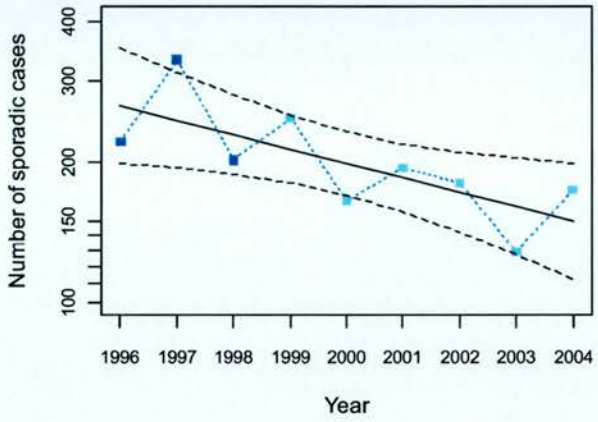
Figure 3.6: Number of total events, by year
The number of total events per year from 1996 to 2004. The black line is the best fit line for the linear model with a log10 transformation of the response variable (total events).



3.4.1.3 Number of sporadic cases

There is a statistically significant decrease in the log transformed number of sporadic cases between 1996 and 2004 ($F_{1,7}$=7.9, p=0.026; Fig. 3.7). However, there is no statistically significant change in the number of sporadic cases or slope of the trends after the onset of enhanced surveillance (p>0.637).

Figure 3.7: Number of sporadic cases, 1996-2004
The number of sporadic cases per year from 1996 to 2004. The best fit line, indicated by the solid black line, is for the linear model with a log10 transformation of the response variable (sporadic cases).

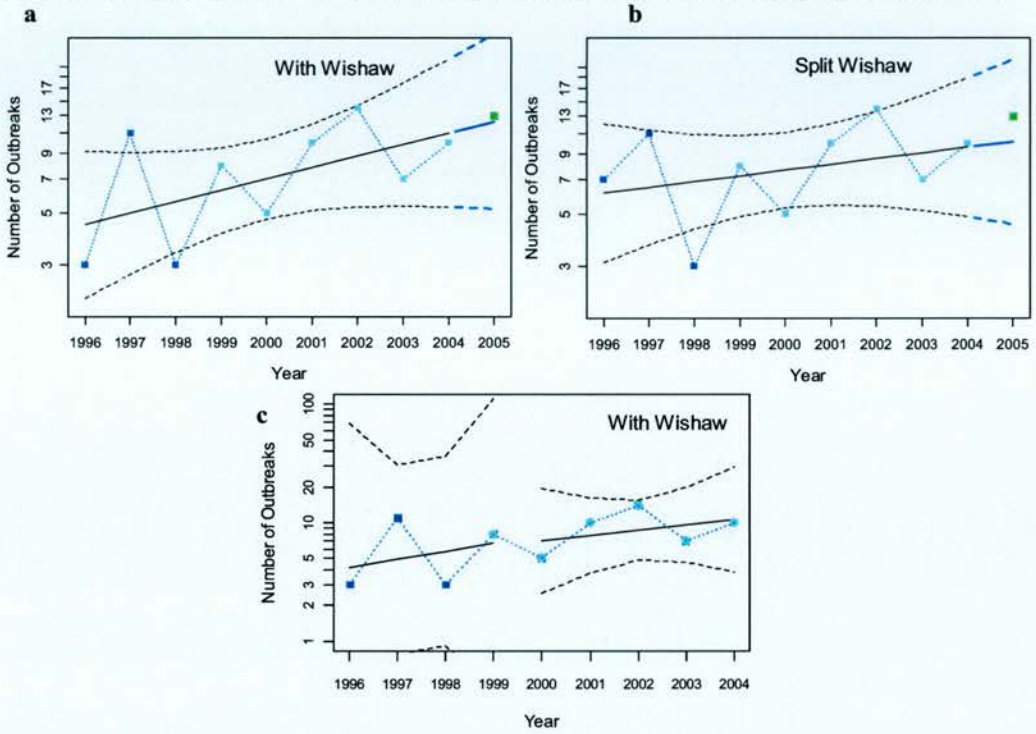


3.4.1.4 Number of *E. coli* O157 outbreaks in Scotland, by year

A linear model can be used to analyse the trend in the log transformed number of outbreaks per year between 1996 and 2004. However, this trend is not statistically significant ($F_{1,7}$=3.1, p=0.124; Fig. 3.8a), and the trend is also not statistically significant when the Wishaw Outbreak was split into five separate outbreaks ($F_{1,7}$=0.9, p=0.382; Fig. 3.8b). The onset of enhanced surveillance does not have a statistically significant effect on the mean number of (p=0.938) or slope of the trend (p=0.856) in outbreaks (Fig. 3.8c). The values predicted by the two models mentioned above for 2005 are very similar and accurate (less than one outbreak from the real values).

Figure 3.8 a-c: Number of *E. coli* O157 outbreaks and trends before and after onset of enhanced surveillance

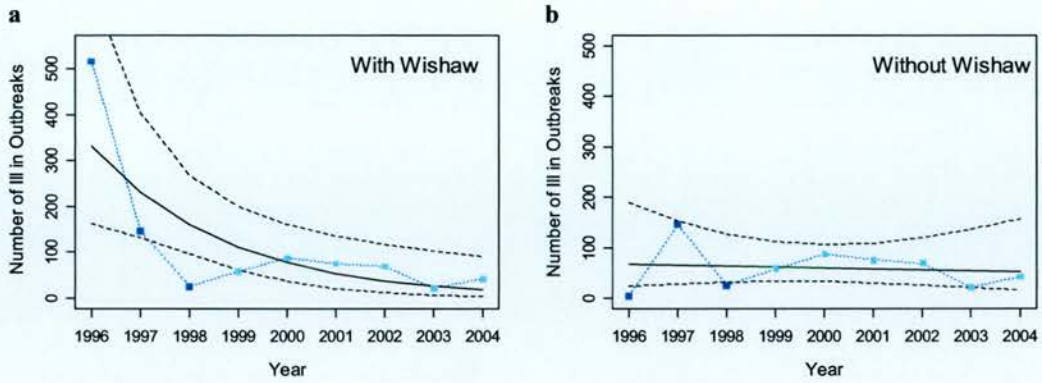
(a) The number of outbreaks per year between 1996 & 2004, with the best fit line for the linear model of the log transformed number of outbreaks. Predictions for the best fit line and 95% confidence interval around that line are shown for the period 2004 to 2005. (b) The number of outbreaks per year between 1996 & 2004 when the Wishaw Outbreak is split into five different outbreaks. (c) The number of outbreaks per year between 1996 & 2004 with the solid black lines indicating separate trends before (1996-99) & after (2000-04) the onset of enhanced surveillance.



3.4.1.5 Number of ill cases from outbreaks, by year

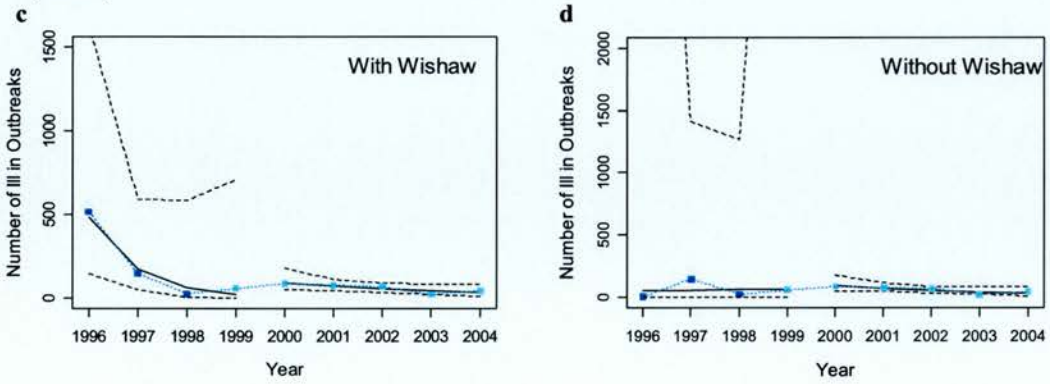
The trends in the number of ill cases from outbreaks are analysed using a Quasipoisson model. There is a statistically significantly decreasing trend ($F_{1,7}$=14.7, p=0.006) in the number of ill cases from outbreaks between 1996 and 2004 (Fig. 3.9a). The trend becomes non-statistically significant when the Wishaw Outbreak cases are excluded ($F_{1,7}$=0.12, p=0.738; Fig.3.9b).

Figure 3.9 a-b: Number of ill cases from outbreaks
The number per year of ill cases in outbreaks between 1996 and 2004 with (a) and without (b) the Wishaw Outbreak Cases included.



In addition, there is a statistically significant interaction between the linear and binomial time variables (Fig. 2.9c), indicating that a statistically significant difference exists between in the slopes of the trends in 1996 -1999 and 2000 – 2004. This difference is barely statistically significant (p=0.044) and is no longer statistically significant when the Wishaw Outbreak cases are excluded (Fig, 2.9d).

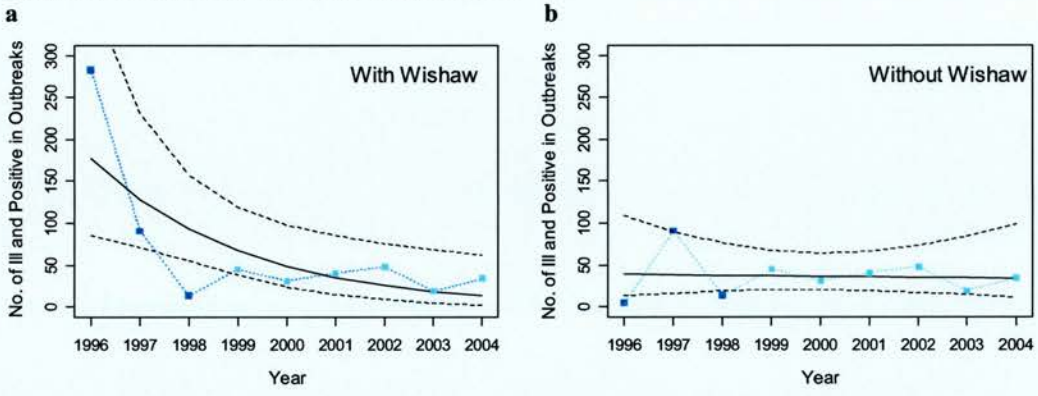
Figure 3.9 c-d: Number of ill cases from outbreaks, trends before & after enhanced surveillance
The number per year of ill cases in outbreaks between 1996 and 2004 with (c) and without (d) the Wishaw Outbreak Cases included with the trends before (1996 -1999) and after (2000-2004) plotted separately.



3.4.1.6 Number of ill and positive [confirmed] cases from outbreaks

The trends for the number of ill and positive cases from outbreaks are very similar to those for the number of ill cases in outbreaks. As with the number of ill from outbreaks, there is a statistically significant downward trend ($F_{1,7}$=11.7, p=0.011) in the number of ill and positive cases from outbreaks when the Wishaw Outbreak is included (Fig. 3.10a), but no statistically significant trend when the Wishaw Outbreak cases are excluded ($F_{1,7}$=0.03, p=0.860; Fig. 3.10b). Also, again there is a barely significant difference in the slopes of the trends between 1996-1999 and 2000 – 2004 (p=0.041) when the Wishaw Outbreak cases are included.

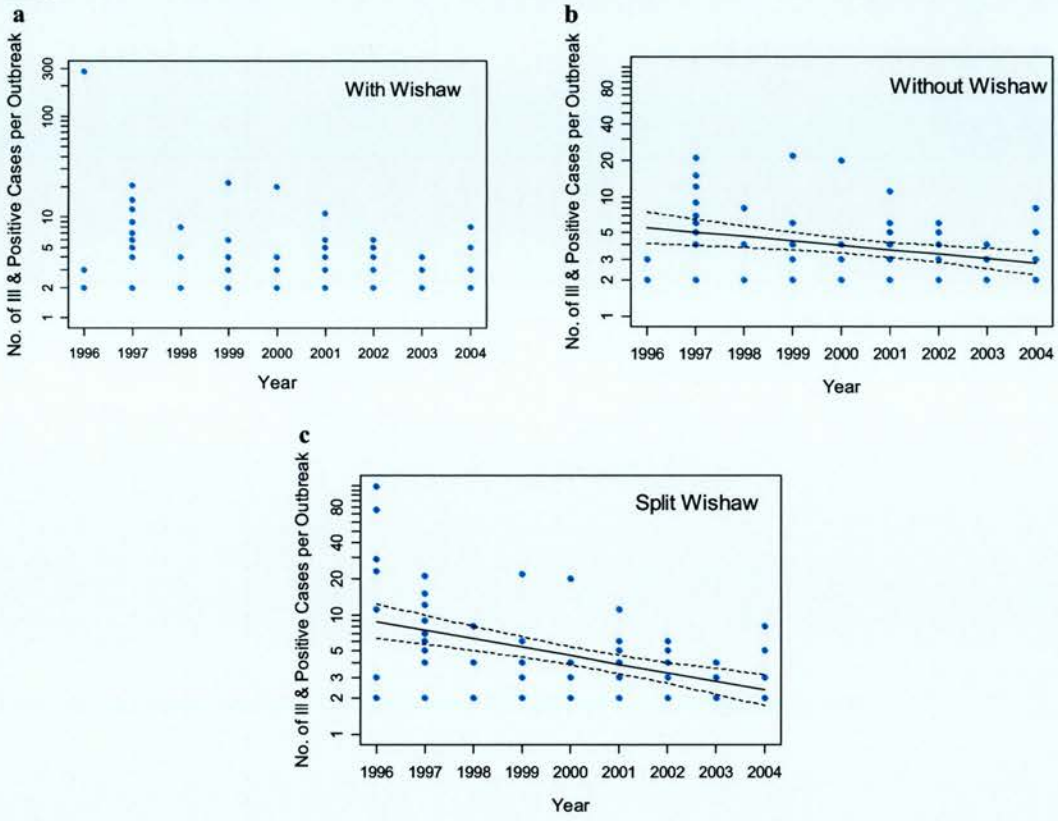
Figure 3.10 a-b: Number of ill and positive cases from outbreaks
The number per year of ill and positive cases from outbreaks between 1996 and 2004 (a) with and (b) without the Wishaw Outbreak Cases included.



3.4.1.7 Number of ill and positive cases per outbreak

When the Wishaw Outbreak is included, the data also cannot be appropriately modelled (Fig. 3.11a), but when the outbreak is excluded and the data is analysed using a linear model, there is a statistically significant decreasing trend in the log-transformed number of ill and positive cases per outbreak ($F_{1,68}$=8.7, p=0.004) between 1996 and 2004 (Fig. 3.11b). As with the number of ill cases, there is also statistically significant downward trend ($F_{1,73}$=25.2, p<0.001; Fig. 3.11c) when the Wishaw Outbreak is considered as five separate outbreaks. There is no evidence (in terms of statistical significance in the analyses used) for any effect of enhanced surveillance when the Wishaw Outbreak is omitted (p>0.669), but there is a statistically significant difference in the slopes between 1996-1999 and 2000-2004 when the Wishaw Outbreak is split into five outbreaks (p=0.006).

Figure 3.11 a-c: Number of ill and positive cases per outbreak
The number of ill and positive cases per outbreak, with each point representing one outbreak. The plots show the data (a) with the Wishaw Outbreak, (b) without the Wishaw Outbreak and (c) with the Wishaw Outbreak split into five separate outbreaks. The best fit lines for the quasipoisson regression of the data are shown in black, with the 95% confidence intervals indicated by dashed lines.

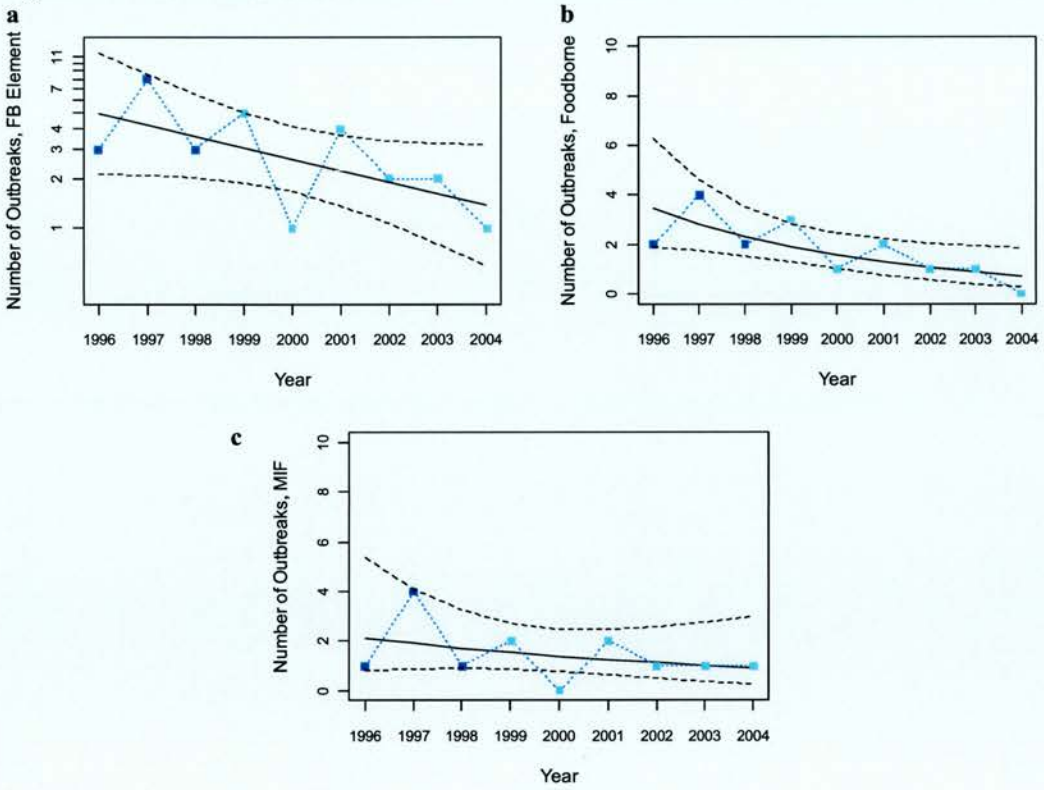


3.4.2 Number of outbreaks – by mode of transmission

There is a statistically significant difference between modes of transmission in the trends in the number of outbreaks ($F_{5,42}$=3.0, p=0.020). Thus the trend for each mode of transmission will be analysed separately.

3.4.2.1 Number of outbreaks – Food related transmission

Figure 3.12 a-c: Number of outbreaks, per year -- foodborne element, foodborne (FB), and multiple modes including foodborne (MIF)
The number of *E. coli* O157 outbreaks for which the mode of transmission is (a) foodborne element, (b) foodborne, and (c) multiple modes including foodborne. For foodborne element the outbreaks are plotted in a log10 scale and the solid line indicates the best fit for the linear regression of the log10 transformed number of outbreaks.

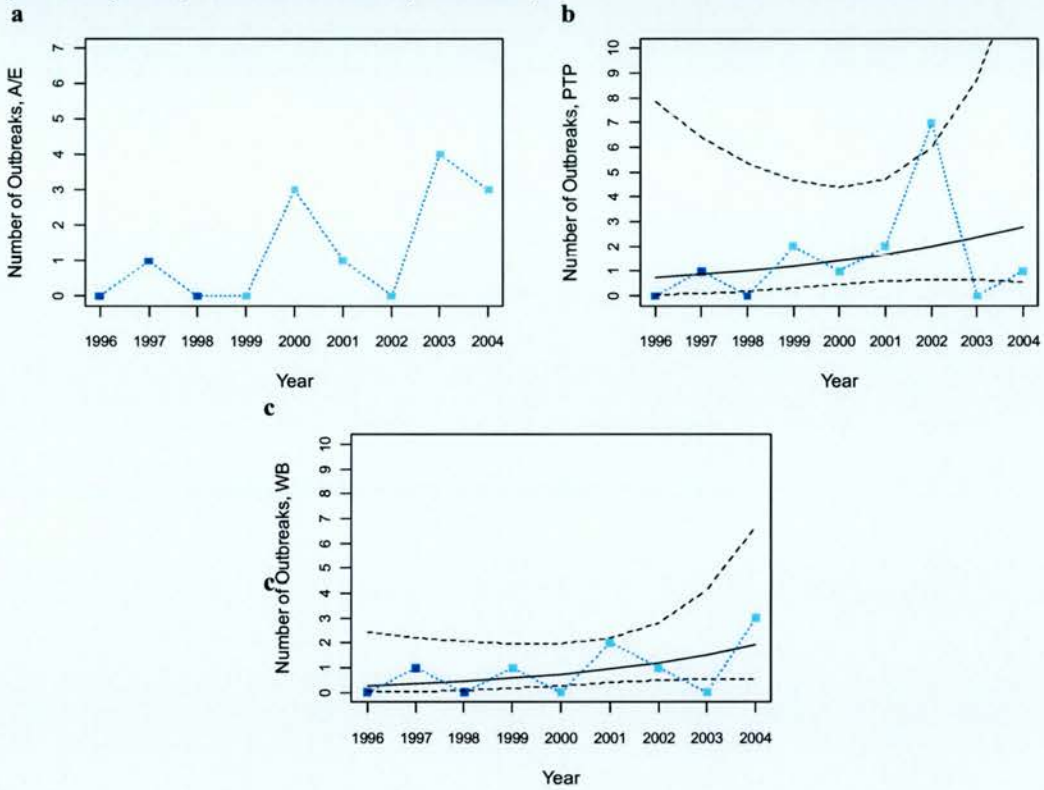


The trend in the number of outbreaks with a foodborne element can be modelled using a linear model, but there is no statistically significant trend in the log10 transformed number of outbreaks with a foodborne element between 1996 and 2004 ($F_{1,7}$=4.6, p=0.070; Fig. 3.12a). However, when outbreaks spread mainly by food, as opposed to those where food is just one of two or more equally important modes of transmission, are examined separately, there is a statistically significant decreasing trend in the number of foodborne outbreaks between 1996 -2004 ($F_{1,7}$=8.7, p=0.021;

Fig. 3.12b). When the outbreaks for which food is one of several transmission modes are examined, there is no statistically significant trend in the number of outbreaks (Fig. 2.12c) between 1996 and 2004 ($F_{1,7}$=1.3, p=0.296).

3.4.2.2 Number of outbreaks – Non-foodborne transmission

Figure 3.13 a-c: Number of outbreaks, per year – Animal/Environmental (A/E), Person to Person (PTP) and Waterborne (WB)
The number of outbreaks in which the mode of transmission is (a) animal/environmental, (b) person to person and (c) waterborne. The best fit lines are for the quasipoisson (animal/environmental and person to person) and Poisson models (waterborne).



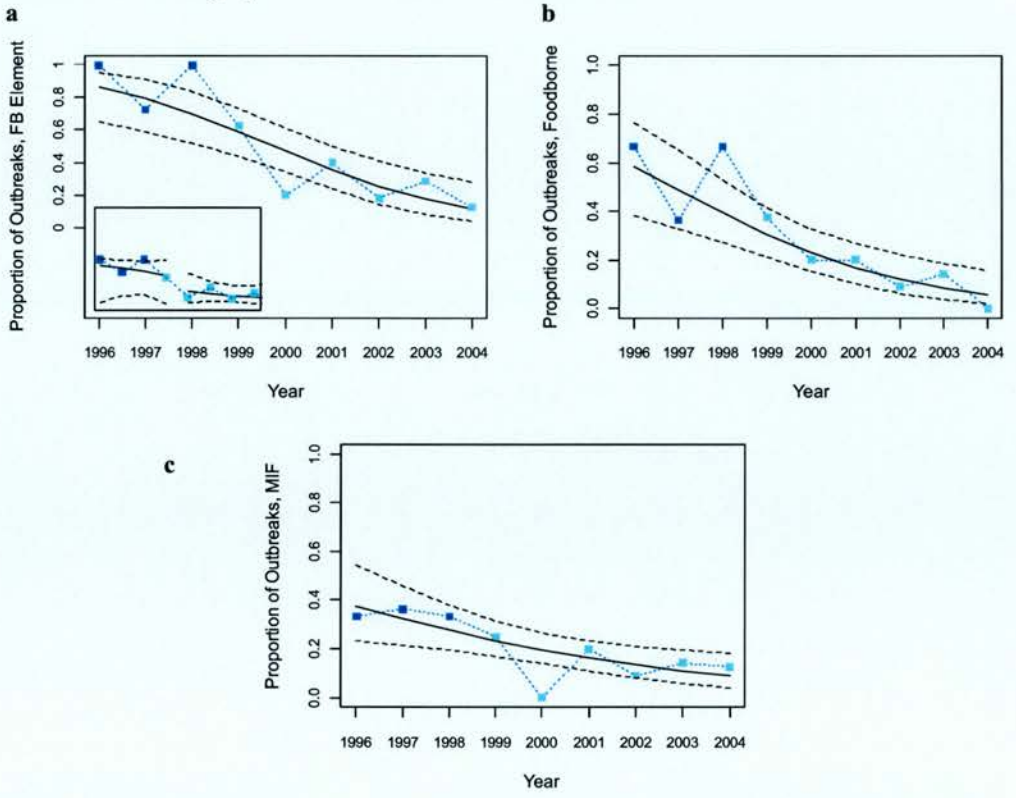
The trend in the number of animal contact or environmental outbreaks between 1996 and 2004 (Fig. 3.13a) cannot be modelled using simple linear models. Using a quasipoisson model, there is no statistically significant trend in the number of person to person outbreaks between 1996 and 2004 ($F_{1,7}$=0.9, p=0.369; Fig. 3.13b). However, there is a statistically significant increasing trend between 1996 and 2002 ($F_{1,5}$=14.6, p=0.012), with only a single person to person outbreak reported in 2003 and 2004 combined. There is no statistically significant trend in the number of waterborne outbreaks between 1996 and 2004 (χ^2=2.80, p=0.094; Fig. 3.13c).

3.4.3 Proportion of outbreaks– by mode of transmission

There is a statistically significant difference between modes of transmission in the temporal trend of the proportions of outbreaks that were spread by each mode of transmission, so the trend in each mode of transmission will be investigated individually.

3.4.3.1 Proportion of outbreaks, food related transmission

Figure 3.14 a-c: Proportion of outbreaks, by year -- foodborne element, foodborne (FB) and multiple modes including foodborne (MIF)
Trends between 1996 and 2004 in the proportion of outbreaks in which the mode of transmission was (a) foodborne element, (b) foodborne and (c) multiple modes including foodborne. In all instances, the model is quasibinomial with the predicted best fit line and 95% confidence interval line around the best fit line shown for the period from 2004 to 2005. The inset in 2.16a shows the trends before (1996 – 1999) and after (2000 – 2004) the onset of enhanced surveillance, with the same axes as the larger plot with 95% confidence intervals indicated.



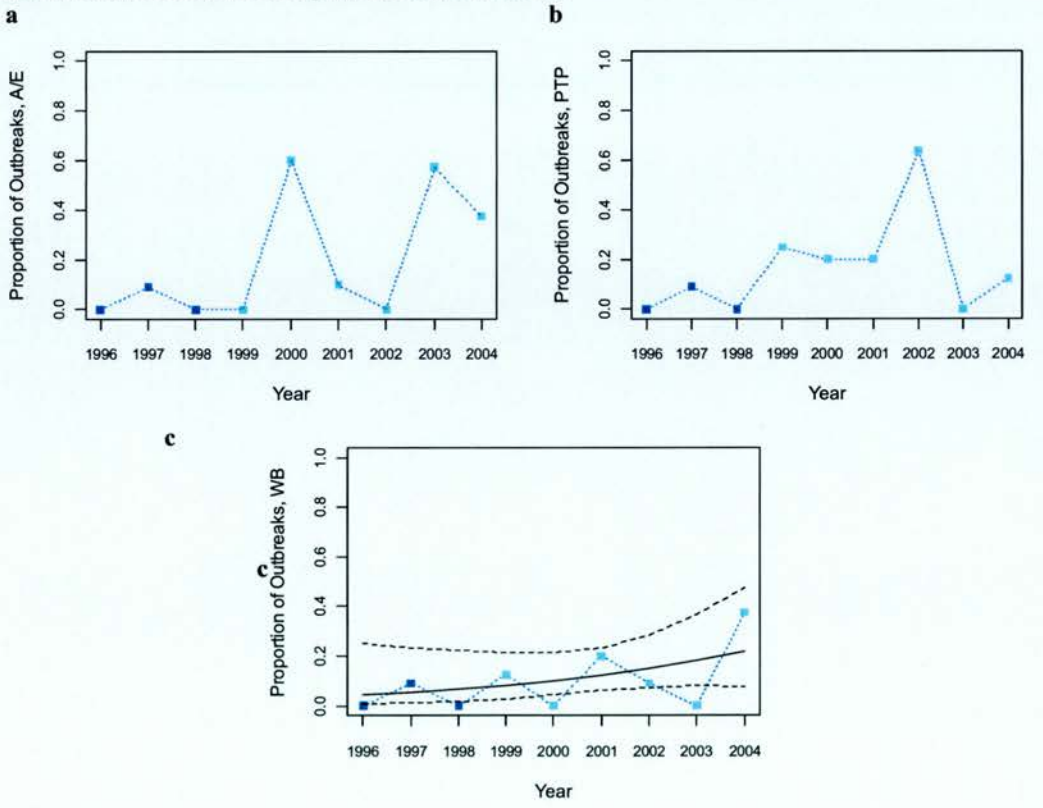
Since 1996, the proportion of outbreaks with food as a main or co-primary mode of transmission has dropped from 1 to less than 0.20 (Fig. 3.14a), and this decreasing trend is statistically significant ($F_{1,7}$=27.3, p=0.001). There are no statistically significant differences in the slopes of the trends (p>0.324, Fig. 3.14a inset) or the means before and after the onset of enhanced surveillance, but there is a noticeable

decrease in the proportion of outbreak that had a foodborne element between 1998 and 2000.

When the outbreaks that are mainly spread by food are examined separately, the proportion of outbreaks that are primarily foodborne decreases statistically significantly between 1996 and 2004 (Fig. 3.14b; $F_{1,7}$=23.5, p=0.002). When both linear and binomial year variables are included in the model, there are no statistically significant differences in the means or slopes of the trends between 1996-1999 and 2000-2004 ($F_{1,5}$=0.7, p=0.434). There is also a statistically significant decreasing trend in the proportion of outbreaks that are spread by multiple modes ($F_{1,7}$=11.9, p=0.011; Fig. 3.14c).

3.4.3.2 Proportion of outbreaks – non food related transmission

Figure 3.15 a-c: Proportion of outbreaks, by year – Environmental/Animal (A/E), Person to Person (PTP) and Waterborne (WB).
Trends between 1996 and 2004 in the proportion of outbreaks in which the mode of transmission is (a) environmental/animal (b) person to person and (c) waterborne. In (c) the model is quasibinomial with the best fit line for the regression shown in black.



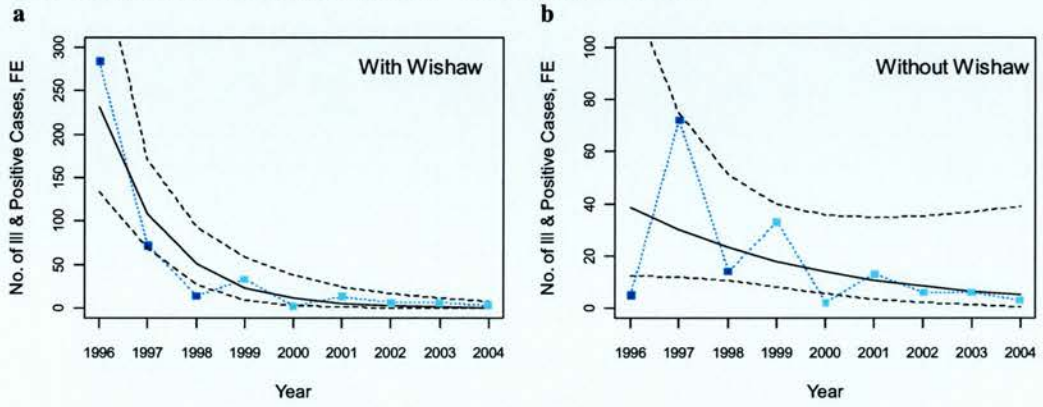
The trends in the proportions of outbreaks that are spread person to person or by animal/environmental exposure (Fig. 3.15a, b) cannot be modelled using simple linear models, but the trend in the proportion of outbreaks that are spread via water can be modelled using generalised linear models with quasibinomial error structure. There is no statistically significant trend in the proportion of outbreaks spread by water ($F_{1,7}$=2.7, p=0.143; Fig. 3.15c) transmission between 1996 and 2004.

3.4.4 Number of ill and positive cases from outbreaks – by mode of transmission

As with the trend in the number of ill cases, there is no statistically significant difference between modes of transmission in the trend of the number of ill and positive cases when the Wishaw Outbreak cases are excluded ($F_{5,42}$=2.2, p=0.077), but there is a statistically significant difference if the Wishaw cases are included ($F_{5,42}$=12.7, p<0.001). Thus, the trends without the Wishaw cases will be analysed separately, with the trends including the Wishaw cases shown for the purposes of comparison.

3.4.4.1 Number of ill and positive cases from outbreaks – food-related transmission

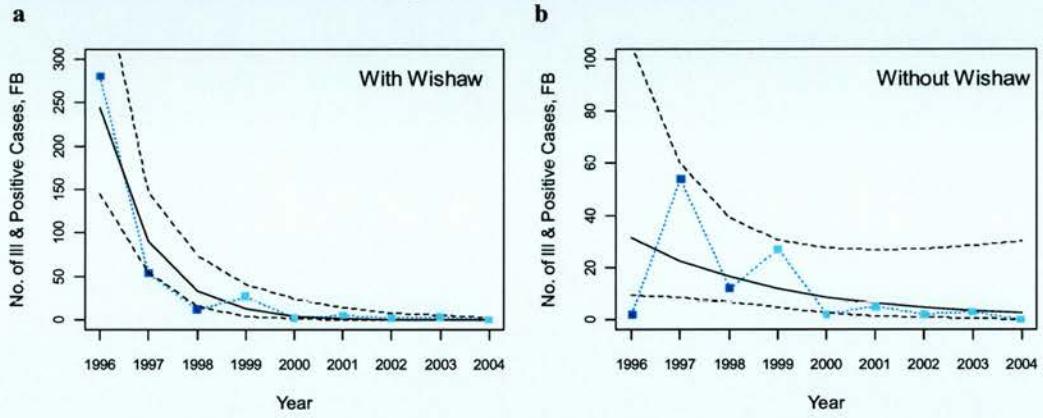
Figure 3.16 a-b: Number of ill and positive cases in outbreaks with a foodborne element
The number of ill and positive cases in outbreaks with a foodborne element (FE) between 1996 and 2004 with (a) and without (b) the Wishaw Outbreak cases included.



With a quasipoisson model, there is a statistically significant downward trend ($F_{1,7}$=49.0; p<0.001) in the number of ill and positive cases from outbreaks with a foodborne element between 1996 and 2004 (Fig. 3.16a). However, once the Wishaw Outbreak is excluded, the trends in the number of ill and positive cases are not statistically significant ($F_{1,7}$=3.6, p=0.101; Fig, 3.16b).

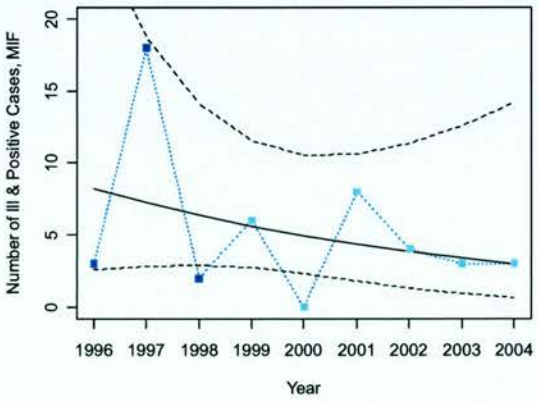
When both the continuous and binomial year variables and an interaction term are inserted in the model, there are no statistically significant differences in the slopes or means of the number ill and positive cases before and after the start of enhanced surveillance ($p>0.133$).

Figure 3.17 a-b: Number of ill and positive cases from foodborne outbreaks
The number of ill and positive cases in foodborne (FB) outbreaks between 1996 and 2004 with (a) and without (b) the Wishaw Outbreak cases included. Predicted best fit lines and 95% confidence intervals for the line are shown for the period between 2004 and 2005.



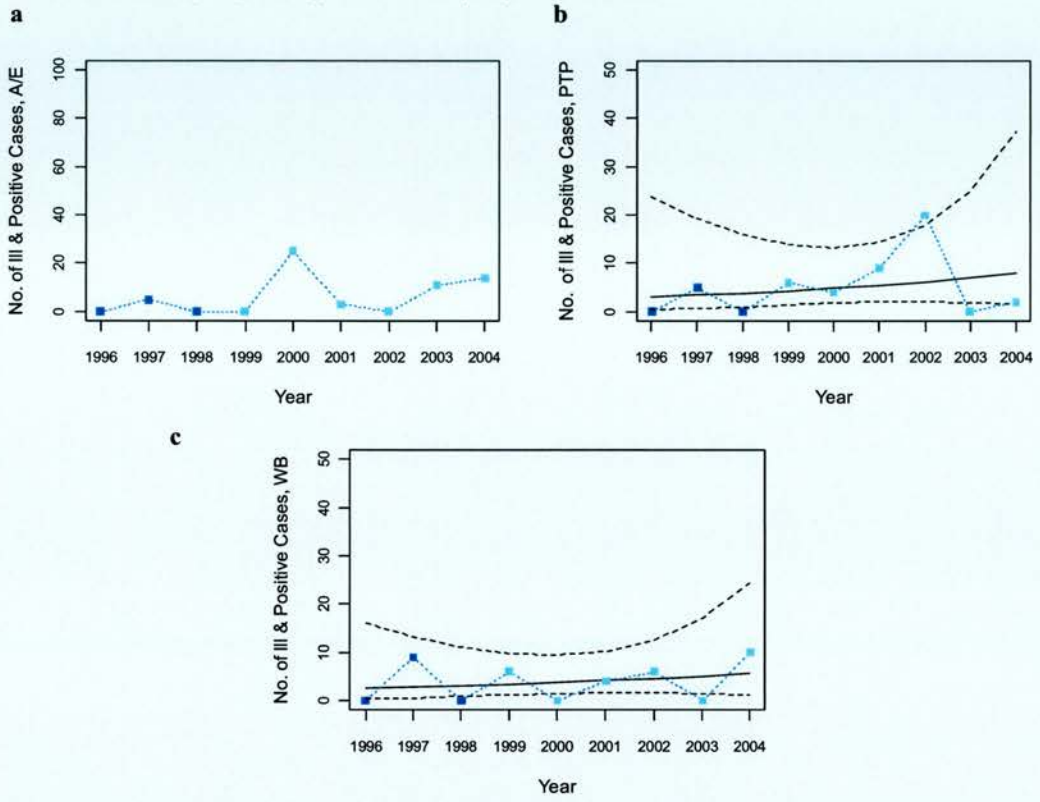
Using a quasipoisson model, there is a statistically significant decreasing trend in the number of ill and positive cases from foodborne outbreaks between 1996 and 2004 ($F_{1,7}=64.5$, $p<0.001$; Fig. 3.17a). The trend becomes non statistically significant when the Wishaw Outbreak cases are omitted from the data set ($F_{1,7}=4.3$, $p=0.077$; Fig. 3.17b). The trend in the number of ill and positive cases from outbreaks spread by multiple modes of transmission including food is not statistically significant ($F_{1,7}=1.1$, $p=0.323$, Fig. 3.18).

Figure 3.18: Number of ill and positive cases from outbreaks spread by multiple modes including food
The number of ill and positive cases in outbreaks spread by multiple modes of transmission, one of which is food (MIF).



3.4.4.2 Number of ill and positive cases from outbreaks – non food related transmission

Figure 3.19 a-c: Number of ill and positive cases from outbreaks – environmental, person to person and waterborne transmission
The number of ill and positive cases from outbreaks spread via (a) animal/environmental (A/E), (b) Person to Person (PTP) and (c) waterborne (WB) transmission.



The trends in the number of cases from animal/environmental outbreaks cannot be appropriately modelled using simple linear models (Fig. 3.19a). When quasipoisson models are used, the trends between 1996 and 2004 in the numbers of ill and positive cases from outbreaks spread through person to person and waterborne transmission are increasing, but not statistically significant ($F_{1,7}<0.6$, $p>0.476$; Fig. 3.19b,c).

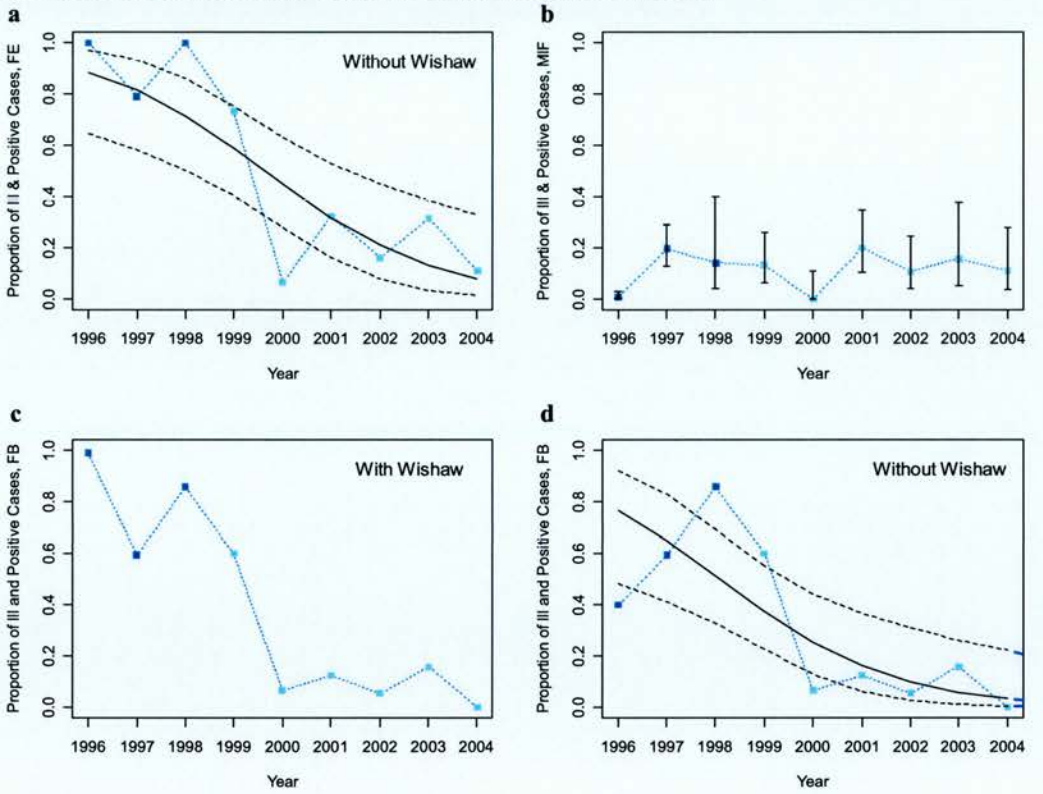
3.4.5 Proportion of ill and positive cases from outbreaks, by mode of transmission

There is a statistically significant difference in the trends of the proportion of ill and positive cases between modes of transmission when the Wishaw Outbreak cases

were excluded ($F_{5,42}$=4.7, p=0.002). Again, there is no adequate simple linear model when the Wishaw Outbreak cases are included.

3.4.5.1 Proportion of ill and positive cases from outbreaks, food related transmission

Figure 3.20 a-d: Proportion of ill and positive cases from outbreaks with a foodborne element and outbreaks with multiple modes of transmission including foodborne
The proportion of ill and positive cases from outbreaks (a) with a foodborne element (FE) excluding the Wishaw Outbreak cases (b) spread via multiple modes of transmission, one of which is foodborne (MIF), (c) spread via food with (d) and without the Wishaw Outbreak cases (FB). 95% confidence intervals are indicated for the trends which are modelled.



The trend in the proportion of ill and positive cases from outbreaks spread by a foodborne element when the Wishaw cases were included could not be modelled described using GLMs with the error structures specified in section 2.3.4, but when the Wishaw Cases were excluded, there was a statistically significant decrease in the proportion of ill and positive cases that were from an outbreak with a foodborne element ($F_{1,7}$=19.5, p=0.003; Fig, 3.20a). In addition, the trend in the proportion of ill and positive cases that were outbreaks spread by multiple methods including food (Fig. 3.20b) could not be modelled. The trend in foodborne outbreaks could not be

modelled if the ill and positive cases from the Wishaw Outbreak were included (Fig. 3.20c). However, if the cases are excluded, there is a statistically significant decrease in the proportion of ill and positive cases that are from foodborne outbreaks ($F_{1,7}$=18.2, p=0.004; Fig. 3.20d).

Where the trends cannot be modelled using GLMs with the specified error structures, visual inspection indicates that there is an apparent shift in the proportion of ill and positive from foodborne outbreaks or those with a foodborne element (Fig. 3.20a,c) . This shift occurs between 1999 and 2000, the juncture at which enhanced surveillance was started. From 1996 to 1998, the complete years in which there was no enhanced surveillance, more than 0.79 (95% CI = 0.70, 0.86) and 0.59 (0.49, 0.69) of ill and positive cases in outbreaks are from foodborne element and foodborne outbreaks, respectively. Since the onset of enhanced surveillance, outbreaks with a foodborne element have accounted for no more than 0.33 (0.20, 0.48) of ill and positive and outbreaks spread by food, no more than 0.16 (0.06, 0.38).

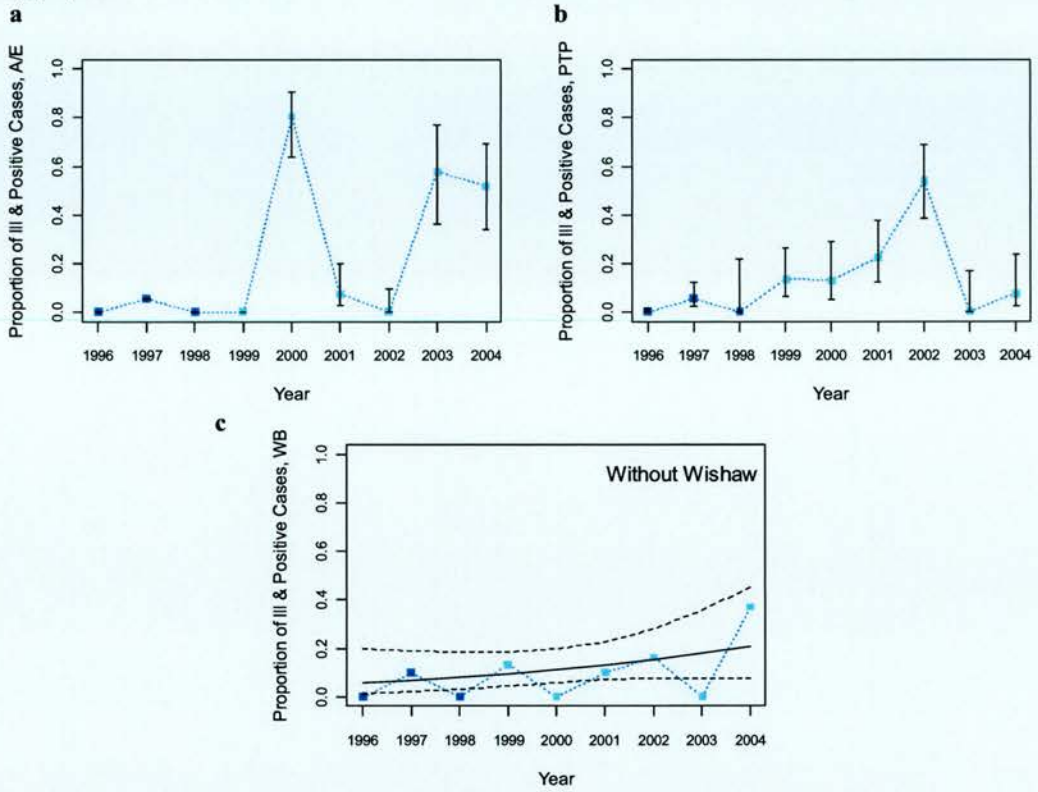
Even when the Wishaw Outbreak is excluded, the proportion of outbreak ill and positive cases that were foodborne is still 0.40 (0.12, 0.77), far higher than any proportion from 2000 onwards and in both instances the decrease in the trend is statistically significant (p<0.001).

3.4.5.2 Proportion of ill and positive cases from outbreaks, non food related transmission: environmental/animal, person to person and waterborne

The trends in the proportion of ill and positive cases in outbreaks that are animal/environmental (Fig. 3.21a) and person to person (Fig. 3.21b) could not be modelled using GLMs and the error structures specified in Section 2.3.4. If the Wishaw Outbreak cases are omitted, the trend in waterborne cases can be modelled using a quasibinomial error structure, but the trend is not statistically significant ($F_{1,7}$=3.5, p=0.104; Fig. 3.21c).

Figure 3.21 a-c: Proportion of ill and positive cases from environmental/animal, person to person and waterborne outbreaks

The proportion of ill and positive cases from outbreaks spread via (a) environmental/animal, (b) person to person and (c) waterborne transmission. 95% confidence intervals are indicated where appropriate.

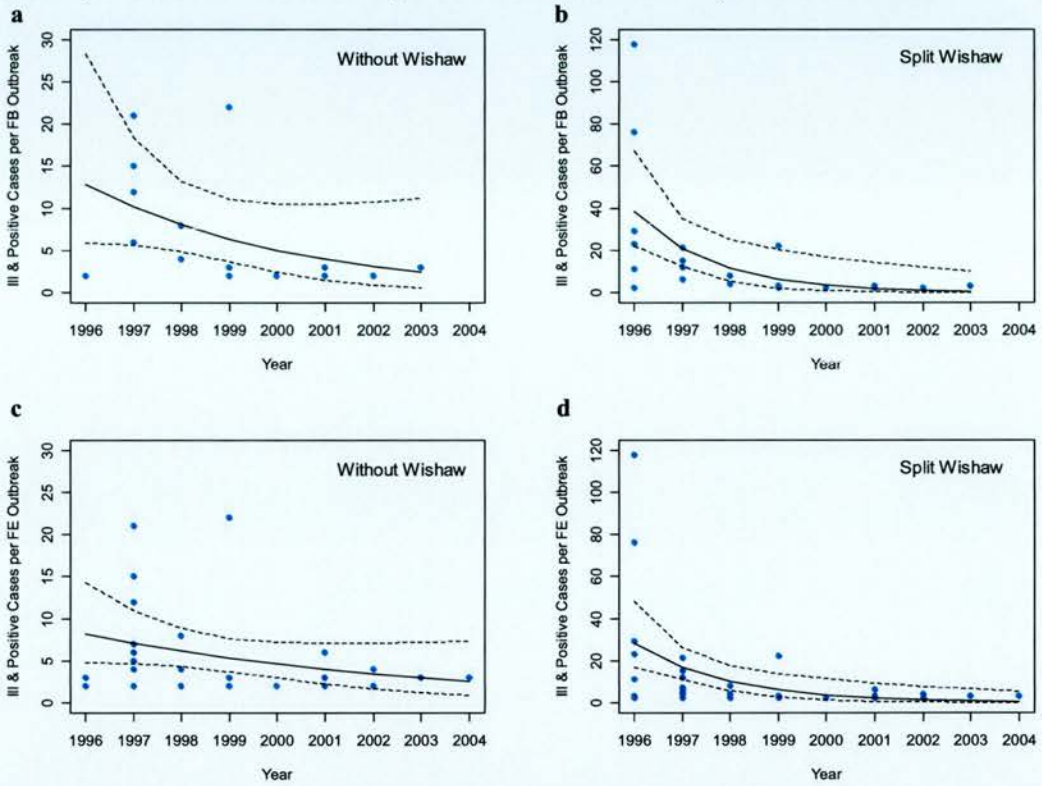


3.4.6 Number of ill and positive cases per outbreak, by mode of transmission

No statistically significant difference in the trends in the number of ill and positive cases per outbreak exists between modes of transmission ($F_{6,56}$=0.8, p=0.560). However, the trends in foodborne element and foodborne outbreaks will be analysed separately because they are of particular interest in this thesis.

3.4.6.1 Ill and positive cases per foodborne/foodborne element outbreak

Figure 3.22 a-b: Ill and positive cases per foodborne outbreak and per outbreak with a foodborne element
The number of ill and positive cases per foodborne outbreak (a), per outbreak with a foodborne element (b), per foodborne outbreak with the Wishaw Outbreak split into five outbreaks (c) and per outbreak with a foodborne element when the Wishaw Outbreak was split (d). The solid line indicates the best fit line for the log transformed linear model. There were no foodborne outbreaks in 2004, so the line should does not represent a model based on the input of the 2004 data.



As with the analysis of the overall ill and positive per outbreak, the Wishaw Outbreak appears to be an outlier, and thus the data cannot be appropriately modelled when the Wishaw data is included. When the Wishaw Outbreak data point is excluded, there is a non-statistically significant decrease number of ill and positive per foodborne outbreak ($F_{1,13}$=3.5, p=0.085; Figure 3.22a). If the Wishaw Outbreak is considered as five separate outbreaks, there is a statistically significant downward trend in the log10 number of ill and positive cases per outbreak ($F_{1,18}$=14.1, p=0.001; Fig. 3.22b).

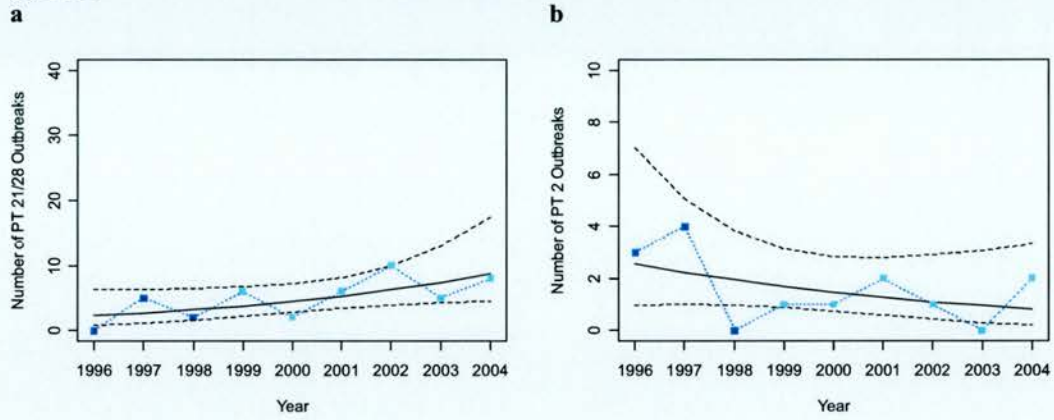
When all outbreaks with a foodborne element are considered without the Wishaw Outbreak, the trend in the number of ill and positive is still decreasing, but again the

change is not statistically significant ($F_{1,26}$=3.2, p=0.087; Fig. 3.22c), but when the Wishaw Outbreak is split into five outbreaks, the decrease in the trend is statistically significant ($F_{1,31}$=15.8, p<0.001; Fig. 3.22d).

3.4.7 Trends in phage types: PT 21/28 and PT

3.4.7.1 Number of outbreaks, PT 21/28 and PT 2

Figure 3.23 a-b: Number of outbreaks -- phage type 21/28 and phage type 2
The number of outbreaks per year between 1996 and 2004 in which isolates are (a) PT 21/28 and (b) PT 2.

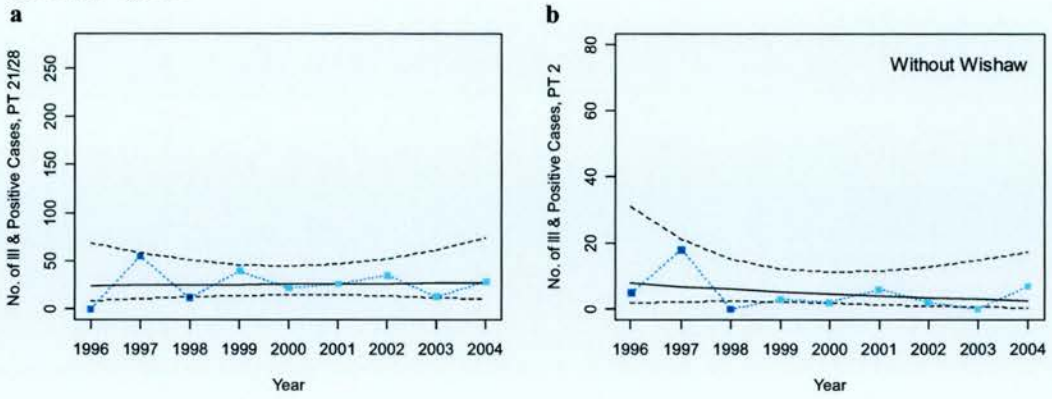


When the trends are analysed using a quasipoisson model, the results indicated that there has not been a statistically significant increase in the number of PT 21/28 outbreaks between 1996 and 2004 ($F_{1,7}$=5.3, p=0.055; Fig. 3.23a). There is not a statistically significant difference between 1996-1999 and 2000-2004 in the means or slopes of the trends (p>0.496). There is not a statistically significant decline in the number of PT 2 outbreaks between 1996 and 2004 ($F_{1,7}$=1.9, p=0.215; Fig. 3.23b).

3.4.7.2 Number of ill and positive cases – PT 21/28 and PT 2 outbreaks

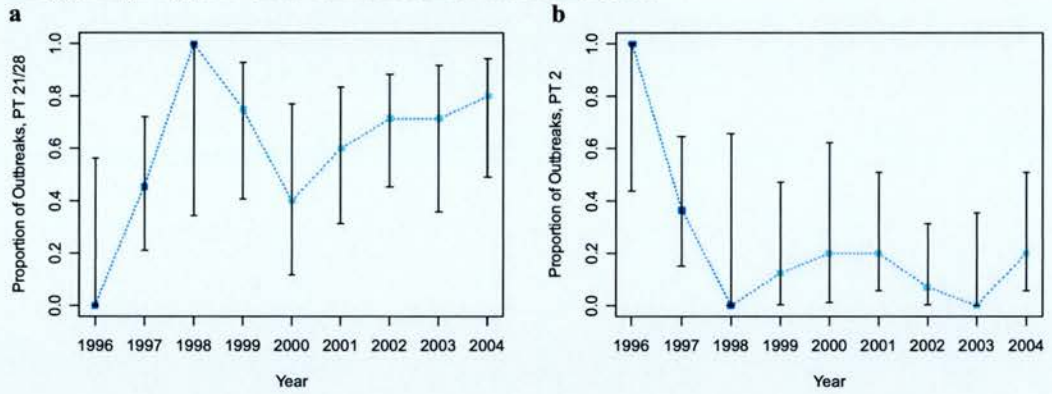
There is not a statistically significant increase or decrease in the number of ill and positive cases from outbreaks with PT 21/28 between 1996 and 2004 ($F_{1,7}$=0.03, p=0.879; Fig. 3.24a). The trend for PT 2 is not statistically significant with or without Wishaw between 1996 and 2004 (p>0.267; Fig. 3.24b).

Figure 3.24 a-b: Number of ill and positive cases in outbreaks – PT 21/28 and PT 2
The number of ill and positive cases from outbreaks where isolates were (a) Phage Type 21/28 and (b) Phage Type 2



3.4.7.3 Proportion of total outbreaks, PT 21/28 and PT 2

Figure 3.25 a-b: The proportion of total outbreaks, PT 21/28 and PT 2
The proportion of outbreaks of a known phage type in which the isolates are (a) Phage Type 21/28 and (b) Phage Type 2. 95% confidence intervals are indicated.

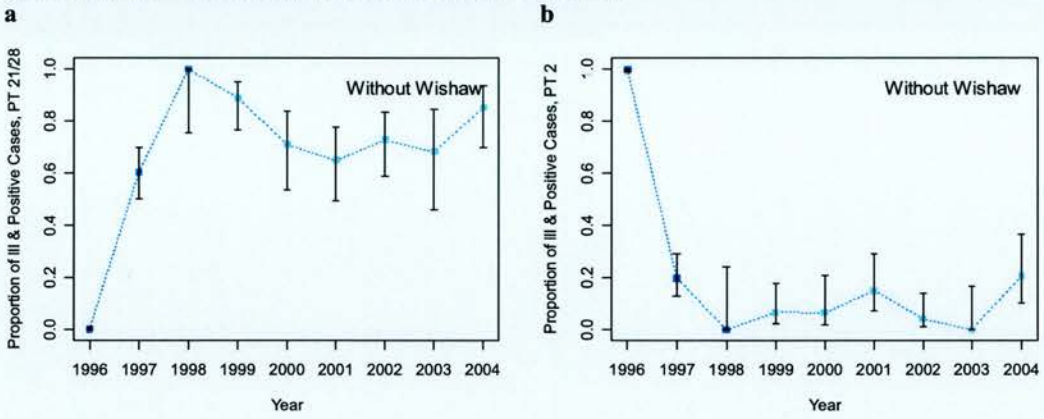


The proportion of outbreaks that are PT 2 appears to have increased (Fig. 3.25a), while in contrast, the proportion of outbreaks that are PT 2 appears to have decreased (Fig. 3.25b), but the trends cannot be appropriately modelled using GLMs and the error structures specified in Section 2.3.4.

3.4.7.4 Proportion of total ill and positive outbreak cases, PT 21/28 and PT 2

The proportion of ill and positive cases from outbreaks with a known phage type for both PT 21/28 and PT 2 cannot be appropriately modelled GLMs and the error structures specified in Section 2.3.4 (Fig. 3.26a, b). A visual inspection of the data suggests, however, that the trends are similar to the trends seen in the proportion of total outbreaks.

Figure 3.26 a-b: The proportion of ill and positive cases from outbreaks - PT 21/28 and PT 2
The proportion of ill and positive cases with a known phage type from outbreaks that are (a) PT 21/28 and (b) PT 2. 95% confidence intervals are indicated.

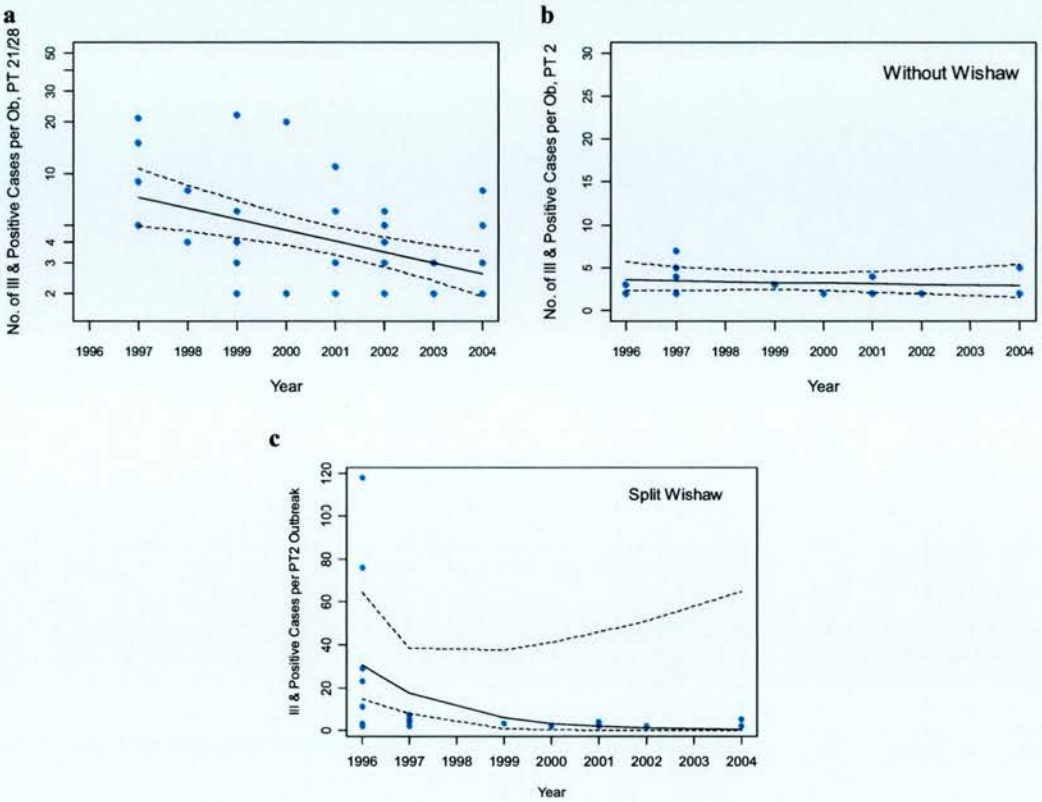


3.4.7.5 Ill and positive cases per outbreak, PT 21/28 and PT 2

There are geometric means of 3.98 ill and positive cases (95% CI = 3.23, 4.90) per PT 21/28 outbreak and 4.15 ill and positive cases (1.98, 8.72) per PT 2 outbreak. No statistically significant difference exists between the two phage types in the log10 number of ill and positive cases per outbreak (Wilcoxon Rank Sum p=0.450), even when the Wishaw Outbreak is excluded (p=0.222) or split into five outbreaks (p=0.506). When a binomial logistic model is used, outbreaks are statistically significantly more likely to be PT 21/28 than PT 2 over time with (z=2.44, p=0.015) or without the Wishaw Outbreak cases (z=2.16, p=0.031).

When a linear model is used, there is a strongly significant decreasing trend in the log transformed size of outbreaks where most or all of the ill and positive cases were PT 21/28 ($F_{1,42}$=13.2, p<0.001; Fig. 3.27a). The data for PT 2 can be modelled without the Wishaw Outbreak cases, and the trend is not statistically significant (p>0.323; Fig. 3.27b). However, after breaking down the Wishaw Outbreak into five outbreaks, there is a statistically significant (p=0.017; Fig. 3.27c) downward trend.

Figure 3.27 a-c: Number of ill and positive cases per outbreak -- PT 21/28 and PT 2
The number of ill and positive cases per outbreak in which the cases are (a) PT 21/28, (b) PT 2 without the Wishaw Outbreak and (c) PT 2 with the Wishaw Outbreak split into five outbreaks. The best fit and 95% confidence interval lines are indicated for each model.



3.5 Discussion

As discussed in Section 1.8 temporal trends have been studied in a wide range of medical conditions including cancer, injury rates and infectious diseases. However, within the scope of *E. coli* O157 research, studies relating to temporal trends have been limited to seasonal trends and descriptive statistics, with little if any published statistical analyses of longer term trends, in Scotland or elsewhere. A statistical analysis of such trends in infection numbers and characteristics is of relevance and importance given the increased focus on identifying appropriate and cost effective prevention and treatment strategies. This chapter was conducted to, where appropriate, model trends in Scottish *E. coli* O157 cases and outbreaks using simple linear models, and to describe these trends. Equally as important was the

identification of trends that could not be statistically modelled and the reasons preventing the creation of appropriate models.

3.5.1 Issues regarding statistical analysis

3.5.1.1 Type of models

For the *E. coli* O157 outbreak trends in Scotland, models with Poisson error structures were most commonly found to be appropriate. Firstly, as discussed in Chapter 2.3.2, Poisson models have long been used in describing infectious disease/ecological trends because they are often the most appropriate for the analysis of data reported in terms of counts (Crawley 2002). Counts, such as the number or cases or outbreaks, are one the most common type of infectious disease data. One important quality of these models, as related to count data, is that by use of the log link function, the model rules out the possibility of obtaining negative fitted values – negative counts being impossible in disease surveillance.

The Poisson error structure assumes a k, the reverse measure of aggregation, of infinity (i.e. random distribution), which is not the case in most of the analyses. However, any over or underdispersion in the sample can be adjusted for by the use of the dispersion factor in the quasi functions, as described in McCullagh and Nelder (McCullagh & Nelder 1989). Thus Poisson error structures are a flexible approach to modelling the trends. Indeed Poisson regression models have been used in many contexts to examine temporal trends in infectious diseases, infections and cancer (Agha et al. 2006; Conway et al. 2006; Kamper-Jorgensen et al. 2006; Maule et al. 2006; Tleyjeh et al. 2005). However, for some infectious disease trends, like that of the number of outbreaks, the increase is log linear – i.e. the rate of increase is not constant over time and. Thus, linear regression, which has also been used in the context of infectious disease trends (Menzies et al. 1994; Miller et al. 2004a), is more appropriate. The binomial error structure was the only appropriate error structure for the data presented as proportions constructed out of counts (e.g. proportion of total outbreaks or ill cases that were a particular mode of transmission) because it allows for data with strict bounds. Additionally, the logit link allows for fitted values with asymptotes at 0 and 1. Thus, a model can be appropriately and easily interpreted when analysing proportions (Crawley 2002). Therefore, for the

purposes of this thesis, generalised linear models with Poisson and binomial error structures and linear models with log transformation of the data proved to be the most appropriate and effective methods for modelling some of the *E. coli* O157 temporal trends in Scotland.

3.5.1.2 Multiple testing

As discussed in Chapter 2.3.5.1, with the large number of tests of statistical significance being performed, there is a risk for statistically significance to occur by chance. With approximately, 65 tests of significance being run on the Scottish data, one would expect 3 or 4 tests to be statistically significant even if the trend is not significant. Thus, caution must be taking in making inferences purely on the statistical results, particularly when dealing with counts of outbreaks where statistical power to detect a trend may be quite low.

3.5.1.3 Wishaw Outbreak and enhanced surveillance

Additional factors which had to be considered were the Wishaw Outbreak in 1996, which involved a disproportionately high number of cases compared to any other outbreaks included in this study (as suggested by its being an outlier in many of the trend analyses presented in this chapter), as well as the onset of enhanced surveillance in 1999/2000. The effect of these factors was assessed by analysing case by case data with and without the Wishaw Outbreak, by splitting the Wishaw Outbreak into several outbreaks by established cohort and by analysing trends before and after the onset of enhanced surveillance.

3.5.2 Where models could not be fitted

This study has established that basic linear regression techniques can be used to analyse some of the yearly temporal trends in Scottish *E. coli* O157 cases and outbreaks. The cases in which trends could not be appropriately modelled could be divided into a few distinct categories which highlighted some of the above mentioned difficulties with the data, as well as issues discussed in Chapter 2. For several situations: the number of environmental outbreaks, the proportion of outbreaks that were environmental, PT2 and PT 21/28, the number of ill and positive cases in environmental outbreaks, and the proportion of total cases that were

environmental, PT2 and PT21/28, the inability of the data to be modelled was directly linked to a lack of data. With only nine (1996 – 2004) data points and often a great variability in the number of outbreaks between years, there was not enough power for meaningful regression analyses. For instance, simulations based on power size calculations from Dupont and Plummer (Dupont & Plummer 1998) and run using nQuery Advisor (Elashoff 2005), indicated that analyses would only have 49% power to detect changes in slope equivalent to one third of the standard deviation of the residuals around the regression line. Additionally analyses run using the TRENDS software packaged, and presented in Section 2.3.5.2, indicated that analyses of trends with low counts and high variation, such as the trend in the number of foodborne cases would have low power. For instance, such an analysis would only have about 50% power to detect a 25% decline in the number of foodborne outbreaks from year to year. The lack of data was also a problem in assessing the influence of the onset of enhanced surveillance because changes in trend slopes that looked of importance were almost always not statistically significant. With no more than five data points on either side of the switch to enhanced surveillance, the lack of statistical significance may be a result more of insufficient power rather than an absence of changing trends. A more statistically powerful technique for measuring the change of trends would be of use in ascertaining the real influence of enhanced surveillance, such as using monthly or weekly data points to increase the power of the analysis. However, data was not available for this thesis at a weekly or monthly level (see Section 2.4.2).

One of the patterns of note in the analyses was the frequent inability to model data involving case numbers when the Wishaw Outbreak was involved. This characteristic, as well as those discussed below, suggests that the Wishaw Outbreak is an outlier, an unusual outbreak that does not fit within the normal pattern in Scotland and may need to be excluded when fitting models. However, the Wishaw Outbreak could also just be at the upper end of the normal size distribution for Scottish outbreaks.

In addition, there is potential evidence that it maybe more appropriate to consider the Wishaw Outbreak as a series of outbreaks rather than one very large outbreak.

Firstly, the trends in the number of ill and positive cases per outbreak (overall, foodborne and foodborne element) could be adequately modelled when the Wishaw Outbreak was split into five outbreaks based on the cohorts indicated in Cowden et al (Cowden et al. 2001), but not when it was included as a single outbreak. Also the trend data was a better fit for the split models than models without the Wishaw cases.

The five cohorts were distinct, "specific groups of people at risk" (Cowden et al. 2001), and were delineated by location where the meat was consumed or purchased: 1) nursing home, 2) pub party, 3) church luncheon, 4) purchased meat from suspect butcher and 5) purchased meat from a premise supplied by the suspect butcher. The idea of Wishaw being compromised of smaller incidents is hinted at in Pennington (Pennington 1998), which says that "the outbreak was comprised of several separate, but related incidents". Certainly many of the very large outbreaks that have taken place around the world have had been comprised of distinct cohorts. In the United States, there have been a number of outbreaks which involved contaminated ingredients served at multiple restaurants of a single fast food chain. These outbreaks, including one involving minced meat in 2002 and 2003 (Tuttle et al. 1999) and another involving an unidentified ingredient in November-December 2006 (Centers for Disease Control and Prevention 2006d), had a common source, but the infections from each restaurant can be considered a separate cohort. Similarly separate cohorts despite a common source of infection can be identified in the recent multi-state spinach outbreak in the US (Centers for Disease Control and Prevention 2006e), as well as in the largest ever United States *E. coli* O157 outbreak, the 1996 Washington County Outbreak (Centers for Disease Control and Prevention 1999a) and the 2005 meat-product outbreak in Welsh schools. (Office of the Chief Medical Officer 2005)

There is further precedence for cohorts being considered separate outbreaks. In Japan, while approximately 10,000 infections in July 1996 - including those in Sakai City - were all linked back to white radish sprouts grown at a single farm, these cases were considered to comprise at least five separate outbreaks (Michino et al. 1999; WHO 1996).

Overall, the findings in thesis which indicate that, for the purposes of trend modelling, it may be more appropriate to consider the Wishaw Outbreak as a series of smaller outbreaks, suggest that the very large outbreaks may in fact have the epidemiological characteristics of a series of smaller outbreaks. As a result, it may of interest to reconsider how an outbreak is defined. For the purposes of investigating and stopping outbreaks, it is important to connect all cases back to their original source. However, examining trends within individual cohorts may be useful in assessing outbreak control and prevention strategies, in studying transmission, and in predicting future outbreak trends.

3.5.3 Modelled trends

3.5.3.1 Overall variables

The results also provided some interesting insights into the ability to model *E. coli* O157 trends in Scotland and the trends themselves. Linear regression after log transformation of data proved to be an appropriate and effective method for modelling the trends in the number of total isolates, total events, number of sporadic cases, outbreak and the number of outbreaks. Poisson regression was appropriate for the number of ill and ill and positive cases from outbreaks and the number of ill and positive cases per outbreak. The results for these variables seem to suggest that there has been a real decrease both in the number of *E. coli* O157 cases not involved in outbreaks, and the size of outbreaks. The decrease in the size of outbreaks might reflect quicker and/improved outbreak investigations and faster implementation of exclusion policies, which would both decrease the possibility of person to person transmission. Additionally, improved attention to food hygiene and hand-washing might be reducing the rate of person to person spread, and thus the size of outbreaks.

Firstly, the results of the study point strongly towards a true, non-Wishaw dependent decrease in the number of *E. coli* O157 isolates over the last decade because there was statistically significant downward trend in the total number of isolates when the Wishaw isolates were included. Also there was a statistically significant decline in the number of total isolate between 1996 and 2005 when the Wishaw isolates were excluded. That the trend without Wishaw was non-statistically significant until 2005 suggests that the Wishaw Outbreak may initially have been an important factor in the

statistically significant downward trend, but that over time, the statistical significance became independent of the inclusion of the Wishaw isolates. However the 42% increase in cases in 2006 once more makes the trend non-statistically significant when the Wishaw Cases are excluded ($F=1$, $p=0.114$).

The case for a significant decrease in *E. coli* O157 is strengthened by the fact that the number of events has also declined statistically significantly between 1996 and 2004. Given that there was a statistically significant decrease in the number of events, there should also be a statistically significant decrease in the number of sporadic cases unless there was a dramatic decrease in the number of outbreaks. Indeed, there was a statistically significant decrease in the number of sporadic cases. Since the results suggest against any statistically significant influence by enhanced surveillance, it appears that there has been a true and statistically significant decrease in the number of total events and sporadic cases reported in Scotland. This may reflect a decrease in contact with the bacteria, due either to a true decrease in the amount of bacteria in the environment and animals, or improvement attention to hygiene. Better hygiene, in terms of hand washing, especially after coming into contact with animals or animal environments, and proper handling of raw vegetables and meat are infection prevention methods promoted by the Scottish Government, Food Standards Agency - Scotland and HPS (Food Standards Agency 2007; Scottish Executive 2007b). Additionally, the reduction in sporadic cases may well reflect a decrease in the number of household outbreaks.

Intriguingly, though there are statistically significantly decreasing trends in the log transformed numbers of both total number of events and the number of sporadic cases, the trend in the number of outbreaks is non- statistically significant. Since there is only a statistically significant decrease in the total number of ill and ill and positive cases from outbreaks each year and no statistically significant difference before and after enhanced surveillance when the Wishaw Outbreak cases are excluded, it would appear the total numbers of ill and ill and positive cases each year from outbreaks have remained relatively stable over the last decade. The lack of a statistically significant decrease in outbreak case numbers could suggest that prevention programs as mentioned above have not been as effective for reducing

outbreak cases as for reducing sporadic cases. More specifically, the measures may be reducing household transmission, but not transmission in public settings such as nursery schools.

However, some caution should be taken in interpreting these analyses, as the decrease in the number of total events was just barely significant. Thus it could be a result by chance due to over-testing. Additionally, of the low counts and high variance in the analysis of total outbreaks, it is possible that a statistically significant decrease in the number of outbreaks may not be picked up due to low power.

The trends in the number of ill and positive cases per outbreak, both overall and for foodborne outbreaks, could only be modelled when the Wishaw Outbreak was excluded or split into five separate outbreaks, with a statistically significant decline over time in the size of outbreaks when the Wishaw Outbreak was split. A decrease in average outbreak sizes may have resulted from a combination of factors such as increased awareness, and more efficient reporting of outbreaks since the first use of the new reporting forms and policies in 1996, in addition to factors beyond the scope of this study, for instance a shift in the dominant age groups of infected persons (Woolhouse 1998) or a change in factors resulting in decreased infectivity of the bacteria (Friedrich et al. 2002).

However, while the average size of outbreaks appears to be decreasing, the numbers of outbreaks are increasing. This suggest, as mentioned above, that while efforts to prevent infections may be more effective, the methods being used have been better at eliminating isolated cases and/or single household clusters rather than preventing outbreaks. The changes in outbreak size and numbers could also be connected to improvements in the ability of epidemiologists and public health to identify linkages between cases. When cases can be linked, more outbreaks may be recognised, especially those with just a few cases. These improvements could be linked to the onset of the enhanced surveillance program and the work of SERL in typing isolates. The ability to make any firm conclusions about trends in outbreaks and sporadic cases is limited by the fact that Health Protection Scotland defines outbreaks as affecting "members of more than one household…". As a result, clusters of cases

within the same household, which would appear to have all the same characteristics as an outbreak, are categorised as sporadic cases. Therefore in this study, the numbers of sporadic cases and events are over-reported whilst the number of outbreaks and outbreak cases are likely to be under-reported. Thus without being able to separate out the household clusters from the sporadic cases, it is difficult to accurately model trends in outbreaks and sporadic cases (defined as cases not epidemiologically linked to other cases) using the methods in this thesis.

An additional part of this study was to determine the appropriateness and accuracy of predictions for the values of the succeeding data points. Predictions, even if only in terms of ranges, for future data points are of interest for two reasons. Firstly, a rough prediction of future case or outbreak numbers could assist in suggesting what future prevention and investigation resources might be needed. Additionally, demonstrating that a model can accurately predict future data points can provide evidence of the model's validity. Data points for 2005 were only known for the number of outbreaks and total cases, in all but a few cases, the data point was within the 95% confidence intervals of the mean prediction. For most of the inaccurate predictions, the actual data points were zero and thus were out of the possible range of a Poisson or binomial model. In addition, predictions were almost uniformly more accurate when the Wishaw Outbreak data was excluded. However, some caution must be taken because the low number of data points results in very wide confidence intervals, which in some binomial models encompassed almost the entire possible range of values. The study seems to suggest that appropriate models, which can be used with some degree of accuracy to predict future data points, can be constructed using Poisson, linear and binomial regression techniques.

3.5.3.2 Mode of transmission

Since the decrease in the predominance of food as a mode of outbreak transmission, as well as the simultaneous increase in outbreaks caused by environmental/farm related contacts has been noted (Locking et al. 2003c), it was also of interest to look at trends by mode of transmission. In order to assess whether the significance of food as a mode of transmission was more strongly associated with instances where it was the dominant mode of transmission or a co-dominant mode of transmission, trends in three food related variables were created (see 3.2.3.1 and analysed). One

variable – 'foodborne', involved only outbreaks where food was the dominant mode of transmission, while 'multiple including foodborne' only included outbreaks where food was co-dominant. The third variable, 'foodborne element' included all outbreaks where food was involved.

When looking at the number of outbreaks, the major issues regarding modelling of trends was insufficient statistical power due both to a low number of data points, and to large degree of variation between the data point values. As mentioned in Section 2.3.5.2 and above, simulations suggest that the power to detect changes in trend was not very high in some analyses. Whilst only trends in environmental outbreaks could not be modelled, all trends had to be assessed with great care. In particular, information on outbreaks provided in Strachan et al (Strachan et al. 2006) suggests that there are environmental aspects to outbreaks not categorised as such. Thus, analyses using the mode of transmission as provided in the data set may not indicate the true nature of Scottish outbreak trends as far as modes of transmission.

Plots of the data suggest that there has been a decrease in outbreaks and in the number of ill and positive cases per outbreak involving food (see section 3.4.6) and an increase where food was not involved, but there was a statistically significant decrease only in the number of outbreaks where food was the dominant mode of transmission. There would appear to be a steady decrease in the number of outbreaks where food was involved at all or where food was dominant, but for these trends where there were low counts, a lack of power may have contributed to the trends not being statistically significant. A similar issue with lack of power may also have influenced the results regarding the number of ill and positive cases. In both cases, however, the decreases where there were multiple modes of transmission including food appear to be less marked.

The results discussed above seem to suggest that there may have been a real decrease in the predominance of food as a sole mode of transmission. However, food still appears to remain an important factor when it is one of two or more modes of transmission involved in an outbreak. This continued importance could be explained by circumstances such as a decrease in the risk of food (e.g. hamburgers, being contaminated prior to delivery to the consumer) but not in the risk of infection being

spread from a person through food or of a few foodborne cases causing many more subsequent infections. This could occur when person who was potentially infected via food then contaminates swimming water (Breuer et al. 2001; Brewster et al. 1994) or transmits the infection via close contact in a nursery (Spika et al. 1986). However, since non-foodborne outbreaks were only routinely reported via HPS forms starting in 1996, it cannot be ruled out that the statistically significant and non significant changes in predominant outbreak modes of transmission might be a result of changes in reporting patterns rather than a true change in outbreak epidemiology. Examining the proportion of outbreaks and the proportion of ill and positive cases in outbreaks by mode of transmission is potentially problematic given that a decrease in one mode infers increases in other modes. However, the results from these analyses can be examined to ascertain whether an increase or decrease in one mode of transmission was matched by an opposing trend in just one or many other types of modes of transmission. In Scotland, there were significant decreases in both the proportion of outbreaks and of ill and positive cases from outbreaks involving food or a foodborne element. In contrast, no statistically significant increase was seen in non-foodborne modes of transmission, suggesting that while food may be losing dominance, no one non-foodborne mode of transmission is becoming predominant. It would be of interest to determine whether there are common characteristics in the non-foodborne outbreaks, for instance connections to animals or farms, but detailed epidemiological information on outbreaks is not available for this thesis.

Visual inspection suggests that enhanced surveillance (see section 3.2.3.3) may have had an important role in the very apparent downward shift in the proportion of ill and positive cases from foodborne outbreaks or those with a foodborne element. This shift, which took place during the years in which the enhanced surveillance program started (1999, 2000), could possibly indicate that there has been a true shift in the causes of *E. coli* O157 outbreaks in Scotland. However it is unlikely that the coordination of the start of enhanced surveillance and the shift was a coincidence, and thus the data would suggest that the data after 1999 reflects the true nature of outbreaks, while the high proportions prior to 1999 were the result of epidemiological investigations that focused primarily on food as a source of transmission. Both the visual and statistical observations, which show a significantly

decreasing trend regardless of the Wishaw Outbreak, strongly argue for a major influence of enhanced surveillance and the related increase in awareness of non-foodborne outbreaks.

It is important to note that the only information available for this thesis on the mode of transmission for Scottish outbreaks included in this chapter was the designation of mode of transmission. Thus, it was not possible to determine the strength of evidence for the choice of mode of transmission. In fact, some outbreaks where environmental or animal contact was not designated as the mode of transmission, may have in fact had an animal/environmental component to infection transmission (Strachan et al. 2006). Thus the analyses of environmental/animal contact trends may not represent the full burden of these modes of transmission. If more information on outbreak modes of transmission is made available, further analyses would be of interest.

3.5.3.3 Phage Types 2 and 21/28

Finally, the results provide further statistical evidence for the previously noted shift in dominance - from PT 2 to PT 21/28 - of the two most common outbreak phage types in Scotland (Locking et al. 2003c). The increase in the number of PT 21/28 outbreaks between 1996 and 2004 was just barely non-significant ($p=0.055$). When the Wishaw Outbreak cases were excluded, there was not a statistically significant decrease in the number of ill and positive cases per year in PT 2 outbreaks between 1996 and 2004 ($p=0.267$). However, given the low counts, there may not have been enough power to detect a statistically significant trend. Of particular interest are the trends in the first three years in which the proportion of outbreaks that were Phage Type 21/28 went from 0 to 1 and the proportion of those that were Phage Type 2 went from 1 to 0. Though these trends could not be appropriately modelled, they reveal a dramatic change in the phage type dominance. This would suggest that while there may have been a real shift in phage type by outbreak, this shift could not be adequately modelled using simple linear models. Since the common phage types in cattle and sheep are the same as those in humans (Paiba et al. 2002), the shift in predominant human phage type may be linked to a shift in the predominant phage

type in cattle and sheep. The predominance of PT 21/28 is of concern because it is associated with severe illness in humans (Halliday et al. 2006).

Intriguingly, there was a significant downward trend in the size of PT 21/28 outbreaks ($p<0.001$), while the size of PT 2 outbreaks showed no significant change over time, remaining small (<10 ill and positive cases). It would be of interest to see whether there is any statistical relationship between PT 21/28 outbreaks and other populations of outbreaks, for instance those spread by food, which have also experienced a significant decrease in size. However, given the small numbers of outbreaks, and given the simulations run to determine power using large n's (see 2.3.5.2) it is unlikely that the analyses would have enough power to reveal any statistically significant relationships. For instance, in the eight years from 1996 to 2003, there were only three outbreaks that were both PT 2 and foodborne; with just seven PT 21/28 and foodborne.

3.5.4 Conclusions

This study would suggest that simple regression models are appropriate for statistical modelling of some of the trends in *E. coli* O157 in Scotland. These models indicate that there has been a statistically significant decrease in non outbreak cases between 1996 and 2004, as well as a statistically significant decrease in cases involving food. Overall outbreak sizes, as well as sizes of outbreaks involving food and involving PT 21/28 have also statistically significantly decreased. However, there has not been a statistically significant change in the number of cases involved in outbreaks, nor in the number of outbreaks. Models could not be constructed for most variables including cases from the Wishaw Outbreak, suggesting that the Wishaw Outbreak may be a statistical outlier. Other issues with model construction resulted from the low number of outbreaks, a factor which reduces the power available in analyses and warrants careful interpretation of analyses and predictions. The following Chapter, Chapter 4, will focus on the analysis of trends in the United States.

Chapter 4 -- *E. coli* O157 trends in the United States

4.1 Introduction

E. coli O157 was first recognized as a human pathogen in the United States after two outbreaks of bloody diarrhoea, both eventually linked to consumption of undercooked beefburgers, in 1982 (Riley et al., 1983; Wells et al., 1983). Notifiable to the federal government since 1994, *E. coli* O157 has been implicated in over 350 outbreaks since 1982 (Rangel et al. 2005), though the vast majority of cases have been sporadic (O'Brien & Adak 2002). Based on analyses done by Hedberg et al which suggest that only approximately 1 in 20 cases are reported (Hedberg et al. 1997), it has been estimated that the bacteria may be responsible for over 73,000 illness each year (Mead et al. 1999), with 85% due to foodborne transmission and approximately one third requiring hospitalization (Mead et al., 1999).

Infection rates have always been highest in the northern states, including California, Oregon, Wisconsin, Michigan, Minnesota, New York and Washington (Griffin and Tauxe, 1991). The most common source of infection is contaminated beefburgers, which accounted for 50% of outbreaks between 1982 and 1995, but outbreaks have also been transmitted via water as well as other foods including raw milk, apple cider and lettuce (Wachsmuth et al., 1997). In recent years, contamination of vegetables - particularly salad greens – has been implicated in several large multi-state outbreaks (Centers for Disease Control and Prevention 2006d).

E. coli O157 outbreaks are a continuing problem in the United States, as illustrated by the series of large outbreaks in the last three years including the multi-state outbreak due to contaminated food items at Taco Bell restaurants in the Mid-Atlantic States (at least 69 confirmed cases) (Centers for Disease Control and Prevention 2006d) and a national outbreak caused by contaminated fresh spinach, mostly sold in bags of ready to eat salads (205 confirmed cases, 3 deaths) (Centers for Disease Control and Prevention 2006h)in the last quarter of 2006. However, as in Scotland, while there have been a number of publications describing outbreak epidemiology (Centers for Disease Control and Prevention 1996b; Centers for Disease Control and Prevention 2004c; Rangel et al. 2005), there has been little statistical analysis of any temporal outbreak trends. A 2004 FoodNet report by Marcus and colleagues, which analysed the trends in overall *E. coli* O157 case trends in the FoodNet sites, is one of

the few exceptions (Marcus et al. 2004a). However that analysis involved only cases reported in the nine FoodNet sites, regions that range from a single city to a whole state, and no actual statistical values were provided in the results. Thus the analyses in this chapter have been conducted in order to provide much more complete and comprehensive data on temporal trends in the whole United States.

4.2 Materials and methods

4.2.1 Data

4.2.1.1 Study data set

The data sets for United States *E. coli* O157 outbreaks in 1996 – 2004 were obtained courtesy of Liz Blanton and Thai An Nguyen at the Enteric Diseases Epidemiology Branch in the Centers for Disease Control and Prevention (CDC). Data for foodborne outbreaks in 1996 and 1997 was taken from the data set provided in early 2005. This data did not undergone the same cleaning – updating to incorporate new data on existing outbreaks, to add new outbreaks and to adjust old case variables so they are equivalent with currently used variables – by the CDC as the data after 1997. Foodborne outbreak data from 1998 through 2004 and all data on non-foodborne outbreaks were taken from updated data sets provided by the CDC in early 2007. The original variables from these CDC data sets are referred to in italics.

The figures on total yearly reported cases were taken from the Summary of Notifiable Diseases --- United States 2003 (Centers for Disease Control and Prevention 2005e). Since data from the PHLIS system were not available after 2001, the NNDSS data were used for the purposes of this study.

4.2.1.2 Time period

Outbreaks which occurred prior to 1996 are not being considered, first because the time period 1996 – 2004 was selected in order to allow ease of comparison with Scotland and Canada, where data were not available before 1996. More importantly, *E. coli* O157 infection was only made federally [i.e. state health agency to federal agency] notifiable in 1994, but 17 states in 1994 and 11 in 1995 still did not require notification [i.e. local/county agency to state health agency] (Centers for Disease Control and Prevention 1995b; Centers for Disease Control and Prevention 1996a),

so no data was reported to the federal government. Since states which did not report in 1994, but did report by 1996 represented approximately 15% of 1996 cases, 1994 and 1995 data was not included due to the possibility that any trend might reflect more the increase in reporting than any underlying epidemiological cause. However all fifty states did not report individual cases to the CDC (Table 3.1) until 2002, so the total number of cases prior to 2002 is most probably an underestimate of the actual number.

Table 4.1: States which did not report data to the NETSS/NNDSS System between 1996 and 2004

States for which data was not reported to the NETSS/NNDSS system between 1996 and 2004	
1996	Pennsylvania, Virginia, West Virginia, Hawaii, Wisconsin, Arizona
1997	Pennsylvania, Virginia, West Virginia, Hawaii
1998	Pennsylvania, Wisconsin
1999	Pennsylvania, Wisconsin
2000	Pennsylvania
2001	Pennsylvania
2002	--
2003	--
2004	--

For example, between 1996 and 2002 when the state did not report individual cases to NNDSS or PHLIS, there were 141 outbreak cases reported from Pennsylvania via the outbreak reporting system. Also, in 1998 the reporting system for foodborne outbreaks, through which most *E. coli* O157 outbreaks are reported, shifted from a paper based system to an online system. At that same time "formal confirmation procedures to finalize reports from each state each year" (Lynch et al. 2006) were instituted. In addition, a revised reporting form went into use starting in 1999. This led to "a substantial increase in the number of reports", and what is termed by the CDC as a "surveillance discontinuity" in 1997-1998 (Lynch et al. 2006). Given the stated effect of the changes in reporting procedures and forms, it is expected that there will be an effect on any temporal trends in outbreak variables. Thus, statistical analyses on the temporal trends in this chapter will be done only on data from 1998 to 2004.

4.2.2 Definitions

Below are the definitions used for the analyses of the United States data. These definitions may vary from those in other countries, in particular Scotland and Canada, and issues regarding these variations will be discussed in Chapter 6.

4.2.2.1 Definition of a case

In the United States, a case is generally defined as a suspect, probable or confirmed incidence of *E. coli* O157 infection.

4.2.2.2 Definition of an outbreak

In the United States, for the purposes of these data sets an outbreak is defined as "two persons becoming ill as the result of a common source" (Lynch et al. 2006).

4.2.2.3 Definitions -- Variables

Definition of total ill cases

For the purposes of this chapter, the number of total ill cases is the number of ill cases from sporadic cases and outbreaks. The values for the variable are taken directly from the NNDSS data provided in the CDC's Summary of Notifiable Diseases (Centers for Disease Control and Prevention 2007g).

Definition of number of ill cases in outbreaks

In this chapter, the number of ill cases in outbreaks refers to cases that are confirmed and/or epidemiologically linked to an outbreak. For all non-foodborne outbreaks and foodborne outbreaks in 1996-1997, data values were taken from the original variable *ill.* Values for foodborne outbreaks between 1998 and 2004 were taken from the original variable *estimated ill*, the sum of the number of lab confirmed cases, probable cases and the number of other cases that were considered by the reporting person or body to be involved.

Definition of number of ill and positive cases in outbreaks

For all non-foodborne outbreaks, and for foodborne outbreaks prior to 1998, the values for number of ill and positive cases were taken from the original variable *CultureConfirmedCases*. From 1998 - 2004, values for foodborne outbreaks were taken from the original variable *Lab Cases Primary*, the number of primary culture

confirmed cases. Since secondary cases are estimated to compromise up to 20% of confirmed outbreak cases (see Chapter 7), the omission of secondary cases in the value for number of ill and positive cases may affect analyses of trends in outbreak size and case numbers in terms of ill and positive cases.

The value for *CultureConfirmedCases* or *Lab Cases Primary* was omitted in 30 outbreaks, 22 of which were in 1998, so the variable (as Ill and positive cases) is limited in use for the purposes of the analyses.

Definition of mode of transmission variable

Values for the variable *Transmission* were provided for all outbreaks. Outbreaks were categorised as foodborne, waterborne, person to person, unknown or environmental/animal contact. Mode of transmission values 'foodborne', 'waterborne' and 'person to person' were taken directly as listed on the original data sets for. The original values, 'animal contact', 'animal exposure' and 'environmental' were combined to create the new category 'environmental/animal contact'.

4.2.3 Reporting issues

4.2.3.1 The Washington County Fair and Wisconsin Watermelon outbreaks

Between 1996 and 2003 there were two outbreaks in the United States that were more than twice as large as any other outbreaks in terms of ill cases, though they were not larger in terms of ill and positive cases. These two outbreaks, which took place in 1999 and 2000 respectively, will be termed the Washington County Fair (WCF) Outbreak and the Layton Avenue Sizzler (LAS) Outbreak.

The Washington County Fair Outbreak took place in August and September of 1999 at the Washington County fairgrounds near Albany, New York. Visitors to the fair were exposed to *E. coli* O157:H7 through consumption of ice or beverages made with water from a shallow, un-chlorinated well (Centers for Disease Control and Prevention 1999a), one of several sources of potable water at the fairgrounds. There were 781 confirmed or suspected cases, 126 of which tested positive (New York State Department of Health 2000)for *E. coli* O157:H7 and 45 for *Campylobacter*. Of

the 126 confirmed cases, 14 progressed to HUS and two died. The most probably source of contamination was confirmed by PFGE typing which established that the PFGE type of the samples from the well was indistinguishable from that of "many" of the patients (New York State Department of Health 2000). All persons were infected at the same location (the fair) and from the same source, so the cases could not be broken into cohorts as done with the Wishaw Outbreak (see Section 3.2.3.2).

In July of 2000, 736 people were sickened in the Layton Avenue Sizzler Restaurant in Milwaukee, Wisconsin, resulting in 63 confirmed cases, 4 cases of HUS and one death, as well as two additional confirmed secondary cases. The epidemiological and microbiological investigations revealed that the infections were most likely to be the result of cross contamination and poor hygienic procedures in the restaurants. It is thought that watermelon and other prepared fruits in the salad bar were contaminated by beef being minced in close proximity within the kitchen of the restaurant. Three minced beef samples tested positive for *E.* coli O157, as did a sealed package of sirloin tri-tip from the same lot as delivered to the restaurant. Isolates from the meat, one of the samples and the infected persons had indistinguishable PFGE patterns (Proctor 2000a).
For the purposes of this study, these two large outbreaks raise concerns because of the number of ill cases involved, more than twice as many ill cases as in each of the other 314 outbreaks in the data set. The presence of the two outbreaks in the middle of the time period being investigated might lead to a distortion of trends present in the data or present problems in finding an appropriate model, as seen with the Wishaw outbreak. In order to determine whether the outbreaks have had any significant effect on the trends being modelled or have affected the ability to model the trends, all analyses involving ill case numbers will also be run omitting each outbreak separately and omitting both outbreaks.

The largest outbreak to date in the United States, in terms of confirmed cases, took place between November 1992 and March 1993 in four states (Washington, California, Nevada and Idaho). Caused by improperly cooked minced beef served at a restaurants that were part of a large fast food chain, the outbreak resulted in four deaths and over 700 confirmed cases (Centers for Disease Control and Prevention,

1993). However because 1992 and 1993 were out with the period of the study, this outbreak was not part of the data set used in the analyses.

4.2.3.2 Outbreak reporting changes

The other important factor which must be considered in the analysis of trends in *E. coli* O157 outbreaks in the United States is the change in the reporting procedures for outbreaks. Prior to the changes in the 2004 form 52.13, the foodborne outbreak reporting form, which do not affect the analysis of the data from 1998 to 2004, waterborne outbreaks may have been reported using form 52.12, which is specifically for waterborne outbreaks. Form 52.12, like the pre-1999 form 52.13, does not ask for the number of confirmed cases, only for the number of exposed and ill cases (Centers for Disease Control and Prevention 2003c). This is as opposed to the current form 52.13 which asks for lab-confirmed, probable and estimated total ill cases. Thus there may be differences in the way confirmed and ill case totals are calculated for foodborne and waterborne outbreaks prior to 2004. This is, however, not anticipated to have a statistically significant effect on the results because data for all waterborne outbreaks was taken from the data set provided in 2007, which had undergone extensive data cleaning by CDC researchers (personal communication, Thai-An Nguyen, 2006). As part of the cleaning process, the data was adjusted to fit within the new variable definitions.

4.2.4 Statistical analyses

4.2.4.1 Descriptive statistics

The data from 1996 to 2004 was described in terms of the overall number of and yearly range of total cases, outbreaks and ill and ill and positive cases from outbreaks. Outbreaks were also described in terms of mode of transmission. Average outbreak size was calculated as the geometric mean number of ill cases per outbreak.

4.2.4.2 Introduction - Models

For this chapter, temporal trends in both count and continuous data from 1998 to 2004 were analysed using GLMs and linear models as specified in Chapter 2.

4.2.4.3 Analyses specific to the United States data

All data was modelled for both the 1998 – 2003 and 1998 – 2004 data sets excluding the data prior to 1998 for reasons described in 3.2.1.3. Trends before and after the onset of the changes to the surveillance system were not modelled as for the Scottish data because only two time points took place before the changes, an insufficient amount of data for a regression mode.

All variables regarding or derived from data regarding the number of cases were also modelled with and without the cases from the Washington County Fair and Layton Avenue Sizzler Outbreaks. This was done in order to determine whether the statistical significance of any trend was due solely to the influence of the cases from these outbreaks. Initial analyses suggested that the Washington County Fair and Layton Avenue Sizzler Outbreaks might have a large impact on certain trends because these two outbreaks were involved so many more ill cases than any other outbreaks during the study time period.

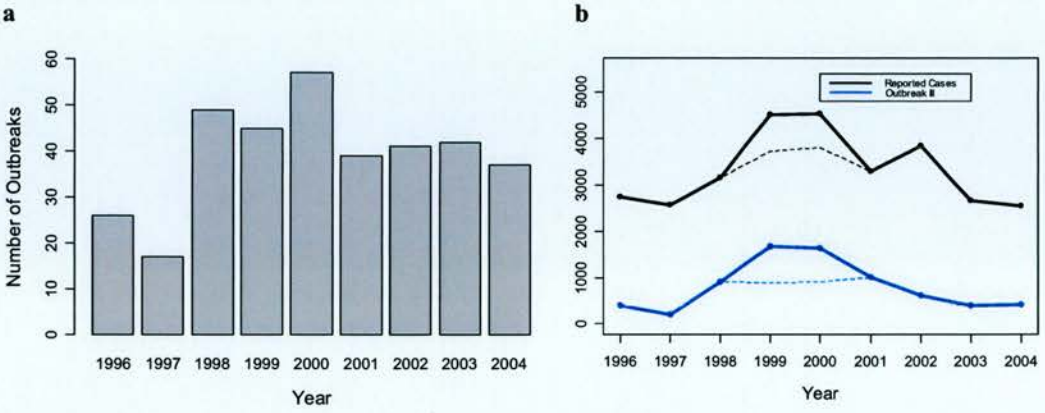
For all variables, if more than one model appeared to be appropriate based on residual plots, models were selected to allow best comparison between data sets with and without the WCF and LAS ill cases.

4.3 Results – Descriptive

4.3.1 1996 – 2004 data

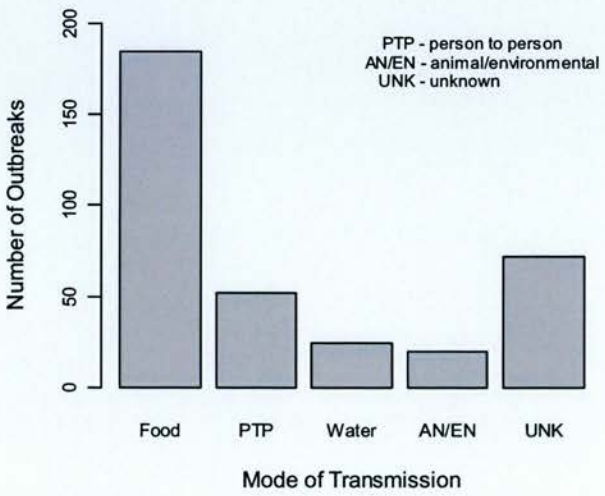
Between 1996 and 2004 there were 29,840 probable and confirmed cases of *E. coli* O157 reported via the NNDSS. During this same time period there were 353 outbreaks, with the number of outbreaks per year ranging from 17 to 57 (Figure 4.1a). Within these outbreaks there were 7309 ill cases (range 203 to 1680; Fig. 4.1b). The number of ill and positive cases was available for 323 outbreaks, totalling 2893 cases. Outbreaks had a geometric mean of 8.57 ill cases (95% CI = 7.62, 9.65).

Figure 4.1 a-b: (a) Number of outbreaks per year and (b) number of reported cases and number of ill cases from outbreaks, 1996 – 2004
Plots showing (a) the number of outbreaks and (b) the number of reported cases and ill cases from outbreaks per year, 1996 – 2004. The number of reported cases is indicated by a black line, the number of ill cases from outbreaks as a blue-dotted line. The dotted lines between 1998 and 2001 indicate the trends when the cases from the two large outbreaks (Washington County Fair and Layton Avenue Sizzler) are excluded.



The highest number of outbreaks - 185 - were spread via foodborne transmission, whilst 52 were spread person to person (Fig. 4.2). For the 24 waterborne outbreaks, 13 were spread via swimming water, 8 through drinking water; the water source was not specified in the remaining three. Of the 20 outbreaks categorised as environmental/animal contact, 17 were transmitted via animal contact or exposure and three via environmental contamination.

Figure 4.2: Number of outbreaks by mode of transmission (1996 – 2004)
The number of outbreaks by mode of transmission, as defined above



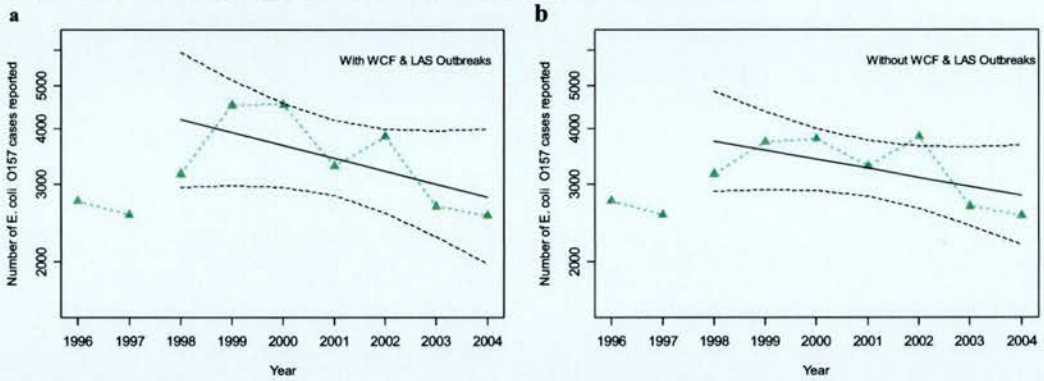
4.4 Results – Trend Analyses

4.4.1 Overall variables

4.4.1.1 Total reported cases from sporadic and outbreak *E. coli* O157 cases in the United States, by year

The temporal trends in the number of total reported *E. coli* O157 cases can be modelled using a linear model with log10 transformation of the response variable. There is not a statistically significant downward trend between 1998 and 2004 ($F_{1,5}$=3.1, p=0.139; Fig. 4.3a). If the Washington County Fair and Layton Avenue Sizzler Outbreaks are omitted, the trend between 1998 and 2004 is still statistically non-significant ($F_{1,5}$=2.8, p=0.155; Fig. 4.3b).

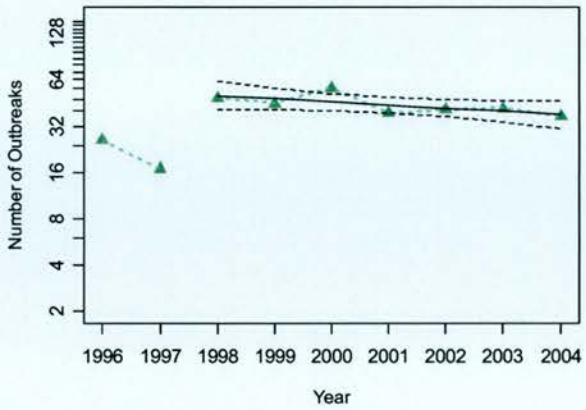
Figure 4.3 a-b: Number of total cases reported to NNDSS: 1996 – 2004
The total number of *E. coli* O157 cases reported via the NNDSS reporting system, by year, between 1996 and 2004 (a) with and (b) without the cases from the Washington County Fair (WCF) and Layton Avenue Sizzler (LAS) Outbreaks. The solid black lines indicate the trends between 1996 and 1998 and between 1999 and 2004, with the dashed lines indicating the 95% confidence intervals around the regression line for the trend between 1999 and 2004.



4.4.1.2 Total outbreaks

The trend between 1998 and 2004 can be modelled using a linear model with log10 transformation of the response variable. However the trend is not statistically significant ($F_{1,5}$=4.4, p=0.090; Fig. 4.4).

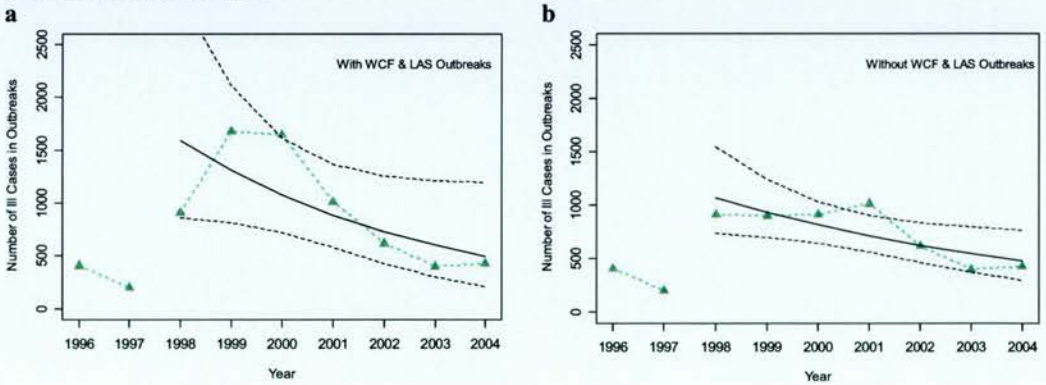
Figure 4.4: Number of outbreaks
The number of *E. coli* O157 outbreaks, per year, between 1996 and 2004. The solid black line indicates the best fit lines, on a log transformed scale, for the trend between 1998 – 2004, with the dotted lines indicating the 95% confidence intervals for the trend between 1998 and 2004.



4.4.1.3 Number of ill in outbreaks

The linear temporal trends in the number of ill cases in outbreaks between 1998 and 2004 can be modelled using a quasipoisson model, with the model a better fit when the two large outbreaks are omitted. There is no statistically significant trend in the number of ill cases in outbreaks between 1998 and 2004 ($F_{1,5}$=6.0, p=0.059; Fig. 4.5a). However, if the WCF and LAS Outbreaks were excluded, there is a statistically significant downward trend from 1998 to 2004 ($F_{1,5}$=8.9, p=0.031; Fig, 4.5b).

Figure 4.5 a-b: Number of ill cases in outbreaks: with and without the WCF & LAS outbreaks
The number of ill cases from outbreaks between 1996 and 2004, with the best fit line and 95% confidence intervals around the trend indicated for 1998 – 2004 when the large outbreaks (a) are included and (b)excluded.

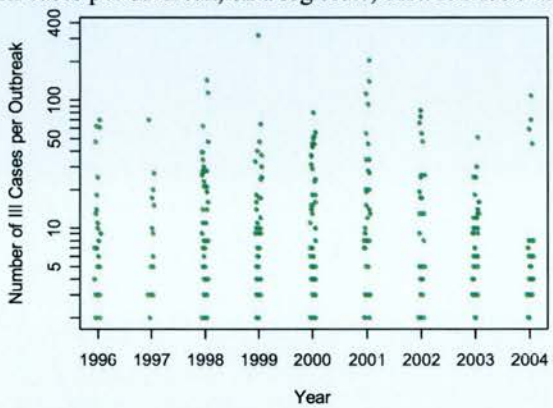


4.4.1.4 Number of ill cases per outbreak

Trends in the number of ill cases per outbreak between 1998 and 2004 (Fig. 4.6), whether or not the two large outbreaks were included, cannot be modelled using the simple linear models.

Figure 4.6: Number of Ill Cases per Outbreak, 1996 – 2004

The number of ill cases per outbreak, on a log scale, between 1996 and 2004



4.4.2 Number of outbreaks

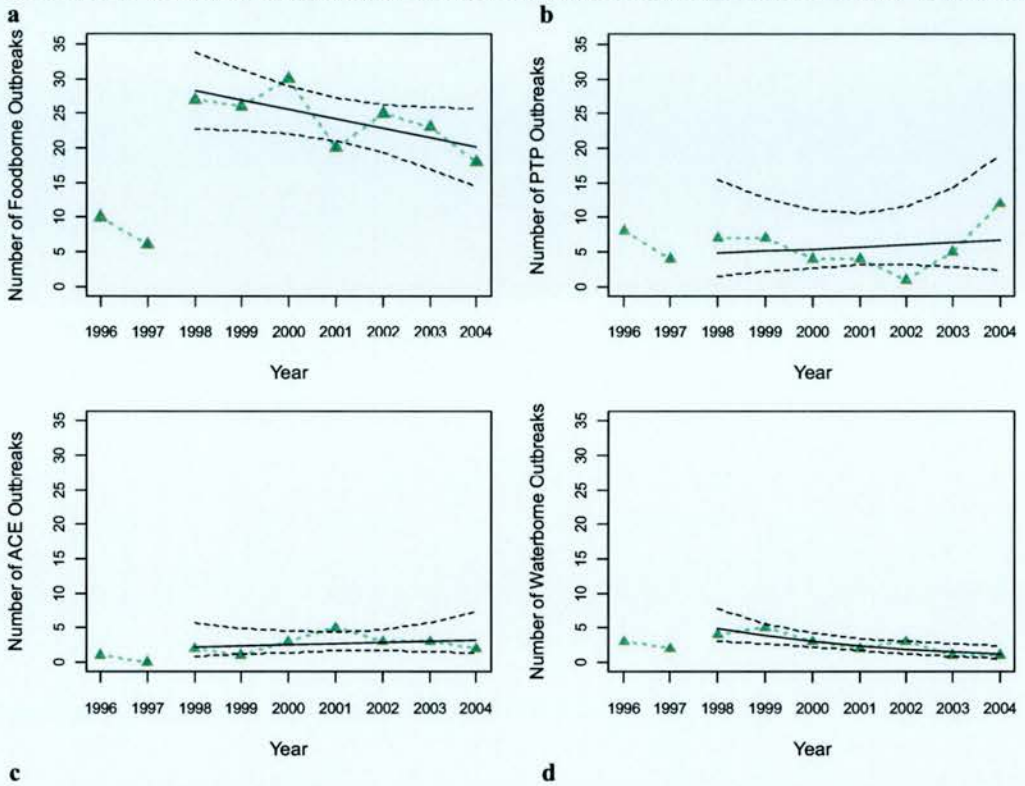
Using an ANCOVA analysis, there is no evidence of any statistical difference in temporal trends of the number of outbreaks between modes of transmission between 1998 and 2004 ($F_{3,20}$=1.7, p=0.190). However, the trend for each mode of transmission will still be modelled separately in order to determine whether any one trend is statistically significant.

4.4.2.1 Number of outbreaks, per year – foodborne, person to person, waterborne and animal/environmental

There is not a statistically significant trend between 1998 and 2004 in the number of foodborne ($F_{1,5}$=4.8, p=0.079; Fig. 4.7a), person to person ($F_{1,5}$=0.18, p=0.691; Fig. 3.7b) or animal/environmental outbreaks ($F_{1,5}$=0.31, p=0.599; Fig. 4.7c). However, the number of outbreaks spread by water has decreased statistically significantly between 1998 and 2004 ($F_{1,5}$=14.1, p=0.013; Fig. 4.7d).

Figure 4.7 a-d: Number of outbreaks, per year – foodborne, person to person, waterborne and animal contact/ environmental

Trends in the numbers of outbreaks between 1998 and 2004 in which the main mode of transmission was (a) foodborne, (b) person to person, (c) animal/environmental and (d) waterborne.



4.4.3 Proportion of outbreaks – by mode of transmission

There is no evidence for a statistically significant difference between modes of transmission in the trends between 1998 and 2004 ($F_{3,20}$=1.5, p=0.238). However, the trends for each mode of transmission are of interest and are modelled separately.

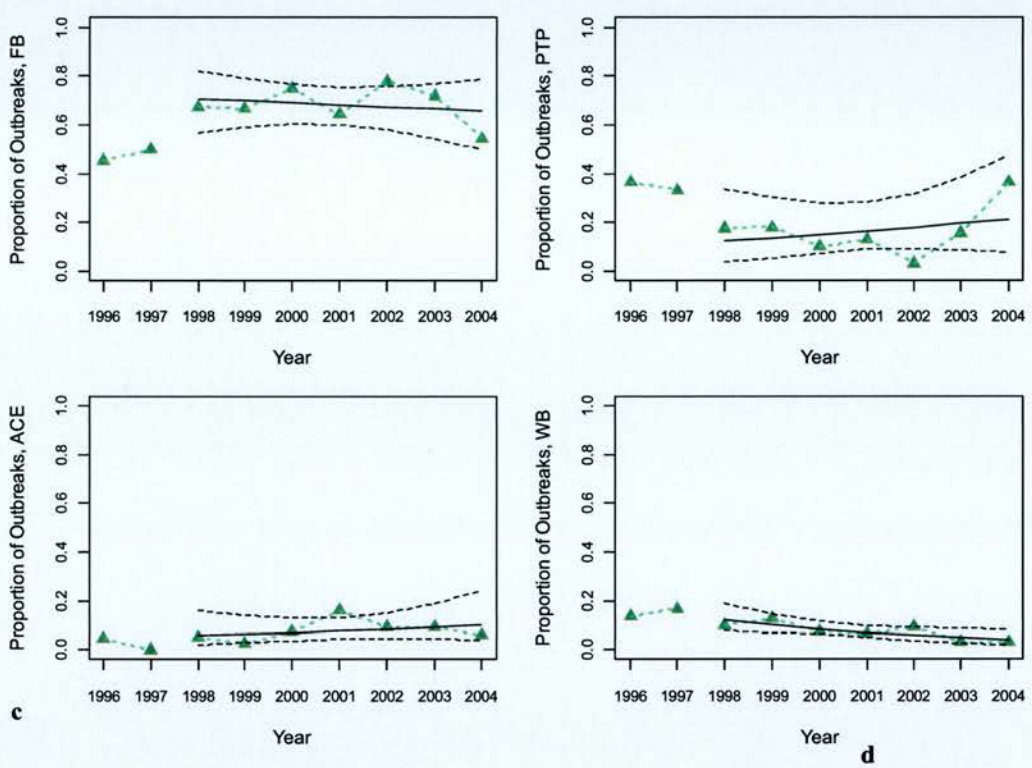
4.4.3.1 Proportion of outbreaks, by year – foodborne, person to person, waterborne and animal contact/environmental

The trends in the proportions of outbreaks with a known mode of transmission which are foodborne ($F_{1,5}$=0.3, p=0.591; Fig. 4.8a), person to person ($F_{1,5}$=0.7, p=0.450; Fig. 4.8b) and animal contact/environmental ($F_{1,5}$=0.78, p=0.417; Fig. 4.8c) can be modelled, but are not statistically significant. However, there is a statistically

significant decline in the proportion of outbreaks which are waterborne ($F_{1,5}$=8.7, p=0.032, Fig. 4.8d).

Figure 4.8: Proportion of outbreaks – foodborne, person to person, waterborne and animal contact/environmental

Trends between 1996 and 2004 in the proportion of outbreaks (out of outbreaks with a known mode of transmission), in which the mode of transmission was (a) foodborne, (b) person to person, (c) animal contact/environmental and (d) waterborne. All models have quasibinomial error structures.



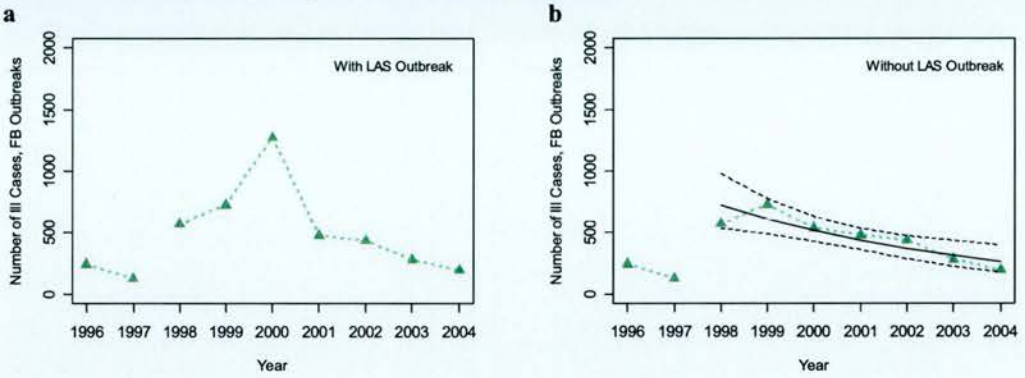
4.4.4 Number of ill cases in outbreaks

There is a statistically significant difference between modes of transmission in the trends in the number of ill cases from outbreaks between 1998 and 2004 when the two large outbreaks are omitted ($F_{3,20}$=11.3, p<0.001), but not when the two large outbreaks are included ($F_{3,20}$=1.95, p=0.153). The trends in the number of ill cases from outbreaks spread person to person, regardless of the inclusion of the WCF outbreak cases and in those spread by food and water when the WCF Outbreak is omitted, can be modelled using the techniques discussed in 2.3.4.

4.4.4.1 Number of ill cases in outbreaks – foodborne

The trend in the number of ill cases from outbreaks spread by food cannot be modelled (Fig. 4.9a). However, when the LAS outbreak is omitted, the trend could be modelled using quasipoisson error structure. There is a statistically significant decreasing trend in the number of ill cases between 1998-2004 ($F_{1,5}$=19.6, p=0.0067; Fig. 4.9b).

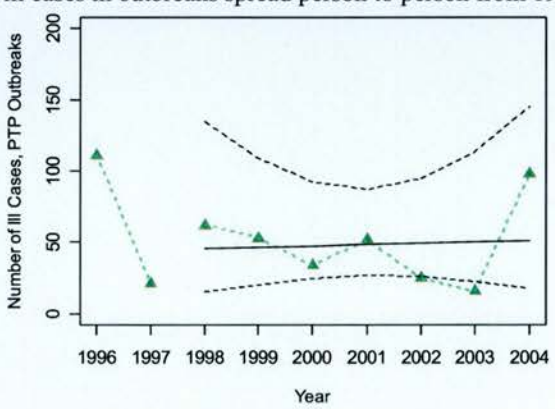
Figure 4.9 a-b: Number of ill cases in outbreaks: foodborne 1996 – 2004
Number of ill cases in outbreaks where the mode of transmission was foodborne, (a) with and (b) without the cases from the Layton Avenue Sizzler Outbreak.



4.4.4.2 Number of ill cases in outbreaks – person to person

When a quasipoisson model is used, there is no statistically significant trend between 1998 and 2004 in the number of ill cases in outbreaks ($F_{1,5}$=0.3, p=0.878; Fig. 4.10).

Figure 4.10: Number of ill cases in outbreaks – person to person
The number of ill cases in outbreaks spread person to person from 1996 - 2004

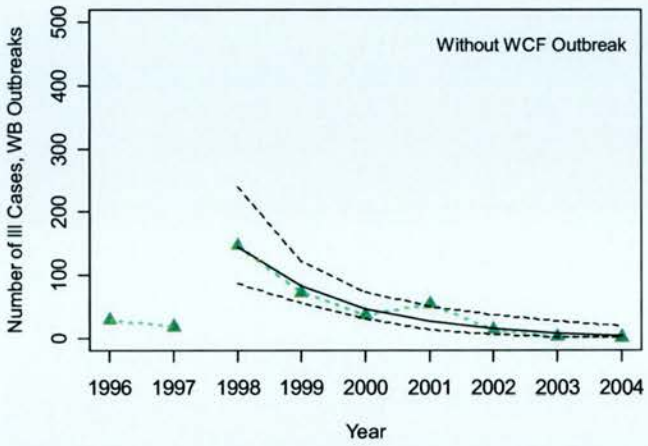


4.4.4.3 Number of ill cases in outbreaks – waterborne

To examine the number of ill cases per year in outbreaks spread by water, quasibinomial models are used. The trend between 1998 and 2004 can only be modelled when the cases from the Washington County Fair Outbreak were excluded.

When the WCF Outbreak is omitted, there is a statistically significant decreasing trend between 1998 and 2004 ($F_{1,5}$=39.1, p=0.002; Fig. 4.11).

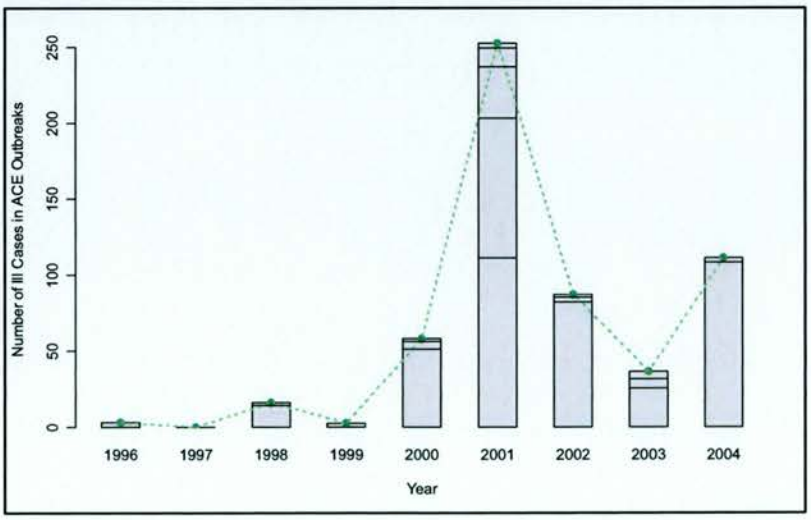
Figure 4.11: Number of ill cases in outbreaks, waterborne – 1996 – 2004
The number of ill cases in outbreaks spread person to person from 1996 – 2004 without the cases from the Washington County Fair Outbreak.



4.4.4.4 Number of ill cases in outbreaks – animal contact and environmental

The trend between 1998 and 2004 in the number of ill cases from outbreaks where the infection was spread via animal or environmental means cannot be modelled (Fig. 4.12) because there were very few outbreaks involved (less than six per year) and the number of outbreaks differed substantially between years given the number of outbreaks involved (range of 1 to 6 outbreaks).

Figure 4.12: Number of ill cases in outbreak – animal contact/exposure and environmental
The number of ill cases in outbreaks spread via animal contact/exposure or environmental means from 1996 – 2004. Within each bar, the each segment represents the cases within one outbreak. The dotted line connects the data points for each year.



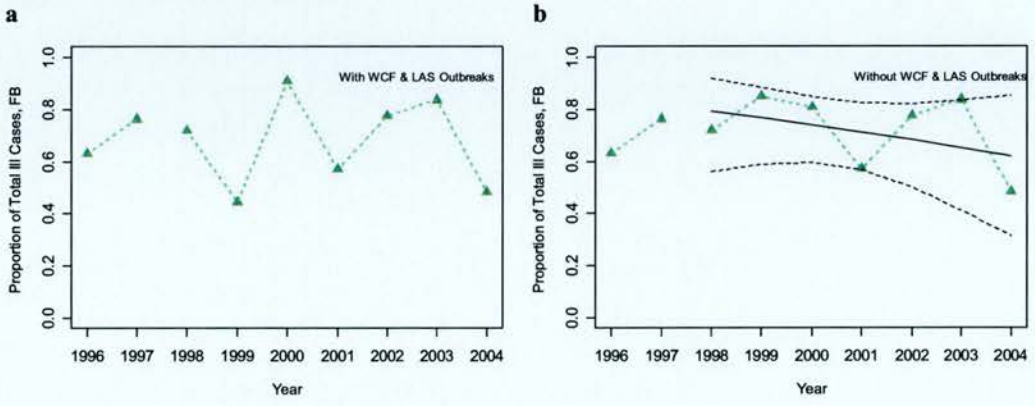
4.4.5 Proportion of ill cases in outbreaks, by mode of transmission

There is a statistically significant difference between modes of transmission in the trends in the proportions of total ill cases from outbreaks with a known mode between 1998 and 2004 when the two large outbreaks are included ($F_{3,20}$=11.3, p<0.001), and when the two large outbreaks are excluded ($F_{3,20}$=5.2, p=0.008). The trends in the number of ill cases from outbreaks spread person to person (Section 4.4.4.2), regardless of the inclusion of the WCF and LAS outbreak cases, and those spread by food (Section 4.4.4.1) and water (Section 4.4.4.3) – only when the WCF or LAS outbreak is omitted - can be modelled using the techniques discussed in 2.3.4.

4.4.5.1 Proportion of ill cases in outbreaks – foodborne

The trends in the proportion of ill cases from outbreaks with known modes of transmission that were spread by food between 1998 and 2004 cannot be appropriately modelled using the binomial error distribution when the cases from the WCF and LAS Outbreaks were included (Fig. 4.13a). When the cases from these two large outbreaks are excluded, the trend can be modelled (Fig. 4.13b), but is not statistically significant ($F_{1,5}$=1.2, p=0.322).

Figure 4.13 a-b: Proportion of ill cases from outbreaks – foodborne
The proportion of ill cases from outbreaks of known mode of transmission between 1996 and 2004 that were spread via food, (a) with and (b) without the cases from the Washington County Fair and Layton Avenue Sizzler Outbreaks.

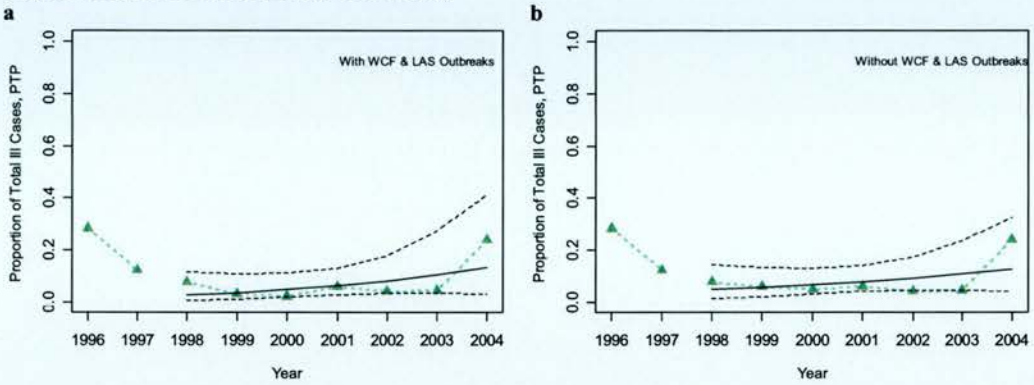


4.4.5.2 Proportion of ill cases in outbreaks – person to person

The trends in the proportion of ill cases from outbreaks with known modes of transmission that were spread person to person between 1998 and 2004 can be appropriately modelled using the binomial error distribution whether or not the cases

from the Washington County Fair and Layton Avenue Sizzler Outbreaks are included (Fig. 4.14a-b). However neither the trend including the large outbreaks ($F_{1,5}$=2.7, p=0.160) nor the trend excluding the two large outbreaks ($F_{1,5}$=1.8, p=0.233) are statistically significant.

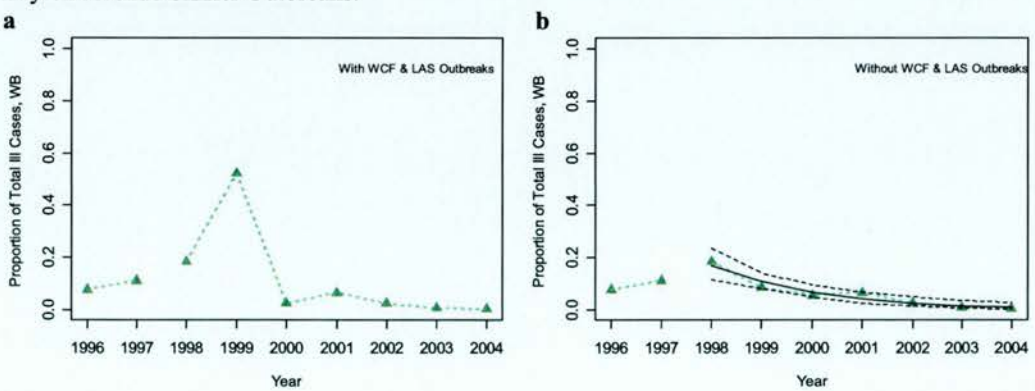
Figure 4.14 a-b: Proportion of ill cases from outbreaks – person to person
The proportion of ill cases from outbreaks of known mode of transmission between 1996 and 2004 that were spread person to person (a) with and (b) without the cases from the Washington County Fair and Layton Avenue Sizzler Outbreaks.



4.4.5.3 Proportion of ill cases in outbreaks – waterborne

An appropriate model cannot be found for the trend in the proportion of ill cases that come from outbreaks with waterborne transmission when the Washington County Fair and Layton Avenue Sizzler outbreak cases are included (Fig, 4.15a). When these cases are excluded, there is a statistically significant decreasing between 1998 and 2004 ($F_{1,5}$=45.6, p=0.001; Fig. 4.15b).

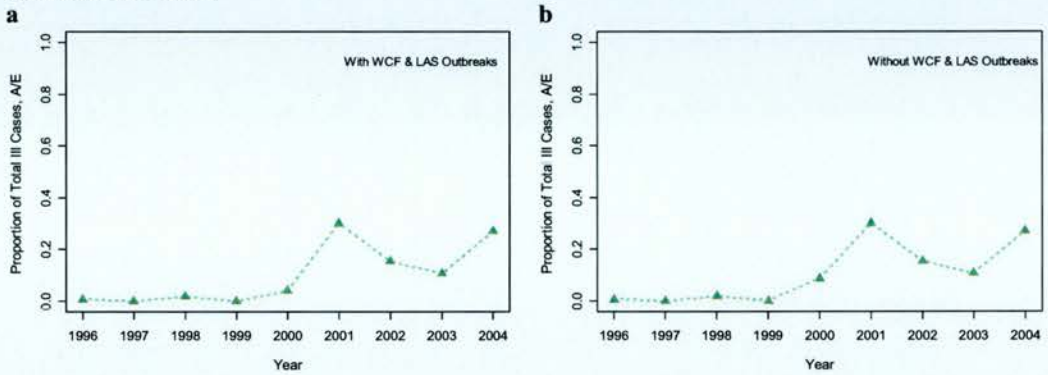
Figure 4.15 a-b: Proportion of ill cases from outbreaks – waterborne
The proportion of ill cases from outbreaks of known mode of transmission between 1996 and 2004 that were spread via water (a) with and (b) without the cases from the Washington County Fair and Layton Avenue Sizzler Outbreaks.



4.4.5.4 Proportion of ill cases in outbreaks – animal contact or environmental

The trends between 1996 and 2004 in the proportion of ill cases in outbreaks where the infection was spread via animal contact/exposure or environmental means cannot be modelled using the binomial error structure (Fig. 4.16a,b).

Figure 4.16 a-b: Proportion of ill cases from outbreaks, animal contact or environmental
The proportion of ill cases from outbreaks of known mode of transmission between 1996 and 2004 that were spread via animal contact/exposure or environmental means (a) with and (b) without the two large outbreaks.



4.5 Discussion

As in Scotland, there has been little statistical analysis of temporal trends in *E. coli* O157 cases or outbreaks in the United States. Rangel and colleagues published a report on the epidemiology of all outbreaks between 1992 and 2002, however it contains only brief, general comments on trends, with no formal statistical analysis of temporal trends (Rangel et al. 2005). An unpublished report from the FoodNet surveillance program (Marcus et al. 2004a), using a Poisson model to look at changes in sporadic and overall case incidence between 1996 and 2002 represents one of the few statistical analyses of temporal trends. The report, however, involves only data from the nine states or regions covered by the FoodNet surveillance program and no actual values for the measure of statistical significance are provided. Given the prominence of *E. coli* O157 infections in the United States in recent years, as illustrated by the two multi-state outbreaks caused by contaminated produce in late 2006, a better understanding of outbreak and case epidemiology, including temporal trends is important. Improved knowledge on temporal trends may help to provide insight into the success or failure of preventative measures and suggest as to which modes of transmission or case types may be of greater importance in terms of prevention and research. It is thus surprising that there has been little published

research on long term trends. Therefore the study in the current chapter was conducted to determine whether it was possible to model temporal trends in U.S. *E. coli O157* case and outbreak numbers using generalised linear models, and if so, whether any of the modelled trends were statistically significant.

The methods and models used in the analyses were identical to that used for the analyses of the Scottish data, but there were unique challenges associated with the United States data. *E. coli* O157 cases have been reportable (to the federal government) in the United States since 1994 (Rangel et al. 2005), however only by 2002 was reporting of individual cases mandatory within all 50 states, with all 50 states reporting case totals to the NNDSS for the first time in 2002. As a result case numbers during the earlier portion of the study time period may not represent the true prevalence. However, a lack of a mandate to report individual cases within a state does not mean that outbreak cases were not reported to the CDC. Thus the number of total cases reported to the NNDSS may not include outbreak cases which were included in CDC outbreak data sets. The NNDDS data set may also vary from the outbreak data set due to the differences in the way dates are assigned to cases (Centers for Disease Control and Prevention 2005e). The difference between data sets, though, is expected to have been minimal as only five states did not report case data to the CDC during the time period of the study and only two states did not report data after 1999, with all states reporting by 2002. These five states do not represent a major proportion of total cases, and in many cases, still reported outbreaks.

More significant were the major significant changes to the surveillance systems in 1998 and 1999 (see Section 4.2.3.2), as well as the composition of the study data set from two separate data sets as detailed in 3.2.1.2. Firstly, the start of the enhanced surveillance program and the switch to new reporting forms resulted in a substantial increase in the number of foodborne *E. coli* O157 outbreaks reported to the CDC beginning in 1998 (Lynch et al. 2006). Since outbreaks where the suspected mode of transmission was non-foodborne were reported to the CDC using state forms until 2006, and the resultant data kept in a separate database (personal communication Thai An Nguyen, CDC), the above changes are less likely to have affected the reporting of non-foodborne outbreaks. However, because there was no standardised

reporting form, the data collected on non-foodborne outbreaks may be less consistent both in terms of definitions and in the strength of evidence associated with the putative mode of transmission.

Additionally, data for all non-foodborne outbreaks and foodborne outbreaks reported between 1999 and 2004 was obtained from an updated data set provided by the CDC (Thai An Nguyen, CDC), while the data for foodborne outbreaks prior to 1998 was obtained from an earlier dataset (Liz Blanton, CDC). This earlier data set had not undergone 'data cleaning', a process which amends the data to reflect the most recent information on case numbers and/or transmission mode, adjusts the entries so that the data corresponds with the most recent variable definitions, and corrects any errors in prior data entry. In particular, for foodborne outbreaks prior to 1998, the number of laboratory confirmed cases was recorded as the number of confirmed cases. Starting in 1998 though, the number of ill and positive cases recorded on CDC form 52.13 only included primary cases. Thus the possibility that any decrease in the number of ill and positive outbreak cases could be due to the exclusion of secondary cases after 1997 must be considered.

Given the major changes to data collection and the subsequent affect on the data, as illustrated by the substantial increase in the number of reported outbreaks, it was determined that it was not appropriate to analyse the temporal trends from 1996 to 2004. For the purposes of this study, therefore, only the data between 1998 and 2004 was included in the temporal trend statistical analyses, though the 1996 – 1997 data points were illustrated and considered in discussion of the analyses where appropriate.

4.5.1 Trend analyses

While it was not possible to model the temporal trends in *E. coli* O157 outbreaks for the entire period between 1996 and 2004 due to the significant changes in the surveillance system, many trends between 1998 and 2004 could be modelled using simple linear or generalised linear models with quasipoisson, Poisson or quasibinomial error structures or log transformation of the data.

4.5.1.1 Where trends could not be modelled

Where appropriate models could not be constructed, there appeared to be two predominant reasons. For trends in the number ill, and the proportion of total outbreak cases each year that were ill, the inclusion of the cases from the two large outbreaks – Washington County Fair and Layton Avenue Sizzler – appeared to be the primary reason for the inability to model the data. While these outbreaks occurred during the middle of the study time-period and thus did not have such a significant potential to affect trends as with the Wishaw Outbreak in Scotland, their disruptive effect on the ability to model the trends suggest that they might be statistical outliers, though not to the degree of Wishaw.

The Washington County Fair Outbreak may have been an outlier in our study because the number of reported number of ill was likely inflated by cases that were actually due to *Campylobacter* - there were 45 persons positive for *Campylobacter* (New York State Department of Health 2000) in the outbreak. Though one of the most common gastrointestinal illnesses in the United States (Centers for Disease Control and Prevention 2005a), *Campylobacter* has only been implicated in 12 waterborne outbreaks between 1998 and 2004 (Centers for Disease Control and Prevention 2004c). Waterborne *Campylobacter* outbreaks between 1998 and 2004 were substantially larger than those caused by *E. coli* O157, whether or not outbreaks with multiple implicated organisms were included (calculations from unpublished data, CDC). Therefore it may be that a quarter or more of the ill cases associated with the Washington County Outbreak were infected with *Campylobacter* rather than *E. coli* O157. Since not all persons were tested for infection, it is clearly impossible to determine which symptoms were a result of which bacterial infections(s), but if it were possible to omit the *Campylobacter* cases from the totals, the reduction in the number of ill cases may mean that the outbreak no longer appears to be an outlier.

The other reason for the inability to fit appropriate models was a low number of outbreaks, which affected primarily the attempts to model the number and proportion of ill cases from outbreaks spread by animal/environmental contact or exposure. With a maximum of just five outbreaks per year and three years where no more than two outbreaks were reported, the amount of data available over the seven-year time

period for analysis may have been too insufficient and too variable for the construction of a regression model with enough power to detect changes in trends (see Section 2.3.5.2). It would appear from a visual inspection of the data however, that while the number of outbreaks spread via animal contact peaked in 2001, there has been an increase in the number of cases in these outbreaks.

Another factor in the inability to model the data from animal/environmental outbreaks was the disproportionate number of ill cases from outbreak cases in 2001, which may be linked to the first appearance of environmentally transmitted outbreaks in that year. Prior to 2001, no outbreaks were considered to be the result of environmental spread, but in 2001 there were two environmental outbreaks with a combined 145 ill cases between them. It is not known whether the category 'environmental' was used prior to 2001, whether no previous outbreaks of this transmission method had been recorded or whether there was increased awareness by public health officials in investigating the causes of outbreaks in agricultural areas. The large numbers of ill cases, as opposed to cases confirmed by laboratory tests, in animal/environmental outbreaks may also reflect the location of the environmental outbreaks, which both involved infections that resulted from airborne or surface contact with bacteria in barns where infected animals were or had been present (Croft et al. 2002; Varma et al. 2003). Since in both outbreaks, food was consumed by all or many of the persons in the contaminated areas, there were many pathways for infection.

4.5.1.2 Modelled trends

Though the incidence of *E. coli* O157 infection has substantially decreased between 1996 and 2003/2004 (Centers for Disease Control and Prevention 2005d) and increased in 2005 and 2006, based on data from 10 states as part of the FoodNet Surveillance (Centers for Disease Control and Prevention 2007f), and there was an increase in the number of reports to NNDSS between 1996 and 2000 (Centers for Disease Control and Prevention 2003a), between 1998 and 2004 there was no statistically significant trend in the number of *E. coli* O157 cases in the United States. However the presence of a statistically significant decrease in overall case numbers between 1999 and 2004, regardless of the inclusion of the two large

outbreaks in 1999 and 2000 ($p<0.047$), and the noticeable increase in number of cases between both 1997 and 1998 and 1998 and 1999, suggest that the impact of enhanced surveillance and substantial reporting forms updates may have extended into 1999. One reason for the effect of enhanced surveillance being spread out over more than one or years might have been that the revised reporting form for foodborne outbreaks was not released until August 1999 (Centers for Disease Control and Prevention 2004a), though the number of outbreaks decreased in 1999 as compared to 1998. After a significant increase in the number of outbreaks between 1997and 1998, resulting from the onset of enhanced surveillance (Lynch et al. 2006), the number of outbreaks remained relatively stable with no statistically significant trend between 1998 and 2004. However, when the two large outbreaks – Layton Avenue Sizzler and Washington County Fair – are excluded, there is a statistically significant decrease in the number of ill cases from outbreaks.

There appears to be a decrease in the role of water as a mode of transmission, as there are statistically significant decreases in the number of waterborne outbreaks and ill cases from waterborne outbreaks as well as the proportion of outbreaks and outbreaks cases that were waterborne. However, given the issues regarding reporting of waterborne outbreaks (see section 4.2.3.2), this apparent decrease may reflect a change in reporting rather than a true increase. Additionally, there appears to be a decrease in the dominance of food as a mode of transmission, with a statistically significant decrease in the number of ill cases from foodborne outbreaks. No statistically significant trend was seen in the number of foodborne outbreaks, but the plot appears to suggest a decreasing trend. Since the power calculations in Chapter 2.3.5.2 specifically indicate that there would be less than 50% power to detect a 25% decrease from year to year, it is very possible that the analysis was not powerful enough to pick up a statistically significant trend.

4.5.2 Conclusion

The results from the analyses in this chapter indicate that some trends in *E. coli* O157 outbreaks and cases in the United States between 1998 and 2004 can be modelled using simple linear models. These models suggest that like the Wishaw Outbreak, the large Washington County Fair and Layton Avenue Sizzler outbreaks do have a

statistically significant effect in some trends, and could possibly be considered as outliers. In addition, whilst there has not been a statistically significant change in the numbers of total confirmed cases (outbreak and sporadic) or outbreaks during the study period, there has been a statistically significant decrease in the number of ill cases from outbreaks. In particular, there have been statistically significant decreases in the number of ill cases from outbreaks spread primarily by food and water. Thus, as in Scotland, food and the occurrence of one or more disproportionately large outbreaks appear to have played significant roles in *E. coli* O157 trends. These factors will also be examined in the next chapter, Chapter 5, when trends in Canada are analysed.

Chapter 5 -- An analysis of *E. coli* O157 temporal trends in Canada: 1996 – 2003

5.1 Introduction – *E. coli* O157 in Canada: 1996 – 2003

The first cases of *E. coli* O157 in Canada were recognized in 1982 (Woodward et al. 2002), when 18 residents of a home for the aged in Ottawa were infected in the third ever reported *E. coli* O157 outbreak (Lior 1983). Notification became mandatory at a national level in 1990 (Wilson et al. 1997); *E. coli* O157 infections are counted together with all verotoxin producing *E. coli* (VTEC) serotypes. Since 1990, reported rates have stayed above 3 cases annually per 100, 000 population (Public Health Agency of Canada 2006) and there have been over 20,000 reported cases of VTEC *E. coli* infection (Public Health Agency of Canada 2006). Between 1994 and 2003, approximately 95% of these cases were O157 (Sockett et al. 2006). Research in the province of Ontario and on data from the National Notifiable Disease Registry suggests that surveillance data is very incomplete and that cases may be underreported at a rate as high as 98% (Michel et al. 1998; Michel et al. 1999; Thomas et al. 2006). A multi-provincial database of outbreaks occurring between 1996 and 2003 was assembled in 2006 and indicated that upwards of 100 outbreaks were reported to provincial health authorities during this period (Sockett et al. 2006).

Ontario has generally had the highest number of cases per year, followed closely by Quebec, and then Alberta and British Columbia (Public Health Agency of Canada 2006).The highest rates of infection have tended to be in Manitoba, Quebec and Alberta (Public Health Agency of Canada 2006). Caution must be taken when looking at the rates in some of the more sparsely populated provinces because a small number of cases can result in a very high rate. For instance, the Northern Territories had only 6 cases in 2000, but because the population is only around 42,000, the rate was 14.82 per 100,000.

As in the U.S., many outbreak-related cases have been reported as being linked to consumption of improperly cooked minced beef or other beef products (Todd 2000). Research has shown that higher livestock density, especially cattle, may be linked to increased occurrence of human STEC infection (Michel et al. 1999; Valcour et al. 2002). Outbreaks have also been traced to raw milk (McIntyre et al. 2002), lettuce (Woodward et al. 2002), apple cider (Health Canada 1999) and salami (MacDonald

et al. 2004). Direct animal contact has been implicated, as has contaminated water such as in a 2001 outbreak linked to a beach in Montreal (Bruneau et al. 2004). The largest reported outbreak in Canadian history was the large Walkerton Outbreak (O'Connor 2002), which involved more than 1000 ill cases. It remains the largest outbreak in North America, and the second largest ever, second only to the1996 Sakai City Outbreak in Japan, which sickened over 6300 people (Michino et al. 1999). (See Section 2.4.1.3 for more detail).

Though overall rates of reported infection have generally been declining since 2000 (Sockett et al. 2006), *E. coli* O157 outbreaks continue to occur in Canada, as a number of outbreaks in 2006, including the minced-beef linked outbreak in Manitoba (Winnipeg Regional Health Authority 2007), one linked to beef donairs in Alberta (Honish et al. 2007) and outbreaks in Ontario (ProMED-mail 2007) indicate. Yet, as in Scotland and the United States, the literature on outbreaks has been mainly limited to descriptive analyses of specific outbreaks (Honish et al. 2007; Michel et al. 1998), with comprehensive studies limited, in large part, by the lack of a systematically collected and validated, long-term multi-provincial database of outbreaks. The recent compilation of a historical national enteric disease outbreak data set, which includes *E. coli* O157 outbreaks, has provided an opportunity for an analysis of long term outbreak trends in Canada. This chapter has been conducted to analyse *E. coli* O157 outbreaks in Canada between 1996 and 2003.

5.2 Materials and methods

5.2.1 Study data sets

5.2.1.1 Outbreak data set

Data set

The data set for Canadian *E. coli* O157 outbreaks in 1996 – 2003 has been obtained from Kathryn Doré at the PHAC. The standard definition of an outbreak in Canada is "an incident in which two or more persons experiences similar illness after a common source exposure. An outbreak is identified through laboratory surveillance or an increase in illness that is unusual in terms of time and/or place. An outbreak is confirmed through laboratory and/or epidemiological evidence." (Foodborne Illness Outbreak Protocol). However, outbreaks with fewer than two cases were reported to

the PHAC data set, which suggests that provincial records were incomplete and/or there may have been alternate outbreak definitions used by some provinces/territories for at least part of the study time period.

The extent of the variation in outbreak definitions between provinces cannot be fully quantified because only a few of the definition(s) used in the data provided by the various provinces in and territories for the report were available in the information provided by PHAC and included in published reports. However, outbreaks with fewer than two cases have been excluded prior to analyses in this study, and thus all outbreaks in this data set are assumed to meet the PHAC definition of an outbreak: an outbreak involving "two or more clinically or laboratory confirmed cases..." (Tinga et al. 2006) and for the purposes of this chapter an outbreak is defined as involving two or more ill cases. In this chapter, an ill outbreak case is defined as an incidence of infection that is clinically confirmed, but not necessarily laboratory confirmed. An ill and positive outbreak case is defined as a laboratory confirmed case.

Data manipulations

For the purposes of this chapter, changes were made to the original data set provided by PHAC, which contained 110 confirmed or suspected *E. coli* O157 outbreaks between 1996 and 2003, as explained below.

Detailed notes on the data and data variables provided by each province/territory to PHAC and the re-coding by the PHAC researchers were made available, and information from these notes and from correspondence with staff at PHAC has been used to adjust and recode the data. In particular, the notes were used in the process of establishing definitions for number of ill and number of ill and positive cases that can best accommodate provincial/territorial data and which are most compatible with the data collected from the United States and Scotland.

Firstly, after discussion with researchers at PHAC, the ten outbreaks for which no information was provided on case numbers - either laboratory, clinical or total - were excluded from the data set. One outbreak was confirmed by further conversation with PHAC to be a duplicate of another outbreak in the data and so was also deleted.

Another outbreak, which had only one confirmed case listed and no data for ill or total cases, was omitted because it did not meet the outbreak definition for this thesis. Data on six outbreaks not included in the data set but reported in detail in the literature, and confirmed as reliable by PHAC, were added to the data set (David et al. 2004; MacDonald et al. 2000; MacDonald et al. 2004; McIntyre et al. 2002; Williams et al. 2000).

Further changes were made to correct inconsistencies in the data and to make variables compatible with those used for the Scottish and United States data, as listed below (Table 5.1).

Table 5.1: Changes made to variables in the Canadian *E. coli* O157 outbreak data set

The changes by outbreak made to the data set received from PHAC prior to analyses.

ID Number	Variable Changed	Old Value	New Value	Explanation/Reference
558	Lab_cases Clin_cases	18 ---	10 17	Corrections as per Galanis et al
3769	Num_cases	11	12	Corrections as per Health Canada, 1999
894	Clin_cases	---	0	Added value as per Honish et al 2005
4930	Clin_cases	69	39	Corrected as lab_cases + Clin_cases should equal Num_cases. Information on data cleaning and comparison between data set and papers on Walkerton Outbreak suggest that in Ontario Clin_cases include Lab_cases.
1161	Clin_cases	5	4	See above
4152	Clin_cases Num_cases	25 25	0 44	Corrected to reflect the fact that there were 44 lab_cases, so at least 44 Num_cases
1477	Num_cases	5	13	Changed to reflect the fact that there were 13 lab_cases, so must have been at least 13 num_cases
1643	Clin_cases	2	0	Corrected so lab_cases + Clin_cases = Num_cases.
1164	Clin_cases	4	0	See above
1168	Clin_cases	3	0	See above
7515, 1425	Clin_cases Lab_cases	0 0	--- ---	Removed improbable values of zero because Num_cases was 3
5543	Lab_cases	273	174	Removed campylobacter cases as per Health Canada, 2000
4968, 2238, 3 904, 4117, 5330, 2398, 2452, 3054, 3540, 3728, 4363, 5800, 5886, 6115, 1961, 2703, 2863, 3752, 3061	Clin_cases			Changed Clin_cases so that it equals Num_cases – Lab_cases (see 4930)
1425, 1477, 1961, 2452, 2863, 3061, 3540, 3752, 4152, 4165, 4224, 5330, 5886, 6115	Mode_Transmission	---	Unknown	Defined all blanks as unknown
3054	Mode_Transmission Mode_Text_Other	Agriculture Cow manure	Environmental	Changed to be compatible with Scottish dataset
910	Mode_Transmission Mode_Text_Other	Agriculture Cattle farm	Environmental	Changed to be compatible with Scottish dataset
2003 Oct – anonymous province/territory	Mode_Transmission Mode_Text_Other	Agriculture Petting Zoo	Animal Contact	Changed to be compatible with Scottish dataset
7536,779,867,7506	Mode_Transmission	Other	Unknown	

The variable Lab_Cases, defined as the number of "Primary and secondary cases for which there are positive lab results", has been re-named *Ill and Positive Cases* and the variable "Clinical_Cases", defined as the "sum of lab-confirmed and clinical cases", was renamed *Ill Cases*. Clinical cases, in this instance, are defined as the number of "Primary and secondary cases for whom there are no lab results, but who have clinical presentation consistent with case definition and who can be linked epidemiologically (by place or time) to the outbreak".

Outbreaks have been assigned to the same categories for the mode of transmission variable as used for the Scottish data, based on the information for putative mode of transmission provided in the original variables mode_transmission and mode_text_other. If the value for mode_transmission in an outbreak was recorded as foodborne, waterborne, unknown or person-to-person, the value was directly copied to the new putative mode of transmission variable. Outbreaks for which mode_transmission was Agriculture were coded as *Environmental/Animal Contact*. Where no value or the value Other was provided for mode_transmission, the new mode of transmission value was *unknown*. Again, it is important to note that all modes of transmission designations are putative as the data is based on information recorded by provincial or regional public authorities and may not have been systematically collected or validated. Thus while not explicitly indicated in the results section, all analyses involving mode of transmission should be considered to involve putative, not confirmed mode of transmission.

The final data set for this thesis has 104 Canadian outbreaks, with variables for number of ill cases, number of ill and positive cases and mode of transmission (foodborne, waterborne, person to person, animal/environmental or unknown).

5.2.1.2 Total cases data set

Data set

Data on the reported confirmed cases (outbreaks and sporadic cases) for 1996 to 2003 has been obtained from the Disease Surveillance On-Line section of the Public Health Agency of Canada website (www.phac.aspc.gc.ca/dsol-smed/index.html). A confirmed case in this data set is, for the purposes of this chapter, considered to be equivalent to an ill and positive case.

Data manipulation

The data reported in 2000 included all the epidemiologically related cases from the Walkerton Outbreak, not just the ill and positive cases (Sockett et al. 2006). Thus the non-confirmed cases are omitted from the data prior to any analyses. The number of non-confirmed – e.g. ill, but not laboratory positive - cases has been calculated by subtracting the number of ill and positive cases as listed in the outbreak data set from the number of ill as listed in the outbreak data set.

5.2.1.3 Study time period

Only data between 1996 and 2003 were included in this thesis in order to allow comparison with the data from Scotland, and also because compiled data on outbreaks in Canada prior to 1996 was not provided in the PHAC data set.

5.2.1.4 Reporting issues – Walkerton Outbreak

With 1346 ill cases, the Walkerton Outbreak is nearly ten-fold greater in size than the next largest outbreak in Canadian history. When only the 147 ill and positive cases, are considered, it is still larger as compared to other reported Canadian outbreaks between 1996 and 2003. Given this relatively large number of ill cases and the results presented in this chapter which suggest that the Walkerton Outbreak may be an outlier, the analyses of the Canadian data will be conducted both including and excluding the Walkerton Outbreak. Since the Walkerton Outbreak occurred in the middle of the time period being analysed and thus does not dramatically increase the first or last data point, it is not expected that it will have the same degree of influence on trends as did the Wishaw Outbreak which occurred in the first year of the study time period (see Section 2.4.1.1), but it is important to ensure that any disproportionate influence is recognised and considered in interpretation of analysis results. Another issue regarding the Walkerton Outbreak is the large number of ill cases attributed to infections with other pathogens – 97 with *Campylobacter jejuni* and 20 with other organisms (O'Connor 2002). In the data set for this thesis, these 117 cases have been excluded from the number of ill cases; it is still possible that even more of the ill cases were the result of these other organisms.

While the Wishaw Outbreak in Scotland could also be analysed as a series of outbreaks (see section 3.2.3.2), a similar analysis will not be performed for the Canadian data because no such identifiable cohorts exist in the Walkerton Outbreak. All but 39 of the 1346 ill persons were infected from direct consumption of water from the common municipal water supply, and the 39 secondary cases can all be traced to contact with a person infected from the water supply.

5.2.2 Variables selected for the analyses

Table 5.2: Variables used in analyses
The variables used in analyses of case and outbreak trends in Canada, 1996 - 2003

Variable	Definition
Total reported cases	Number of laboratory confirmed verotoxin-producing *E. coli* cases
Number of Outbreaks	Number of outbreaks
Number of ill cases from outbreaks	Total number of cases from outbreaks, including those clinically confirmed, but not laboratory confirmed
Number of ill cases per outbreak – overall and foodborne	Total number of cases per outbreak (all outbreaks, and those spread via food)
Number of Outbreaks spread via: a) Food b) Water c) Animal/Environmental d) Person to person (PTP)	Number of outbreaks in which the putative mode of transmission, as categorised per thesis definitions, (mode) was: Food Water Animal/Environmental Person to person
Number of ill cases from outbreaks spread by a, b, c or d	Number of total cases from outbreaks where the putative mode of transmission, as categorised per thesis definitions, was a, b, c or d
Proportion of outbreaks spread via a, b, c or d	Proportion of outbreaks with known mode, as per the thesis definition, in which the mode of transmission was: a, b, c or d
Proportion of ill cases from outbreaks spread via a, b, c or d	Proportion of cases from outbreaks in which the mode was a, b, c or d

The variables for the analysis of the Canadian data were selected to allow analyses of basic trends in Canadian *E. coli* O157 cases and outbreaks, as well as comparison between trends in Canada, Scotland and the United States. Firstly the trends in the number of reported VTEC cases will be modelled (Table 5.2), then trends in the number of ill cases in outbreaks (Table 5.2) and finally trends in the number of ill cases per outbreak (Table 5.2).

Also to be modelled are trends in the putative modes of transmission, in terms of both the numbers and proportions of total outbreaks and ill cases each year. In particular, the number of outbreaks and ill cases from each putative mode of

transmission will be modelled, as will the proportion of outbreaks and ill cases associated with each mode of transmission. The final variable to be modelled will be the number of ill cases per outbreak spread via food. Since the mode of transmission for approximately one third of outbreaks is listed as 'unknown' in the thesis data set, the above proportions will be calculated excluding those outbreaks where mode of transmission is listed as 'unknown'. All variables including or derived from numbers of ill cases will be analysed with and without the Walkerton Outbreak cases, to check for any influence on trends by the outbreak as discussed in section 5.2.1.4.

5.2.3 Analyses

5.2.3.1 Descriptive analyses

The data from 1996 to 2003 will be described in terms of the variables mentioned in Table 5.2: the annual numbers of total reported cases, ill outbreak cases and outbreaks will be described, as will be the breakdown of outbreaks by mode of transmission.

5.2.3.2 Statistical analyses

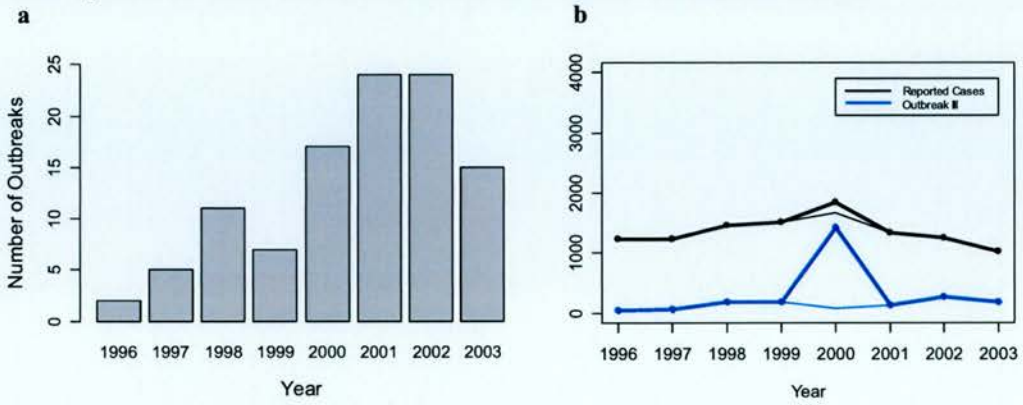
For this chapter, temporal trends between 1996 and 2003 are analysed using GLMs and linear models as specified in Chapter 2. All variables involving case numbers including waterborne cases were analysed with and without the cases from the Walkerton outbreak included (see Section 2.4.1.3).

5.3 Results – Descriptive

Between 1996 and 2003, there were 12,042 cases of verotoxigenic *E. coli* reported to Health Canada (Public Health Agency of Canada 2006), the majority of which were serotype O157 (Sockett et al. 2006). This total includes cases from 104 outbreaks, with the annual number of outbreaks ranging from 2 to 24 (Fig. 5.1a). Within these outbreaks there were 2485 ill outbreak cases reported to PHAC (yearly range 41 to 1429) (Fig. 5.1b). The number of ill and positive cases was only available for 92 outbreaks, for which there were 865 total cases between 1996 and 2003. Outbreaks had a median of 5 ill cases (range 2 to 1346) and 3 ill and positive cases (range 0 to 174).

Figure 5.1a-b: (a) Number of outbreaks per year & (b) Number of reported cases and outbreak ill & positive cases

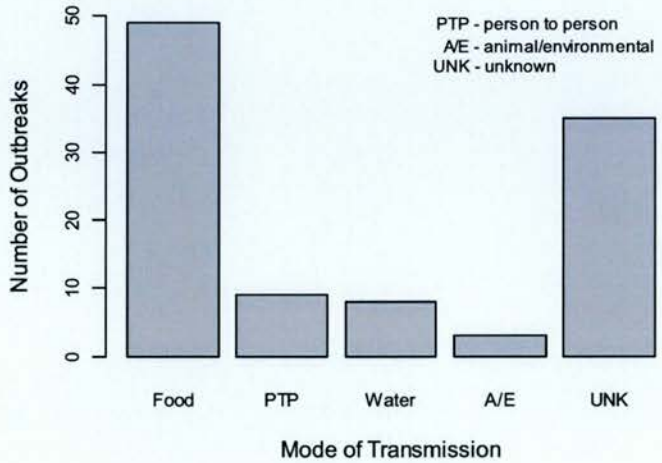
In (a), the number *E. coli* O157 outbreaks per year and in (b) reported VTEC cases and *E. coli* O157 ill cases from outbreaks per year. The number of reported cases is indicated by a thick black line and the number of ill cases in outbreaks by a thick blue line. The narrow lines between 1999 and 2001 represent the trends when the cases from the Walkerton Outbreak are omitted.



Of the 104 outbreaks, 69 (66.3%) had a putative mode of transmission recorded. Within the 104 total outbreaks, those categorised as foodborne were most common (44.7%), followed by unknown (34.3%), with person to person, animal/environmental and waterborne transmission each implicated in more than 2% of outbreaks (Figure 5.2).

Figure 5.2: Number of outbreaks by mode of transmission

The number of outbreaks by mode of transmission



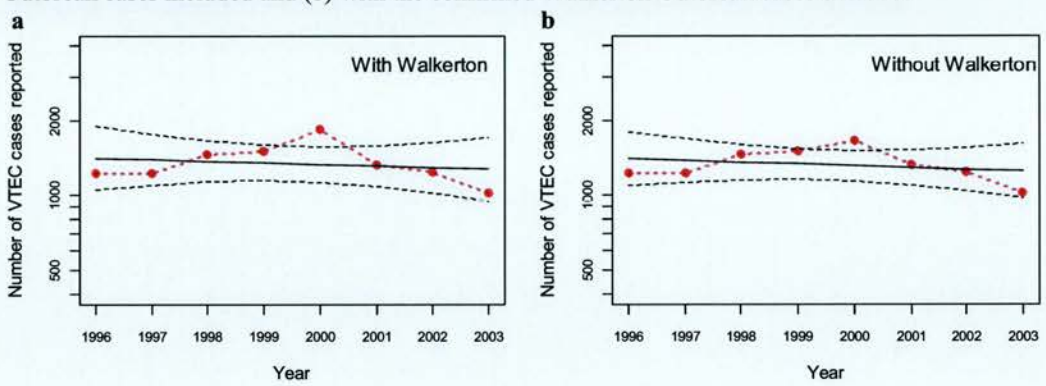
5.4 Results – Statistical analyses

5.4.1 Total case and outbreak variables

5.4.1.1 Total reported VTEC cases

Temporal trends in the number of total reported VTEC cases, regardless of whether the cases from the Walkerton Outbreak are included, can be modelled using linear models with log10 transformation of the data. There are no significant trends in the number of VTEC cases reported to PHAC when the Walkerton Outbreak is included ($F_{1,6}$=0.3, p=0.62; Fig. 5.3a) or excluded ($F_{1,6}$=0.5, p=0.53; Fig. 5.3b).

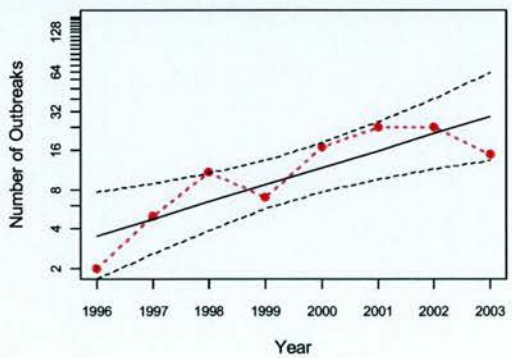
Figure 5.3 a-b: The number of VTEC cases reported to PHAC (a) with and (b) without the Walkerton Outbreak
The number of VTEC cases reported to Health Canada, by year, (a) with the confirmed Walkerton Outbreak cases included and (b) with the confirmed Walkerton Outbreak cases excluded



5.4.1.2 Number of *E. coli* O157 outbreaks, by year

A linear model is used to analyse the trend in the log transformed number of outbreaks per year between 1996 and 2003. The analysis reveals that this trend is statistically significant and increasing ($F_{1,6}$=14.5, p=0.01, Fig. 5.4).

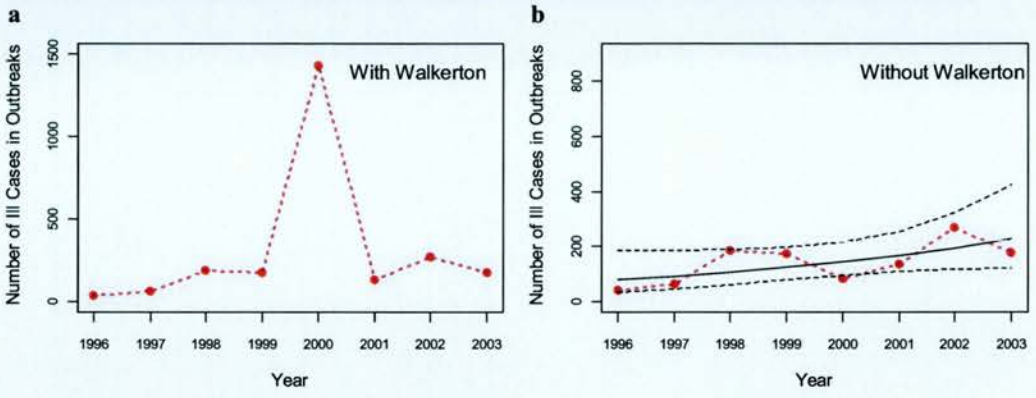
Figure 5.4: Number of *E. coli* O157 outbreaks
The number of *E. coli* O157 outbreaks per year, with the best fit line for the model.



5.4.1.3 Number of ill cases in outbreaks, by year

The temporal trend in the number of ill cases in outbreaks cannot be appropriately modelled when the Walkerton Outbreak cases are included (Fig. 5.5a). When the 1346 Walkerton Outbreak cases are excluded, and a GLM with a quasipoisson error structure is used, the trend in the number of ill cases in outbreaks per year between 1996 and 2003 is not statistically significant ($F_{1,6}$=4.3, p=0.08; Fig. 5.5b).

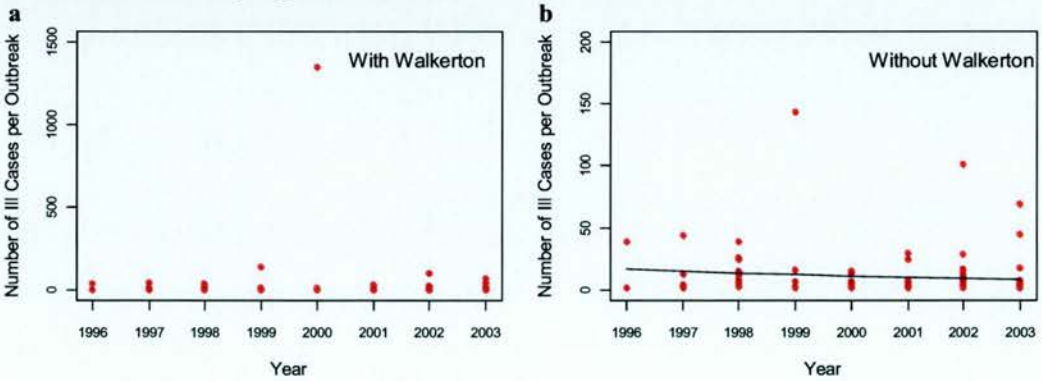
Figure 5.5 a-b: Number of ill cases in outbreaks, with and without the Walkerton Outbreak
The number of ill cases in outbreaks (a) with and (b) without the Walkerton Outbreak included.



5.4.1.4 Number of ill cases per outbreak

The temporal trends in the number of ill cases per outbreak cannot be appropriately modelled when the Walkerton Outbreak is included (Fig. 5.6a). After the Walkerton Outbreak cases are excluded, the trend can be modelled using a GLM with quasipoisson errors, but it is not statistically significant ($F_{1,6}$=1.1, p=0.29; Fig. 5.6b).

Figure 5.6 a-b: Number of ill cases per outbreak
The number of ill cases per outbreak (a) with and (b) without the Walkerton Outbreak. The lines are the best fit for the quasipoisson model.



5.4.2 Number of outbreaks, by mode of transmission – overall differences

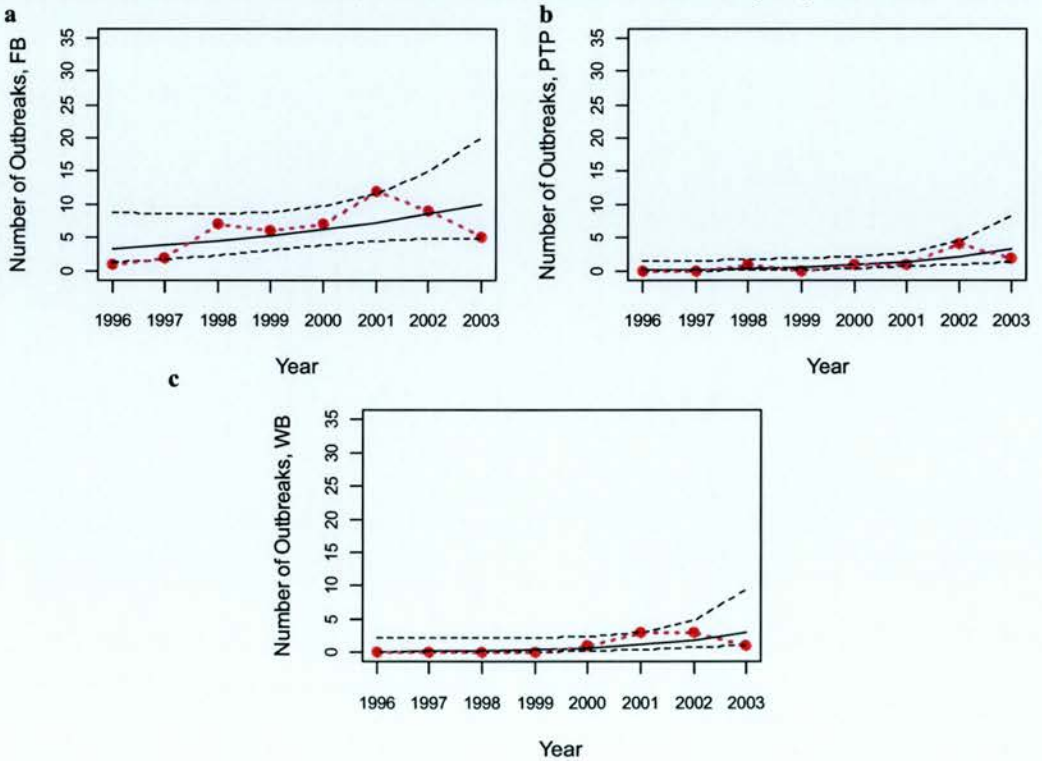
Over time, there is a statistically significant increase in the number of outbreaks per year (see section 5.4.1.2), but there is no evidence of any statistically significant difference between modes of transmission in the temporal trends of the number of outbreaks ($F_{3,24}$=2.4, p=0.09). However, the individual trends will still be fitted because the trends for each mode of transmission are of interest.

5.4.2.1 Number of foodborne, person to person and waterborne outbreaks

The trends in the number of outbreaks by mode of transmission can be modelled using a GLM with quasipoisson or Poisson errors. No statistically significant trend exists in the number of foodborne outbreaks each year ($F_{1,6}$=3.8, p=0.10; Fig. 5.7a), but there are statistically significant increasing trends in the number of person to person ($F_{1,6}$=9.7, p=0.02; Fig. 5.7b) and waterborne outbreaks (z=2.2, p=0.03; Fig. 5.7c). It should be noted that no waterborne outbreaks were recorded until 2000.

Figure 5.7 a-c: Number of outbreaks: foodborne (FB), person to person(PTP) & waterborne(WB)

Trends between 1996 and 2003 in the number of outbreaks in which the mode of transmission was (a) foodborne, (b) person to person and (c) waterborne. With the exception of waterborne for which a Poisson error structure was used, the best fit lines are for models with quasipoisson error structure.



5.4.3 Proportion of outbreaks, by mode of transmission – overall differences

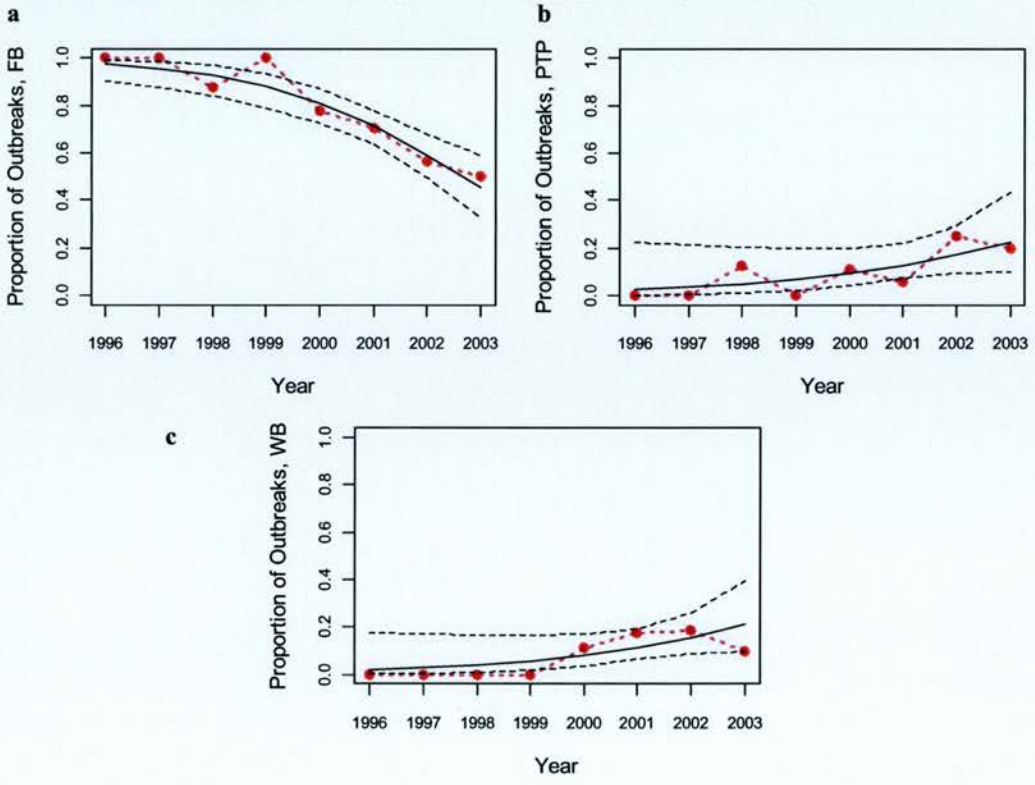
There is a statistically significant difference between modes of transmission in the temporal trends of the proportion of outbreaks ($F_{3,24}$=15.1, p<0.001). Therefore, the individual trends will be fitted below.

5.4.3.1 Proportion of outbreaks: foodborne, person to person and waterborne

When GLMs with quasibinomial error structures are used, there is a statistically significant decreasing trend between 1996 and 2003 in the proportion (out of outbreaks of known mode of transmission) of outbreaks that are foodborne ($F_{1,6}$=37.1, p<0.001; Fig. 5.8a). The trends in the proportions of outbreaks that were person to person ($F_{1,6}$=4.4, p=0.08; Fig. 5.8b) and waterborne ($F_{1,6}$=5.5, p=0.06; Fig. 5.8c) transmission are not statistically significant.

Figure 5.8 a-c: Proportion of outbreaks: foodborne (FB), person to person (PTP) and waterborne (WB)

Trends in the proportion of outbreaks (out of outbreaks with a known mode of transmission), in which the mode of transmission was (a) foodborne, (b) person to person and (c) waterborne.



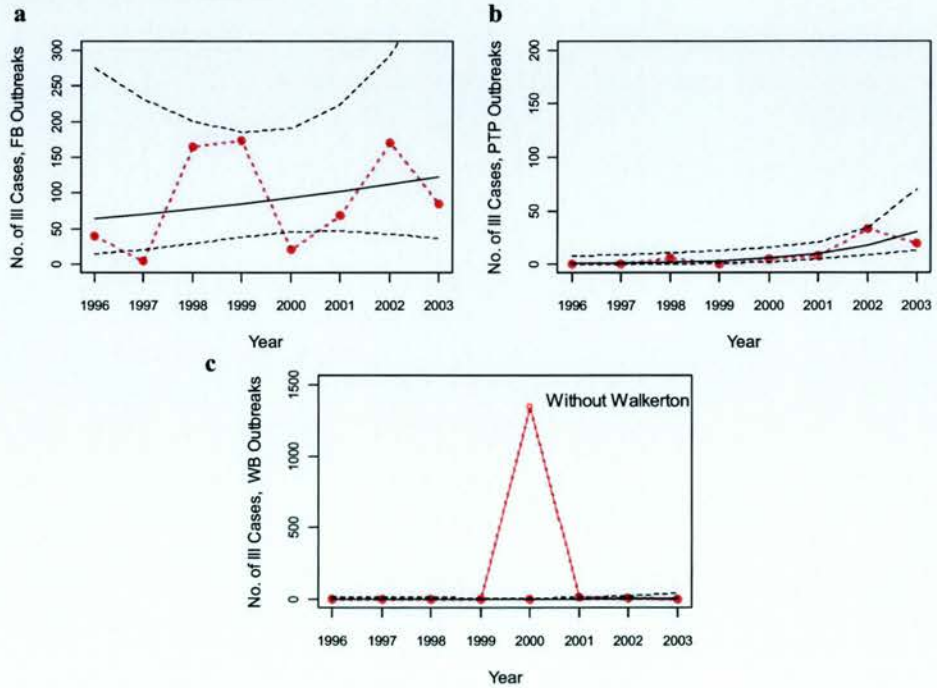
5.4.4 Number of ill outbreak cases

The differences in trends of the number of ill cases in outbreaks between modes of transmission when the Walkerton Outbreak cases were included will not be modelled because an adequate simple linear model cannot be found for the overall number of ill cases (see section 5.4.1.3). However, when the Walkerton Outbreak cases are excluded, there is a significant difference between modes of transmission in temporal trends ($F_{3,24}$=3.2, p=0.04). Thus the trends will be individually modelled.

5.4.4.1 Number of ill outbreak cases: foodborne, person to person & waterborne

The trend in the number of ill cases in foodborne and person to person outbreaks can be adequately modelled using GLMs with quasipoisson error structures. However, while the trends in the number of ill cases from foodborne ($F_{1,6}$=0.5, p=0.50; Fig. 5.9a) and waterborne outbreaks (without Walkerton) ($F_{1,6}$=4.5, p=0.08; Fig. 5.9c) are not statistically significant, the trend in the number of ill cases in person to person outbreaks increases statistically significantly ($F_{1,6}$=14.7, p=0.01; Fig 5.9b). The trend in the number of ill cases in waterborne outbreaks with the Walkerton outbreak cannot be adequately modelled using simple linear models (Fig. 5.9c).

Figure 5.9 a-c: Number of ill cases in outbreaks: foodborne, person to person and waterborne
The number of ill cases in outbreaks where the mode of transmission is (a) foodborne, (b) person to person and (c) waterborne. In plot c, the bright red dot and line indicate the plot if the Wishaw Outbreak cases are included



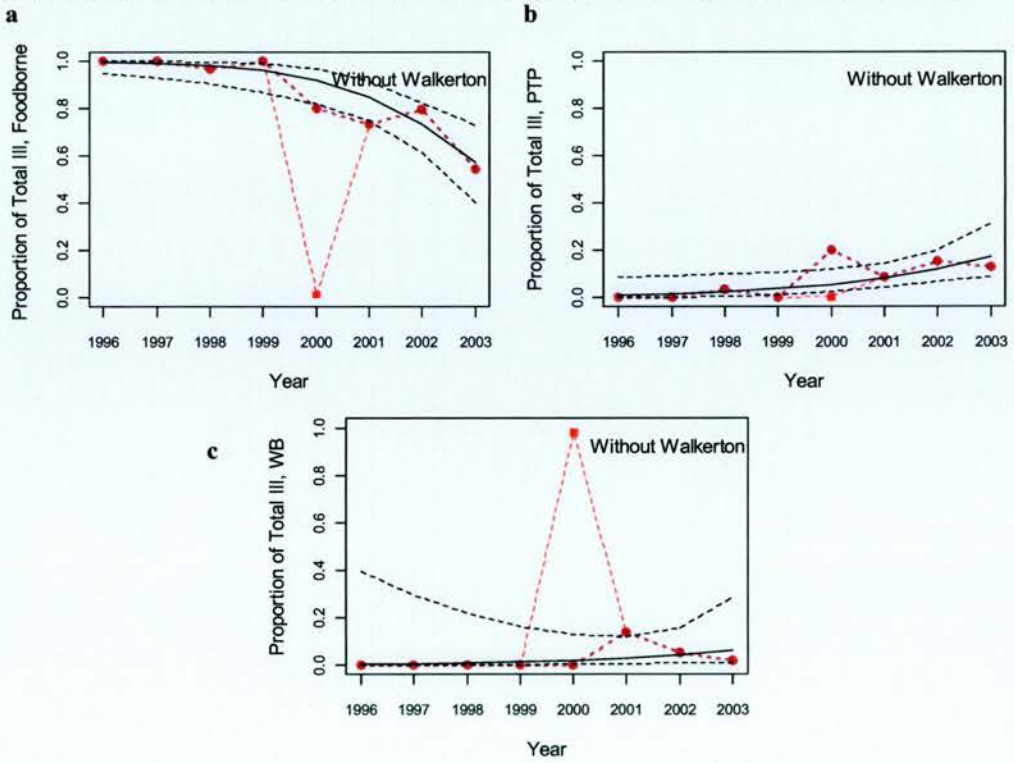
5.4.5 Proportion of ill cases in outbreaks, by modes of transmission

When the cases from the Walkerton Outbreak are excluded, there is a statistically significant difference between modes of transmission in the trends in the proportion of outbreaks (with known modes of transmission) ($F_{3,24}$=17.9, p<0.001). Therefore the individual trends are modelled separately.

5.4.5.1 Proportion of ill outbreak cases: foodborne, person to person & waterborne

When the cases from the Walkerton Outbreak are excluded, the trends for the proportion of ill cases from both foodborne ($F_{1,6}$=32.0, p=0.001; Fig. 5.10a) and person to person ($F_{1,6}$=9.9, p=0.02; Fig. 5.10b) outbreaks are statistically significant. However, the trend in the proportion of ill cases that are waterborne is not significant ($F_{1,6}$=1.5, p=0.27, Fig. 5.10c).

Figure 5.10 a-c: Proportion of ill outbreak cases: foodborne, person to person & waterborne
The proportion of ill cases from outbreaks of known mode of transmission where the mode of transmission was (a) food, (b) person to person and (c) waterborne. The light red dots and lines indicate what the proportion of ill cases would be with the Walkerton Outbreak cases included.



5.5 Discussion

In Canada, analysis of temporal trends in *E. coli* O157 outbreaks, either descriptively or statistically, has been limited. Woodward and colleagues provide a listing of major outbreaks between 1985 and 2000 in their overview of *E. coli* O157 in Canada (Woodward et al. 2002), however any in depth analysis of outbreaks has been hampered by a lack of a nationally representative, systematically collected and validated outbreak data set. Summary tables and listings of outbreaks in 1994 and 1995 were published in 2000 by Health Canada and a report on outbreaks in 1997 and 1998 was issued in 1999, but the latter did not provide a line listing of outbreaks (Tinga et al. 2006). The National Microbiology Laboratory has issued yearly reports on confirmed cases of enteric diseases, available online since 1999 (National Laboratory for Enteric Pathogens 2002; National Laboratory for Enteric Pathogens 2006). These reports include outbreak listings, however the reports state that the lists are not complete, in some years may include case clusters which are not epidemiologically linked and the listings do not include data on the number of ill cases or specifically indicate the mode of transmission (National Laboratory for Enteric Pathogens 2002; National Laboratory for Enteric Pathogens 2006). In addition, only a descriptive summary of the data is provided. The new national outbreak data set has been collated and summarised in an internal report by Tinga and colleagues in 2006 (Tinga et al. 2006), but the report provides only outbreak counts for individual pathogens and analyses of putative modes of transmission and outbreak summaries are broken down by overall pathogen type and not by pathogen.

Though national case rates have decreased in recent years, rates have remained over 3 per 100,000 and over 5 per 100,000 in some provinces (Public Health Agency of Canada 2006) and outbreaks, at least two in the past year relating to meat products (Honish et al. 2007; Winnipeg Regional Health Authority 2007), are still reported. Given the continuing role of *E. coli* O157 as a clinically significant enteric pathogen, research into temporal trends is of importance to:

- determine whether trends in overall case numbers are reflected in outbreak trends
- indicate whether current preventative measures may be having an effect on case and outbreak trends and

- suggest which future directions in prevention and research might be most appropriate.

The new availability from PHAC of a pan-Canadian database is a first step towards making such analyses possible. The purpose of the chapter is therefore to determine whether temporal trends in Canadian *E. coli* O157 cases and outbreaks can be modelled using simple generalised linear models and if so, which of the trends are statistically significant.

Despite using the same methods and materials as with Scotland (Chapter 3) and the United States (Chapter 4), the Canadian data involved the greatest challenges with regards to analysis because of the nature of the recently compiled PHAC data set. The data set is acknowledged to be incomplete (Tinga et al. 2006), with only four provinces providing data for the entire period (1996 – 2003). Additionally, some outbreaks that are summarised in peer-reviewed journals are not included in the data set. Also neither a consistently used outbreak definition, nor a standard outbreak reporting form existed at the time of data collection. Thus, the exact data collected – in terms of variables and variable definitions - varied between and within provinces, and so required extensive work by PHAC researchers, and further work in this thesis to classify the data using post-hoc standard definitions. As detailed in section 5.2.2.1, the data set has been adjusted/recoded to best conform to definitions that are compatible with the data collected from the United States and Scotland and to include outbreaks detailed in the peer-reviewed literature, but not included in the original data set. Information from the National Laboratory for Enteric Pathogens reports was not included because data was only provided on laboratory confirmed cases.

Since it was known for which years each province provided data (Tinga et al. 2006), it was possible to determine whether these omissions might have had a significant impact on the results of the analyses. Much of the missing data was from provinces or territories where very small numbers of total VTEC cases were reported to PHAC through the NND database. Thus, the only missing data that was likely to have statistically significant effects on analyses was that of three provinces from which large numbers of cases were reported, and from which three or more years of data

were missing. The potential effect of this missing data was examined by running additional analyses of some variables with the data from each of the three provinces omitted in turn and then with the data from all three provinces omitted (not shown).

The lack of a standardised reporting form also may have impacted the determination of mode of transmission because the selection of mode by public health official could have been influenced by the options provided. Also, no information was given about the strength of evidence linking the putative mode of transmissions to outbreaks. For instance, in one provincial database there was no option for either waterborne or foodborne when entering mode of transmission (Tinga et al. 2005). Recoding of data by PHAC attempted to address this issue by basing the determination of mode of transmission on all available data. However in situations where mode was ambiguous or not given, the mode of transmission had to be coded as 'unknown', which may have, in part, accounted for nearly a third of the outbreaks having mode of transmission listed as unknown. Again, where mode of transmission was provided, its validity was rarely substantiated and the investigator's level of certainty was not known.

While the notifiable disease data on total VTEC cases has been collected since 1990 in a standardised fashion, it still presented challenges because data are not subdivided by VTEC serotype in the Notifiable Diseases Online system (Public Health Agency of Canada 2006). Also the outbreak cases reported to this system likely varied from the outbreak cases captured in the outbreak database, as suggested by the differences in the outbreaks listed in the dataset provided for this thesis (Tinga et al. 2006) and the yearly summaries by the National Laboratory for Enteric Pathogens (National Laboratory for Enteric Pathogens 2002). Thus it is not considered appropriate to use the data to determine the number of sporadic cases or the proportion of cases that came from outbreaks. However, the data is still considered to be sufficient as an approximation of the total number of *E. coli* O157 cases because non-O157 serotype cases comprise less than 5% of the total (Woodward et al. 2002).

5.5.1 Discussion of statistical analyses

For all analyses involving the number or proportion of ill cases, overall or by mode of transmission, the Walkerton Outbreak may have been an outlier as appropriate

simple linear models could not be constructed when the cases were included. With 174 laboratory confirmed cases, the Walkerton Outbreak was larger than all but one other outbreak in the data sets analysed in Chapters 2 – 5; the outbreak is only disproportionately large in terms of ill cases. However, the large number of ill cases is attributable, in part, to large number of infections from other organisms during the Walkerton Outbreak, including 97 confirmed *Campylobacter* cases (O'Connor 2002), and 20 confirmed cases from other pathogens. The presence of these confirmed cases of infection by other organisms suggests that there are likely to be non-confirmed, non- *E. coli* O157 cases amongst the number of ill cases. Thus, the outbreak's true influence could be much better ascertained in analyses based on the number of ill and positive cases. However, such analyses were not possible because data on the number of ill and positive cases was missing for 12 of the 104 outbreaks in the data set. In particular, since nine of these outbreaks took place between 1997 and 1999 inclusive, analyses of trends in the number of ill and positive cases may not be accurate prior to 2000, the year when the Walkerton Outbreak occurred.

Alternatively, a more accurate value for the number of ill cases might be ascertained based on knowledge about the average size of waterborne outbreaks for both *E. coli* O157 and *Campylobacter* and/or the average ratio of ill to ill and positive cases in *E. coli* O157 outbreaks. As there have been no other large waterborne outbreaks reported in Canada, the best comparisons are with outbreaks in the United States. The outbreak in Cabool, Missouri (Swerdlow et al. 1992), also the result of contaminated municipal water supply, and the Washington County Outbreak in New York (Centers for Disease Control and Prevention 1999a), in which a fairground water supply became contaminated, both involved large numbers of ill cases where only a portion of cases were ever tested for infection. The latter outbreak also involved a number of *Campylobacter* cases. In Cabool, where no other pathogen was isolated, *E. coli* O157 was isolated from approximately 60% of patients tested – 19 of 25 with bloody diarrhoea and 2 of 7 with non-bloody diarrhoea. These figures could potentially be used to estimate the number of ill cases with *E. coli* O157 infection in Walkerton based on the known number of cases tested, the percentage positive and the distribution of symptoms if more information was available about the symptoms of the ill persons associated with the Walkerton Outbreak.

The total number of reported *E. coli* O157 cases per year in Canada appears to have peaked in 1989 (Woodward et al. 2002), with the number reported to have decreased in recent years (Sockett et al. 2006). The results from this study, however, do not reveal a statistically significant trend away from no change in the number of total VTEC cases. Yet, while trends in the number of ill cases from outbreaks and the number of ill cases per outbreak were not statistically significant, there was a significant increase in the number of *E. coli* O157 outbreaks reported between 1996 and 2003. Caution must be taken in interpreting the analysis of the number of outbreaks because, though excluding any one of the three provinces discussed above as having potentially important missing data does not make a significant difference, excluding all three provinces results in the trend in the number of outbreaks being non statistically significant. This suggests that the increasing trend may, in part, be caused by the greater amount of missing data in the earlier part of the time period covered in the data set. Additionally, there appears to be an increase in the number of ill cases (excluding Walkerton), but the simulations in Chapter 2.3.5.2 suggest that that there may not have been enough power to detect a statistically significant trend.

There is no overall statistically significant difference in the trend in number of outbreaks between modes of transmission, though there are statistically significant increases over time in the numbers of outbreaks spread person to person and by water. The trends in the proportion of outbreaks did vary statistically significantly between outbreaks, with there being a statistically significant decline in the proportion of outbreaks that were foodborne. However, the statistical and epidemiological significance of these trends should be interpreted with caution because the trends may be statistically significantly affected by the data that is known to be missing, particularly from the earlier section of the data set time period.

Though a statistically significant trend does not exist in the number of ill cases from outbreaks (without Walkerton), there is a statistically significant difference in trends in the number of ill cases between modes of transmission. While trends in the number of foodborne and waterborne outbreaks show no statistical significance, there has been a statistically significant increase over time in the number of ill cases from outbreaks where the predominant mode of transmission was person to person

contact. There is also a shift in the proportions of outbreak ill cases, with an increase in the proportion of ill cases coming from person to person outbreaks and a decrease in the proportion from foodborne outbreaks, both statistically significant.

If these changes in predominant outbreak mode of transmission reflect actual case numbers and not reporting artefacts (e.g. incomplete reporting), it could reflect a true shift in outbreak epidemiology from outbreaks being primarily spread via food to outbreaks being predominantly spread via water and person to person contact, and/or an increasing thoroughness in outbreak investigations. In Scotland the importance of non-foodborne modes of transmission has been recognized, with a listing of outbreaks between 1994 and 2003 appearing to show a shift away from food being a predominant factor in outbreaks (Strachan et al. 2006). In Canada, the data seem to suggest that such a shift may also have taken place, since as per the data set, only one person to person outbreak was reported prior to 2000, the first waterborne outbreak was reported in 2000 and the first outbreak involving animal or environmental contact was not reported until 2001. Studies have also established a positive association between cattle density and human infection in parts of Canada (Michel et al. 1999; Valcour et al. 2002), with routes of infection related to cattle including well water contamination (O'Connor 2002) and direct contact, with the potential for further person to person spread (Wilson et al. 1997).

However, the lack of outbreaks attributed to person to person, waterborne and animal/environmental spread prior to 2000 (with the one exception) and the resulting statistically significant trends, may be linked to biases within data collecting. Firstly, *E. coli* O157 outbreaks in Canada have been long associated with minced beef (Todd 2000), but only more recently linked to animal contact and other food items (Woodward et al. 2002). Thus earlier outbreak investigations may have been more likely to have focused on foodborne modes of transmission. The trends may also be linked to the 'large proportion' of missing data or missing values in the data set (Tinga et al. 2006). With approximately a third of outbreaks having an unknown mode of transmission, real trends may be masked by the lack, or ambiguity, of information which resulted from surveillance systems which used variables that did not effectively capture data about mode of transmission in outbreaks involving

pathogens like *E. coli* O157. In one province, the possible values for mode of transmission included neither food nor water. Instead, options such as 'animal/person', 'faecal-oral', 'indirect' and 'droplet' (Tinga et al. 2005) were used. These options do not provide enough information to properly classify transmission in terms generally used in regards to gastrointestinal outbreaks, such as foodborne, person to person and waterborne. Since outbreaks where the information on mode of transmission was not clear have to be recoded as unknown, the number of unknown outbreaks is likely to have been overestimated. Such an example illustrates why it is vital to have surveillance systems and/or database variables that are suited to the infection or illness involved, and which use variables and values that can adequately capture information for future reference, and where data collection methods are explicit, rigorous, systematic and where the resultant information can be validated.

In addition, the reported modes of transmission may also reflect the trends in the particular provinces/territories which provided data rather than the entire country, over the time period of the study. For instance, since there was more missing data in the first half of the study period, in particular from provinces with high rates of infection as per the data on the reported confirmed cases from the Notifiable Diseases Online database, the trends seen in this chapter may reflect best the trends in province/territories with lower rates of infection and/or provinces/territories that provided more complete information.

5.5.2 Conclusions

Overall the analyses suggest that in Canada there may have been an increase in the number of *E. coli* O157 outbreaks, accompanied by the emergence of non-foodborne outbreaks (i.e. waterborne, person to person), as reflected in both case and outbreak numbers and proportions. If these results are true, this may suggest that current prevention efforts (personal communication, Carole Tinga, 2007) may not be effectively targeting outbreaks spread by non-foodborne means. However, given that only five provinces provided data for every year in the study period, and outbreak listings in other sources (National Laboratory for Enteric Pathogens 2002) suggest that even this data is not complete, these results need to be interpreted with caution. It is hoped that the publication by PHAC of a national enteric disease outbreak

summary, assembled using data from the current data set and other sources, the introduction of a web-based outbreak summary data (which will provide a consistent format in which to record outcomes of outbreak investigations), as well the results of this study, will encourage both a renewed effort to build a standardised, comprehensive prospective national enteric outbreak summary system as outlined by Tinga et al. (Tinga et al. 2006).

Having now analysed the individual trends in *E. coli* O157 cases and outbreaks for Scotland, United States and Canada, the next chapter will go on to compare data and trends across countries in order to determine whether statistically significant differences in trends occur between countries.

Chapter 6 -- Comparison of Temporal Trends in *E. coli* O157: Scotland, United States and Canada

6.1 Introduction

Having now analysed the temporal trends in *E. coli* O157 within each of the three countries included in this section of the thesis (Scotland, Canada and the United States), the focus will now shift back to the concept of comparison between countries. As introduced in the Prospectus and in Chapter 2, the goal of the first section of this thesis is to determine whether or not there are significant differences in temporal trends between countries. The presence or absence of statistically significant differences in trends could suggest as to the underlying factors that influence *E. coli* O157 epidemiology. For instance, a lack of many statistically significant differences might suggest that *E. coli* O157 epidemiology is primarily influenced by factors that are similar across countries.

Though there have a been a few between-country statistical comparisons of temporal trends in diseases such as cancer (Venzon & Moolgavkar 1984), gonorrhoea (Van der Heyden et al. 2000) and childhood diabetes (EURODIAB ACE Study Group 2000), no such statistical analysis has ever been done involving *E. coli* O157. To date, the only comparisons of *E. coli* O157 across countries have been purely descriptive. In 1994, Waters and colleagues published a broad overview of *E. coli* O157 in Scotland and Alberta from 1987 to 1991 (Waters et al. 1994). In the paper, comparisons were made on a purely descriptive basis with the authors highlighting a number of similarities between the two countries. Reports from the EnterNet *Salmonella* and *E. coli* O157 surveillance network have included tables of data from member countries, but as of yet, no statistical comparisons have been published (European Commission 2006; European Commission 2007).

As discussed in Chapter 2, the lack of comparisons between countries is likely to caution in regards to comparing data collected via different surveillance systems. However, with the onset of improved and/or enhanced surveillance programs across Scotland, Canada and United States in the last decade, it seems appropriate for a comparison to now be attempted. Thus, in this chapter, the temporal trends in variables for which appropriate models could be found for Scotland (Chapter 3), United States (Chapter 4) and Canada (Chapter 5) will be now be statistically compared. The results of these analyses will be presented, along with a discussion of

the issues involved in comparing data from different countries and how these issues affect the interpretation of the results.

6.2 Materials and methods

6.2.1 Data sets

The final total case and outbreak data sets used in Chapters 2, 3, 4 and 5 will be compared, with data from 1996 to 2004 used for Scotland, 1998 to 2004 for United States and 1996 to 2003 for Canada. 1996 and 1997 data from the United States was omitted due to a major change in reporting procedures, as detailed in section 4.2.1.2. See sections 3.2.1, 4.2.1 and 5.2.1 for detailed descriptions of the data sets.

6.2.2 Variables

Variables for the analyses are selected based on the appropriateness of comparison between Scotland, the United States and Canada. As listed in Table 6.1, the variables which will be compared are:

- number of total confirmed cases
- number of ill cases per outbreak
- number of outbreaks
- number of ill cases.

The number of ill cases and outbreaks will be examined both overall and by mode of transmission (again, as with the individual analyses, mode of transmission was based on the data provided. The strength of the evidence linking the putative mode of transmission to the outbreak is not always known). Also compared were the proportions of outbreaks and ill cases from outbreaks with a known mode of transmission which are foodborne, waterborne and person to person.

For the total number of confirmed cases (outbreak and sporadic)(Table 5.1), data is taken from the HPS data set (Scotland), NNDSS data set (United States) and the PHAC data set (Canada). For Canada, the non-confirmed cases from the Walkerton Outbreak are omitted prior to any analysis. The Scottish and United States data sets include only confirmed O157 cases, while the Canadian data set is known to contain a small percentage (~3%) of non-O157 VTEC cases (Health Canada 2000b; Sockett et al. 2006; Woodward et al. 2002).

Data for the total numbers of outbreaks (Table 6.1) are taken from the respective outbreak data sets provided by each country and detailed in sections 3.2.1, 4.2.1 and 5.2.1. Outbreak case number data – yearly case totals and cases per outbreak - are analysed only in terms of the number of ill cases because of the large number of missing values for the number of ill and positive case variable in both the United States (29 out of 310) and Canadian data sets (12 out of 104).

In analyses involving (putative) mode of transmission (Table 6.1), comparisons are carried out between the three main modes of transmission as categorised in Chapters 2, 3, 4 and 5: foodborne, waterborne and person to person. Outbreaks spread by animal or environmental means were not compared due to the low number of outbreaks spread by animal/environmental means in Canada (3 out of 104) and the lack of appropriate models for the trends in Scotland.
In contrast to Scotland where outbreaks can be categorised as 'multiple modes including (or excluding) foodborne', in the United States and Canada, there is no separate category for outbreaks with co-dominant modes of transmission. However since descriptions of outbreaks categorised as 'foodborne' do occasionally refer to other modes of transmission (and the strength of evidence for mode of transmission designation is not known), the mode of transmission category 'foodborne' would appear to be less specific in the United States and Canada than in Scotland. In order to account for this difference in categorisation, Scottish outbreaks designated as either foodborne or multiple including foodborne are combined into one category, referred to previously in Chapter 3 as 'foodborne element' and referred to in this chapter as simply 'foodborne'.

Outbreaks in the Scottish data set, which are categorised as 'multiple modes excluding foodborne', indicating that transmission has taken place through two or more co-dominant non-foodborne modes, are omitted from the mode of transmission analyses. These outbreaks are excluded because the category 'multiple excluding foodborne' does not exist in the American or Canadian data sets, and not enough information on non-foodborne outbreaks in the U.S. and Canada is included in the data sets to determine which, if any, would have met the criteria for 'multiple modes, not including foodborne'.

Table 6.1: Definitions for variables used in the comparison of trends between Scotland, the United States and Canada
Variables used in the comparison of trends. * indicates those variables where the definition comes from the established surveillance definition in the respective country, though the variable name may not be the term used in each country.

Variables	Scotland	United States	Canada
Total confirmed cases*	# of cases confirmed by faecal isolate	# of confirmed or probable cases. Probable cases also include clinically compatible cases that are epidemiologically linked to probable or confirmed cases or that are Shiga toxin positive and lab confirmed cases pending confirmation of H7 or Shiga toxin production.	# of laboratory confirmed verotoxin-producing *E. coli* case
Outbreak*	'in which two or more linked cases experience the same illness' and involves 'members of more than one household or residents of an institution'	Two persons (or more) becoming ill due to a common source	Involving two or more clinically or laboratory confirmed cases
Ill case*	Case confirmed by faecal sample or serum Ab + # of cases where the person was ill and epidemiologically linked to the outbreak, but the presence of *E. coli* O157 was not confirmed	Estimated number of total ill cases. [For non-foodborne outbreaks and 1998-2004 foodborne outbreaks, variable may not include secondary cases]	An incidence of infection that is clinically confirmed, but not necessarily laboratory confirmed
Number of ill cases per outbreak	Number of ill cases per each outbreak as defined above	Number of ill cases per each outbreak as defined above	Number of ill cases per each outbreak as defined above
Modes of transmission • Foodborne	Spread mainly by food or by multiple modes including food*	Food is the predominant transmission route	Food is the primary mode of transmission for the primary cases within the context of the geographic area or setting experiencing the outbreak*
• Waterborne	Spread mainly by water*	Water is the predominant transmission route	Water is the primary mode of transmission…*
• Animal/Environmental	Spread mainly by environmental contact/exposure or animal contact	Transmission via animal contact, animal exposure or environmental contamination (surface or aerosol)	An agricultural source or animal contact or animal exposure is the main mode of transmission…
• Person to person	Spread mainly via person to person	Person to Person contact is the	Person to person contact is the primary

• Unknown	contact* Method of spread not defined	predominant transmission route Transmission route not specified or specified as unknown	mode of transmission…* Primary mode of transmission for the primary cases is unknown or not specified
Proportion of outbreaks (by mode of transmission)	Proportion of outbreaks spread by each mode of transmission	Proportion of outbreaks spread by each mode of transmission	Proportion of outbreaks spread by each mode of transmission
Proportion of ill cases (by mode of transmission)	Proportion of ill cases in outbreaks spread by each mode of transmission	Proportion of ill cases in outbreaks spread by each mode of transmission	Proportion of ill cases in outbreaks spread by each mode of transmission

6.2.3 Analyses

6.2.3.1 Descriptive

All analyses involving case numbers are carried out excluding the cases from the large outbreaks (Wishaw Outbreak in Scotland, Washington County and Layton Avenue Sizzler Outbreaks in the United States and the Walkerton Outbreak in Canada) because the analyses in Chapters 3, 4 and 5 suggest that these outbreaks may be outliers in terms of case numbers, and thus the data may be best analysed with these outbreaks omitted.

For each country, the geometric mean number of ill cases per outbreak and the geometric mean number of total confirmed cases (without the large outbreaks) and outbreaks per year are calculated using R (R Development Core Team 2007). 95% confidence intervals for the geometric means are calculated using a function written for R. ANOVA analyses are used to test for overall statistically significant differences between countries in the log-transformed data. Where the overall difference is statistically significant, pair-wise comparisons between countries are performed using contrasts. The mean numbers per year of ill cases from outbreaks (overall, foodborne, waterborne and person to person) are calculated for each country. Also calculated are the mean proportions of both ill cases and outbreaks (from outbreaks with a known mode of transmission) that are foodborne, waterborne and person to person. Calculations are done using the GENMOD procedure in SAS v9.1 (SAS Institute Inc. 2003). The GENMOD procedure is also used to compare the means between countries using a GLM with a log link and Poisson or Binomial error structure.

For each of the above variables, the geometric mean or means are calculated for the rate per 100,000 persons per year, rates being calculated by dividing the variable by the 2003 population (see Table 2.1) rounded to the closest 10,000. The geometric means or means are compared between countries as specified above.

Additionally, the data is presented by country in terms of the overall proportions of both total cases and total outbreaks spread by each mode of transmission during the entire study period. For each mode of transmission, the proportion of total outbreaks and cases are compared overall between countries by using the R function prop.test

(The R Project 2007) which, in this instance, uses a Chi-squared test with a 2 by 3 contingency table to test the null hypothesis that the probabilities of success in two or more groups are the same (Armitage et al. 2002). Where there is overall statistical significance ($p<0.05$), the prop.test function is used to compare proportions between countries.

6.2.3.2 Correlation

For each country in the study (Scotland, United States and Canada), the number of ill cases and ill and positive cases for each outbreak are plotted against each other. Outbreaks for which the number of ill and positive cases is not known are excluded from the analyses. The correlation between the two variables is compared using a correlation test with Kendall's *tau*. Kendall's *tau* was selected because a non-parametric test that was distribution free (Armitage et al. 2002) was required due to non-normal distributions of the number of ill cases and the number of ill and positive cases. Stats Direct (StatsDirect Ltd. 2007) was used for calculations of Kendall's *tau* and the corresponding 95% confidence intervals.

6.2.3.3 Trends

Analysis of Covariance (ANCOVA) analyses, which combines regression and analysis of variance (Crawley 2005), are used to test for statistically significant differences between countries in the temporal trends in the variables mentioned in section 6.2.2. If there is a statistically significant interaction between the time and country explanatory variables in the ANCOVA, a statistically significant difference in temporal trends between countries is considered to exist. In addition, differences between countries and modes of transmission in the trends are tested for by including a mode of transmission variable and looking for three way interaction between time, country and mode of transmission. However, tests with three way interactions were only used for trends in the proportion of outbreaks or ill cases from outbreaks. Three way tests were not used in the analyses of trends in the number of outbreaks or ill cases from outbreaks because not all modes of transmission were considered in these analyses due to low counts. Thus it would not be appropriate to compare trends between all modes of transmission.

Two separate data sets are used for the ANCOVA models. The first, which is used for comparing the number of ill cases per outbreak, contains an entry for each outbreak with variables for country, mode of transmission, year and number ill. The second data set is used for comparison of trends involving the numbers or proportions of total confirmed cases, ill cases and outbreaks. It includes a separate entry for the number or proportion of outbreaks or ill cases for each country for each of four modes of transmission (food, water, person to person and animal/environmental) for each year. Years in which data was not available or omitted from previous analyses – 1996 and 1997 for the United States and 2004 for Canada - are not included in the data set for the analyses.

For the ANCOVA, year and country – and mode of transmission when appropriate - were entered as explanatory variables along with the interaction between the two variables, using the following log-linear model (number of outbreaks used as example):

Number of Outbreaks = $\exp(a_i + b_i * \text{Year})$ where i is different for each country

R code: Model <- glm(Outbreaks~ Year + Country + Year*Country, family=quasipoisson, data=….), where * indicates interaction between the two variables

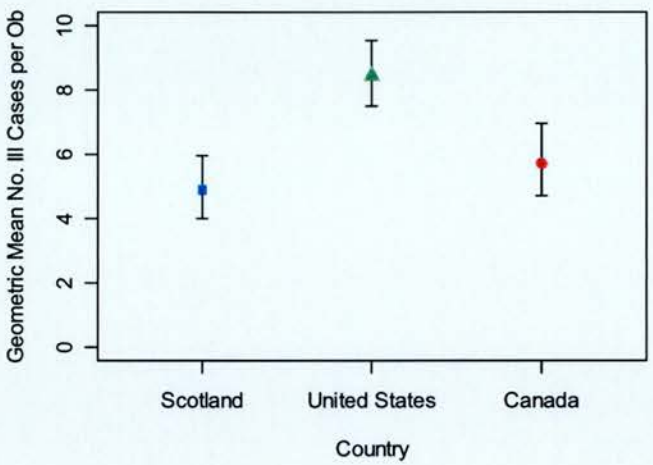
If the p value for F was less than 0.05, contrasts were used to look for significant difference in temporal trends between countries.

6.3 Results

6.3.1 Descriptive analyses – overall

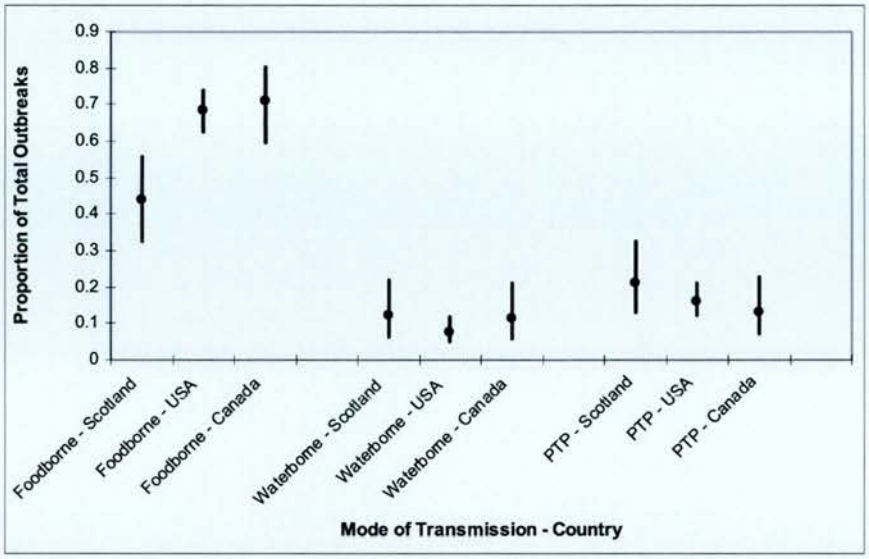
The geometric mean number of ill cases per outbreak is 4.90 (95% CI = 4.02, 5.97) in Scotland, 8.5 (95% CI =7.48, 9.54) in the United States [1998 – 2004], and 5.71 (95% CI=4.69, 6.95) in Canada (Fig. 6.1). There is a statistically significant difference between countries in the geometric mean number of ill cases per outbreak ($F_{2,482}$=8.5, p<0.001). While there is not a statistically significant difference in the geometric mean number of ill cases per outbreak between Scotland and Canada (t=0.82, p=0.416), outbreaks in the United States have a statistically significantly higher geometric mean number of ill cases per outbreak than those in Scotland (t=3.46, p<0.001) and Canada (t=2.91, p=0.004).

Figure 6.1: Geometric mean ill cases per outbreak, by country
The geometric mean number of ill cases per outbreak (Ob) for each country. The vertical bars indicate the 95% confidence intervals around the geometric mean.



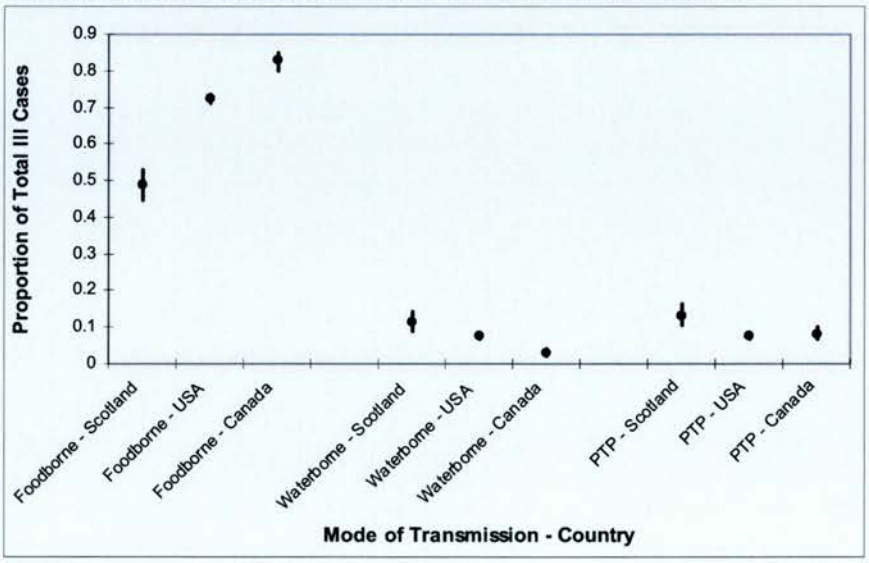
When the outbreaks for which the mode of transmission is not known are excluded, Scotland has a statistically significantly lower proportion of outbreaks during the time period being studied that are foodborne than either the United States or Canada (p<0.004; Fig. 6.2), but there are no overall statistically significant differences between the countries in the proportions of outbreaks that are waterborne or person to person (p>0.05; Fig. 6.2)

Figure 6.2: Proportion of total outbreaks by mode of transmission and country
The proportion of total outbreaks (with a known mode of transmission) spread by foodborne, waterborne and person to person (PTP) transmission by country. The vertical lines indicate the 95% confidence intervals.



There is a statistically significant difference in the proportions of ill cases in all modes of transmission ($p<0.001$; Fig. 6.3) between the three countries.

Figure 6.3: Proportion of total ill cases by mode of transmission and country
The proportion of total ill cases (with a known mode of transmission) spread by foodborne, waterborne and person to person (PTP) transmission by country with the large outbreaks from each country omitted. The vertical lines indicate the 95% confidence intervals



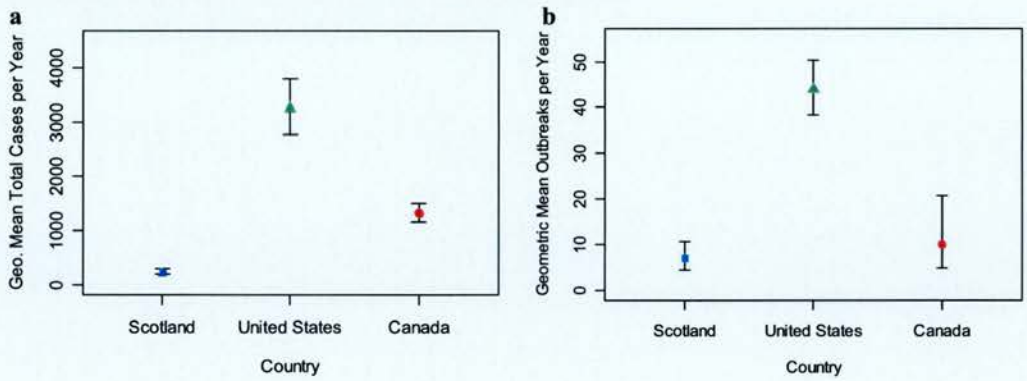
Compared to Canada and the United States, Scotland has a statistically significantly lower proportion of ill cases from outbreaks that are foodborne ($p<0.001$; Fig. 6.3),

but a statistically significantly higher proportion that are waterborne and person to person ($p<0.003$; Fig. 6.3). Similarly, in comparison to the other two countries, Canada has a statistically significantly higher proportion of ill cases from outbreaks that are foodborne ($p<0.001$; Fig. 6.3), and a statistically significantly lower proportion of ill cases outbreaks that are waterborne and person to person (except the United States) ($p<0.004$; Fig. 6.3).

6.3.2 Descriptive analyses – by year

If the large outbreaks are excluded, Scotland has the lowest geometric mean total confirmed cases per year and the United States, the highest, with the differences between countries in the geometric means all statistically significant ($p<0.001$; Fig. 6.4a). However, the United States has a statistically significantly lower geometric mean rate per 100,000 people of total confirmed cases (1.1, 95% CI =0.96, 1.3) than either Scotland (4.6, 3.7 to 5.7) or Canada (4.1, 3.6 to 4.7) ($p<0.001$).

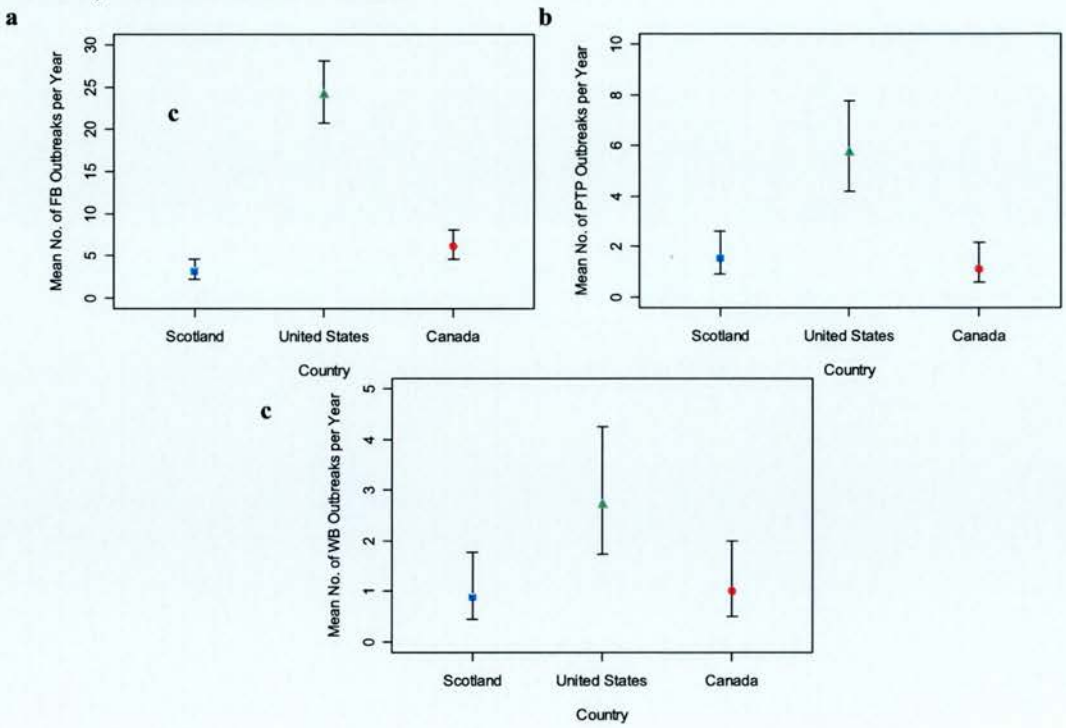
Figure 6.4 a-b: Geometric mean number of (a)total confirmed cases and (b) outbreaks
The geometric mean number of (a) Total confirmed cases excluding the large outbreaks and (b) Outbreaks. Black bars indicate the 95% confidence intervals.



The United States has a statistically significantly higher geometric mean number of outbreaks than either Scotland or Canada ($p<0.001$; Fig. 6.4b). Scotland though, has a statistically significantly higher mean rate of outbreaks (0.14, 0.10 to 0.21) than either Canada (0.03, 0.02 to 0.06) or the United States (0.02, 0.01 to 0.02)(see Table 6.2). The geometric mean number of outbreaks per year in Canada is likely to be higher than shown here because data was missing from at least one province and/territory for each year of the time period.

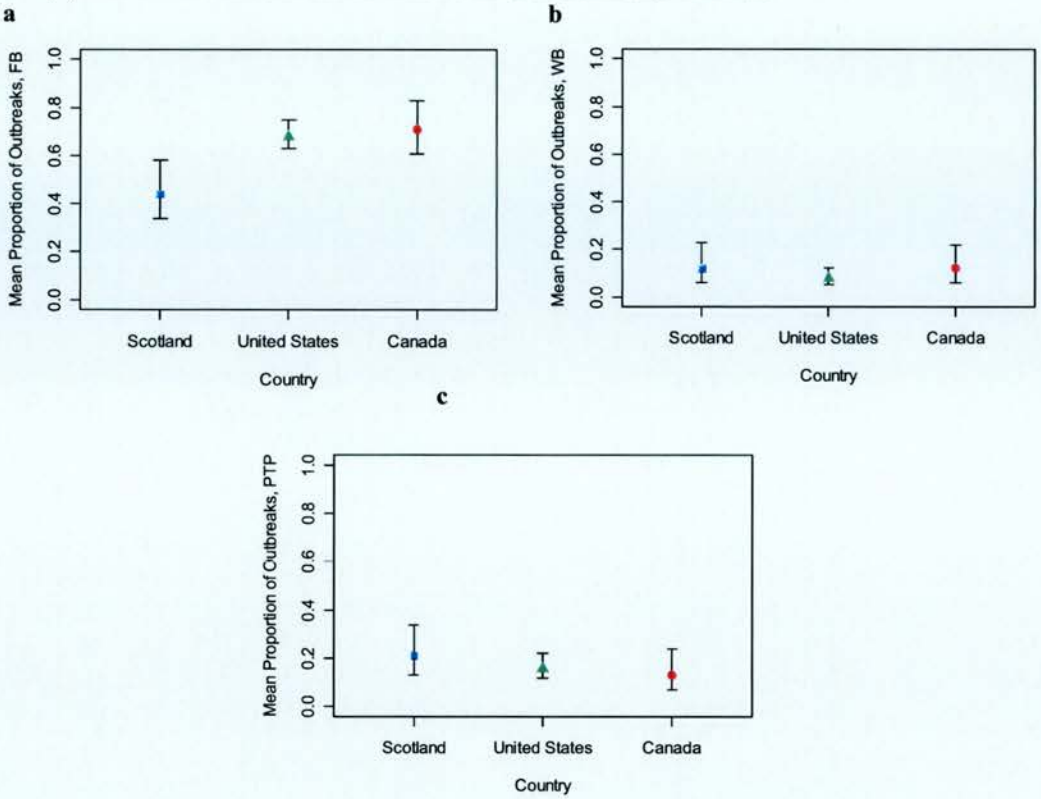
There are statistically significant differences in the mean numbers of foodborne outbreaks (see Table 6.2) between all three countries ($p<0.007$; Fig. 6.5a). For person to person outbreaks, a statistically significantly difference in the means exists between the United States and both Scotland and Canada ($p<0.0001$; Fig. 6.5b). Similarly, a statistically significantly difference between mean numbers of waterborne outbreaks (see Table 6.2) exists only between the United States and both Canada and Scotland ($p<0.018$; Fig. 6.5c). Scotland has the highest mean rate of outbreaks for all three modes of transmission, followed by Canada and then the United States, but none of the differences are statistically significant ($p>0.05$).

Figure 6.5 a-c: Mean numbers of foodborne (FB), person to person (PTP) and waterborne (WB) outbreaks per year in Scotland, United States and Canada
The mean numbers of (a) foodborne, (b) person to person and (c) waterborne outbreaks per year in Scotland, the United States and Canada



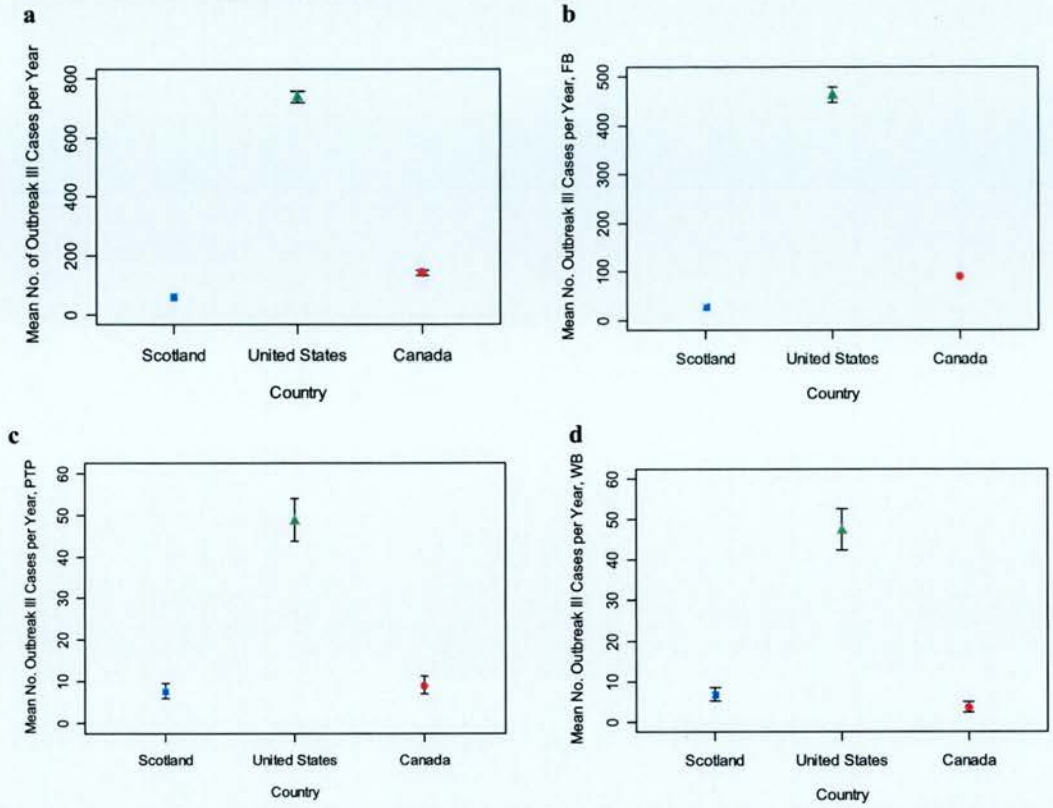
When only outbreaks with a known mode of transmission are considered, the mean proportion of Scottish outbreaks that are foodborne (see Table 6.2) is statistically significantly lower than that in the United States or Canada ($p<0.003$: Fig. 6.6a). For the proportion of outbreaks that are person to person and waterborne, no statistically significant differences in mean proportions exist between countries (Fig. 6.6b, c).

Figure 6.6 a-b: Mean proportion of yearly outbreaks that are foodborne and waterborne
The mean yearly proportion of outbreaks with a known mode of transmission that are (a) foodborne and (b) waterborne. The black lines indicate the 95% confidence intervals.



There are statistically significant differences in the mean number of ill cases per year between all three countries ($p<0.001$; Fig. 6.7a), but no statistically significant difference in the mean rate of ill cases ($p=0.053$). The mean number of ill cases per year from foodborne outbreaks was also statistically significantly different between the countries ($p<0.001$; Fig. 6.7b). For ill cases from person to person and waterborne outbreaks, the mean number in the United States is statistically significantly different from the means in both Canada and Scotland ($p<0.001$; Fig. 6.7c,d). When the mean rate of ill cases per year is considered for foodborne, waterborne and person to person outbreaks, there are no statistically significant differences between countries ($p>0.05$).

Figure 6.7 a-d: The mean number of ill cases from outbreaks – overall, foodborne (FB), person to person (PTP) and waterborne (WB)
The mean number of ill cases from (a) all outbreaks, (b) foodborne outbreaks, (c) person to person outbreaks and (d) waterborne outbreaks, all with the large outbreak cases excluded. The black lines indicate the 95% confidence intervals.



The mean proportions of ill cases from foodborne outbreaks are all statistically significantly different from each other ($p<0.001$; Fig.6.8a), while the mean proportions of ill cases from person to person and waterborne outbreaks in Scotland are significantly higher than that in Canada or the United States ($p<0.004$: Fig. 6.8b,c).

Figure 6.8 a-c: Mean proportion per year of ill cases from outbreaks – foodborne, person to person and waterborne

The mean proportion of ill cases from (a) foodborne, (b) person to person and (c) waterborne outbreaks, excluding the cases from the large outbreaks. Where visible, the black lines indicate the 95% confidence intervals.

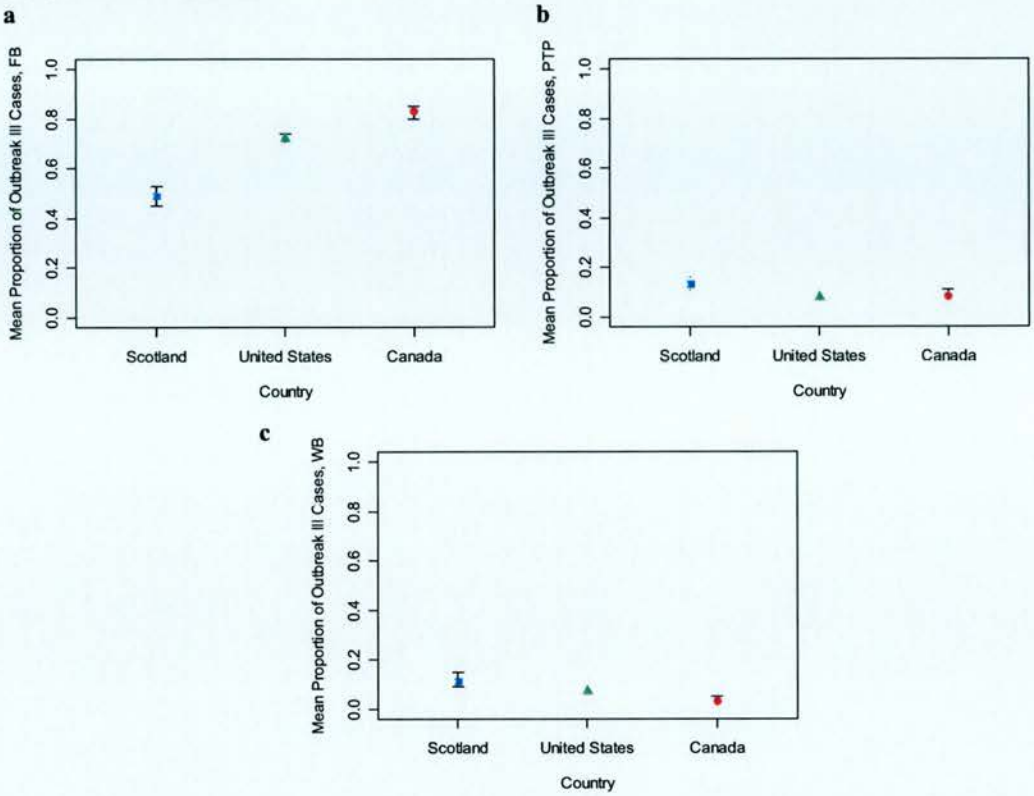


Table 6.2: Geometric mean or mean numbers and rates per year

The geometric mean numbers and rates per year for total confirmed cases and total outbreaks, and the mean numbers and rates per year for outbreaks (foodborne (FB), person to person (PTP) and waterborne (WB)) and ill cases from outbreaks (total, foodborne (FB), person to person (PTP) and waterborne (WB)). Mean proportions are given for the proportion of outbreaks and ill cases that are foodborne, waterborne and person to person. 95% confidence intervals are indicated for all means and geometric means.

	Mean/Geometric Mean Number			**Mean/Geometric Mean Rate**		
	Scotland	United States	Canada	Scotland	United States	Canada
Total confirmed cases	232.8 (186.6, 290.4)	3250.8 (2782.0, 3798.6)	1323.7 (1166.2, 1502.6)	4.6 (3.7, 5.7)	1.1 (1.0, 1.3)	4.1 (3.6, 4.7)
Outbreaks	7.0 (4.5, 10.7)	43.9 (38.2, 50.3)	10.1 (4.9, 20.6)	0.1 (0.09, 0.2)	0.01 (0.01, 0.02)	0.03 (0.02, 0.06)
FB Ob	3.2 (2.2, 4.6)	24.1 (20.8, 28.1)	6.1 (4.6, 8.1)	0.06 (0.01, 1.18)	0.01 (2×10^{-6}, 28.2)	0.02 (1×10^{-4}, 2.9)
PTP Ob	1.6 (0.9, 2.6)	5.7 (4.2, 7.8)	1.1 (0.6, 2.2)	0.03 (0.001, 1.28)	0.002 (1.1×10^{-10}, 35560.8)	0.004 (3×10^{-8}, 417.8)
WB Ob	0.89 (0.4, 1.8)	2.7 (1.7, 4.3)	1.0 (0.5, 2.0)	0.02 (0.001, 2.4)	0.001 (2.8×10^{-14}, 3.2×10^{7})	0.003 (1.3×10^{-8}, 754.5)
Proportion Ob FB	0.44 (0.34, 0.58)	0.68 (0.63, 0.75)	0.71 (0.61, 0.83)	---	---	---
Proportion OB PTP	0.21 (0.13, 0.33)	0.16 (0.12, 0.22)	0.13 (0.07, 0.24)	---	---	---
Proportion Ob WB	0.12 (0.06, 0.23)	0.08 (0.05, 0.012)	0.12 (0.06, 0.22)	---	---	---
Ill cases from Outbreaks	59.6 (54.7, 64.8)	740 (720.1, 759.8)	142.4(134.3, 150.9)	1.2 (0.65, 2.1)	0.25 (0.06, 1.1)	0.45 (0.16, 1.26)
Ill cases - FB	28.0 (24.8, 31.7)	462.7 (447.1, 478.9)	90.5 (84.1, 97.3)	0.55(0.23, 1.3)	0.12 (0.01, 1.0)	0.28 (0.08, 1.0)
Ill cases - PTP	7.6 (6.0, 9.6)	48.6 (43.7, 54.0)	9.0 (7.1, 11.3)	0.15 (0.03, 0.81)	0.01 (1.6×10^{-5}, 9.6)	0.03 (0.0005, 1.8)
Ill cases - WB	6.6 (5.1, 8.5)	47.3 (42.5, 52.7)	3.4 (2.3, 4.9)	0.13 (0.02, 0.8)	0.01 (1.4×10^{-5}, 10.2)	0.01 (1.2×10^{-5}, 9.0)
Proportion cases FB	0.49 (0.45, 0.53)	0.72 (0.71, 0.74)	0.83 (0.80, 0.85)	---	---	---
Proportion cases PTP	0.13 (0.11, 0.16)	0.08 (0.07, 0.08)	0.08 (0. , 0.11)	---	---	---
Proportion cases WB	0.11 (0.09, 0.15)	0.07(0.07, 0.08)	0.03 (0.02, 0.05)	---	---	---

6.3.3 Correlation analyses

When the correlation between the number of ill and ill and positive cases per outbreak was calculated using Kendall's Tau Test, for Scotland tau=0.756, for United States tau= 0.661 and for Canada tau=0.667(Table 6.3; Fig. 6.9).

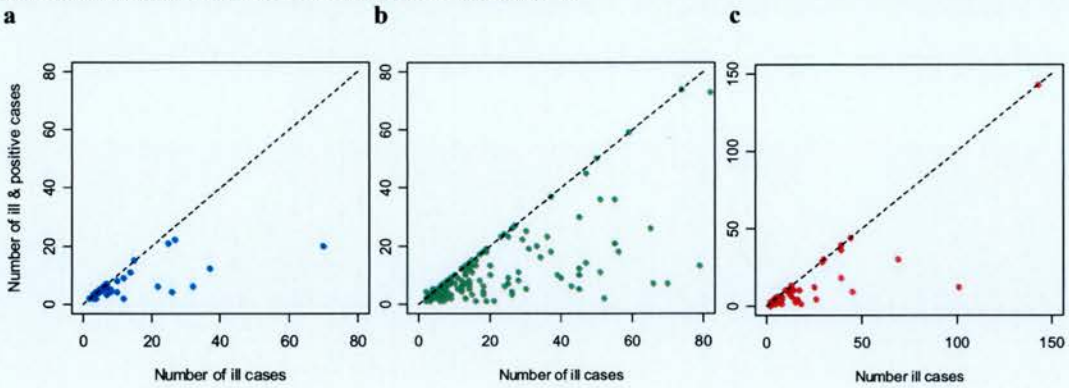
Table 6.3: Coefficients (Kendall's Tau) for the correlation between the number of ill and ill and positive cases per outbreak

Kendall's Tau for the correlation between the number of ill cases and ill and positive cases per outbreak. 95% confidence intervals are indicated in brackets. Only outbreaks for which both the number of ill and ill and positive cases are known are included.

Country		Kendall's Tau
Scotland		0.756 (0.657, 0.855) (n=70)
United States		0.661 (0.607, 0.716) (n=279)
Canada		0.658 (0.546, 0.770) (n=91)

Figure 6.9 a-c: Plots of the number of ill cases against the number of ill and positive cases per outbreak

Plots of the number of ill cases per outbreak against the number of ill and positive cases per outbreak for (a) Scotland 1996 – 2004, (b) United States 1998-2004 and (c) Canada 1996-2003. The large outbreaks are omitted, and the dotted line indicates y=x.

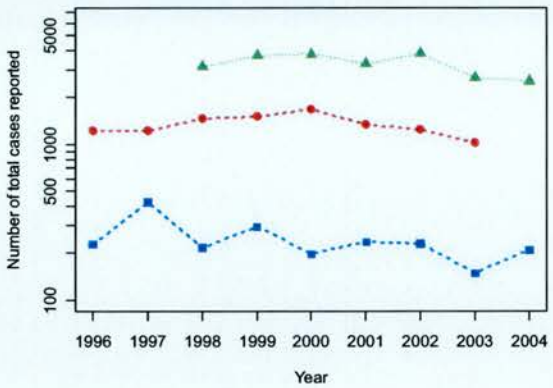


6.3.4 Trend analyses

6.3.4.1 Total confirmed cases

When the trends in the log transformed number of total confirmed cases from 1996 to 2004 in Scotland, 1998 to 2004 in the United States and 1996 to 2003 in Canada are compared, there is not a statistically significant difference between the countries in the slopes ($F_{2,18}$=0.6, p=0.560; Fig. 6.10) .

Figure 6.10: The number of total confirmed cases by country
The number of total confirmed cases by country. Scotland is indicated by blue squares, United States by green triangles and Canada by red circles. The Canadian data is for all VTEC, the rest for *E. coli* O157 only.

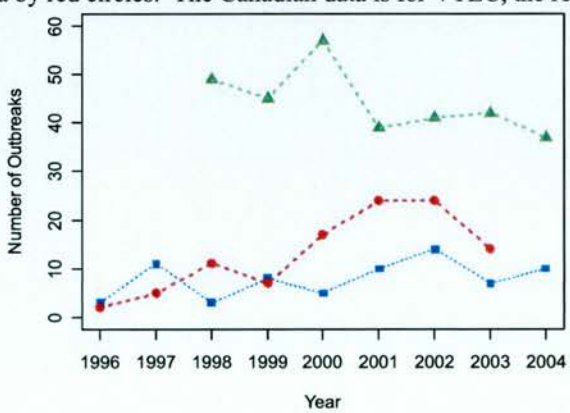


6.3.4.2 Number of total outbreaks

Overall

In contrast, there is a statistically significant difference between the three countries in the trends of the log transformed number of outbreaks ($F_{2,18}$=5.5, p=0.014; Fig. 6.11), with only the trends in Canada and the United States differing statistically significantly (p=0.004). The Canadian trend increases statistically significantly, while the trends in the United States and Scotland do not.

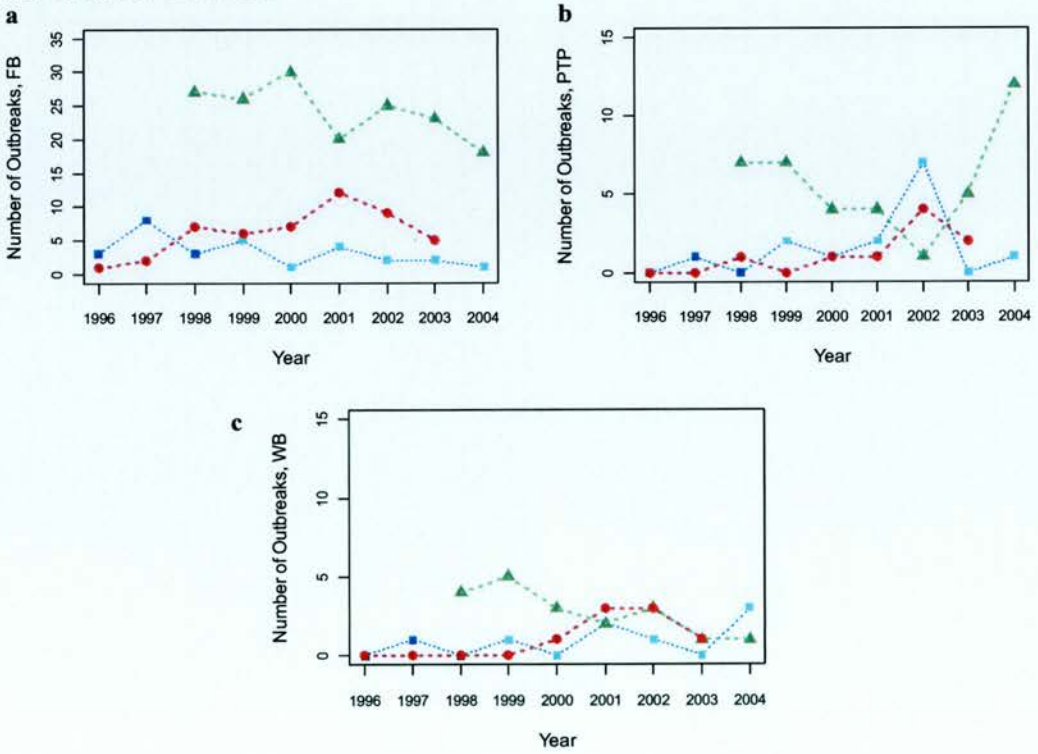
Figure 6.11 a-b: The number of total outbreaks by country
The number of outbreaks by country. Scotland is indicated by blue squares, United States by green triangles and Canada by red circles. The Canadian data is for VTEC, the rest for only *E. coli* O157.



By mode of transmission

If the outbreaks are analysed by mode of transmission, there is a statistically significant difference between countries in the slopes of the trends in the number of foodborne outbreaks ($F_{2,18}$=5.8, p=0.011; Fig. 6.12a), with Canada having a statistically significant different trend slopes from the United States and Scotland (p<0.013). Though none of the slopes are individually statistically significant, the Canadian trend would appear to be increasing, whilst the American and Scottish trends appear to be decreasing. When trends in person to person outbreaks are examined, there is not a statistically significant difference between countries in the trend slopes ($F_{2,18}$=1.2, p=0.337; Fig. 6.12b). The slopes for the trends in the number of waterborne outbreaks do vary statistically significantly between countries ($F_{2,18}$=8.4, p=0.003; Fig. 6.12c). It is not unexpected that both Canada and the United States have statistically significantly different slopes from Scotland (p<0.012) because while the Scottish trend is not statistically significant, the American trend is significantly increasing and the Canadian trend significantly decreasing.

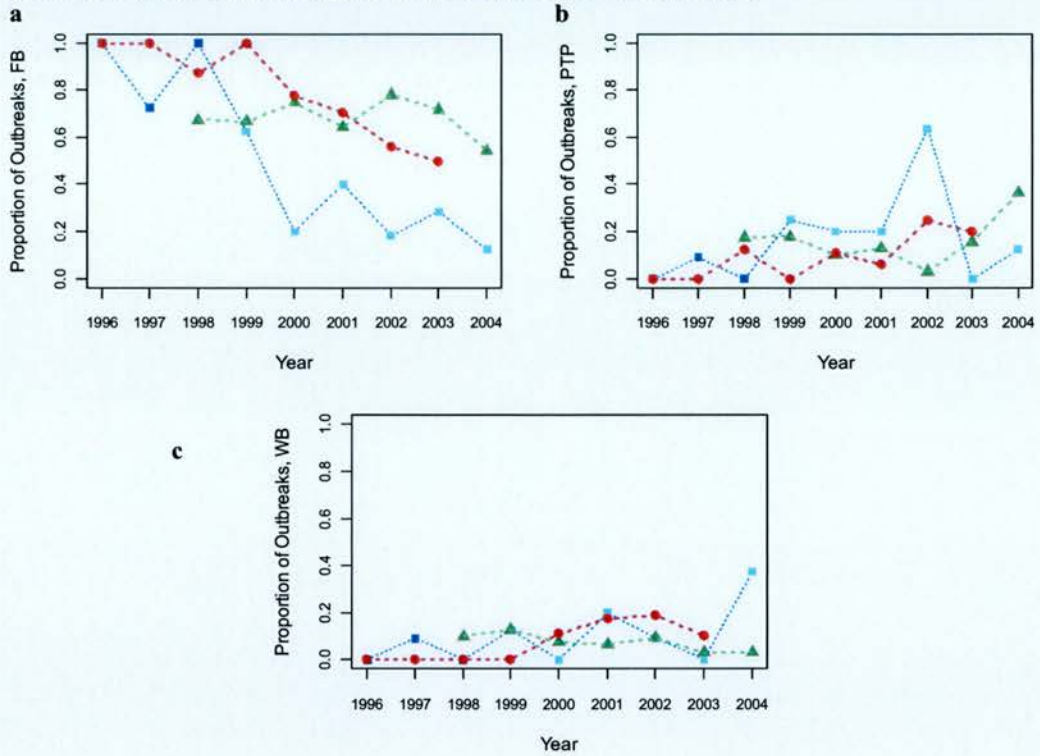
Figure 6.12 a-c: The number of total outbreaks by country and mode of transmission
The number of (a) foodborne, (b) person to person and (c) waterborne outbreaks in Scotland, the United States and Canada. Scotland is indicated by blue squares, United States by green triangles and Canada by red circles.



6.3.4.3 Proportions of outbreaks – by mode of transmission

When outbreaks of a known mode of transmission are considered, there is a statistically significant difference overall between countries in the slopes of the trends of the proportion of outbreaks that are different modes of transmission ($F_{6,72}$=4.1, p=0.001).

Figure 6.13 a-c: The number of total outbreaks by country and mode of transmission
The proportion of outbreaks with a known mode of transmission that are (a) foodborne, (b) person to person and (c) waterborne in Scotland, the United States and Canada. Scotland is indicated by blue squares, United States by green triangles and Canada by red circles.



The slopes of the trends in the proportion of outbreaks that were foodborne and waterborne differ statistically significantly between countries (p<0.006; Fig. 6.13 a,c) and in both instances, the slope of the trend in the United States is statistically significantly different from the slopes in the other two countries (p<0.015). Scotland and Canada both have significantly decreasing trends in the proportion of outbreaks that are foodborne. However the trend in the trend in the United States is not statistically significant. The opposite is true for waterborne outbreaks: there has been a statistically significant decline in the proportion of outbreaks that are waterborne in the United States, with no significant trend in the other two countries. The trends in

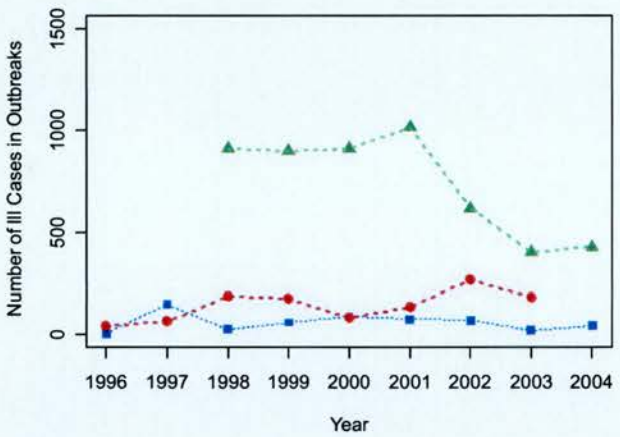
the proportion of outbreaks that were person to person were not compared because no appropriate simple linear model was found for Scotland (Fig. 6.13b).

6.3.4.4 Number of ill cases

Overall

When the cases from the large outbreaks are excluded, the overall difference in the trend slopes is statistically significant ($F_{2,18}$=5.4, p=0.015; Fig. 6.14), with a statistically significant difference between the United States and Canada (p=0.005). As illustrated in Figure 6.14, there is a statistically significant decline the number of ill cases in the United States, while the trends in Canada and Scotland do not change statistically significantly between 1996 and 2003/2004.

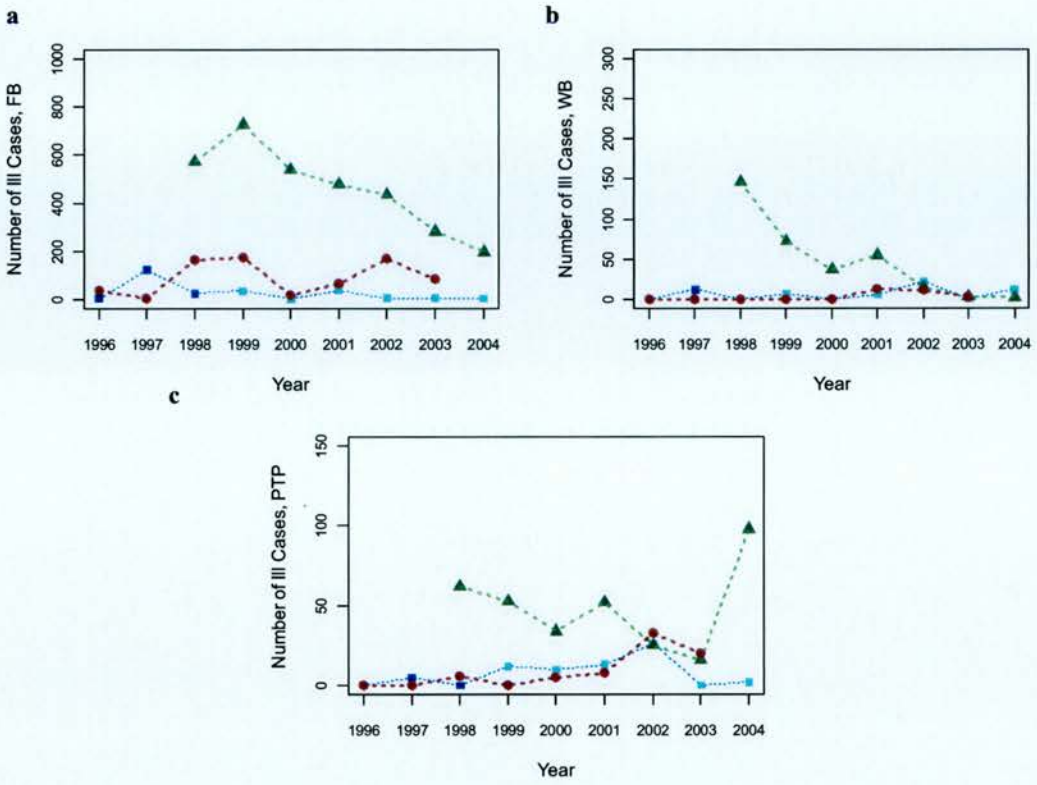
Figure 6.14: The number of ill cases from outbreaks by country
The number of ill cases from outbreaks by country without the cases from the large outbreaks. Scotland is indicated by blue squares, United States by green triangles and Canada by red circles.



By mode of transmission

For the trends in ill cases from foodborne and person to person outbreaks, the slopes do not vary statistically significantly between counties ($F_{2,18}$>2.5, p>0.080; Fig. 6.15a,c). For ill cases from waterborne outbreaks, there is a statistically significant difference between countries in trend slopes ($F_{2,18}$=13.2, p<0.001; Fig. 6.15b). The slope of the United States trend is statistically significantly different from that in the other two countries, as it decreases statistically significantly in contrast to the non-significant (flat) trends in Canada and Scotland.

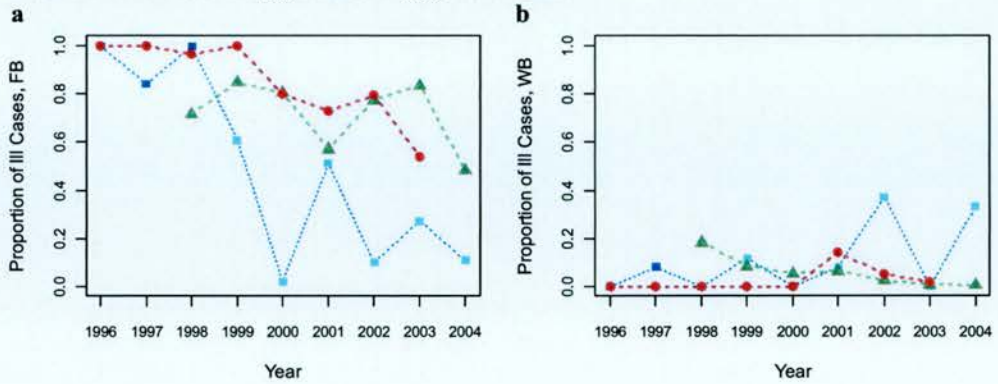
Figure 6.15 a-c: The number of ill cases from outbreaks by country and mode of transmission
The number of ill cases from outbreaks by country where the mode of transmission is (a) foodborne, (b) waterborne and (c) person to person.



6.3.4.5 Proportion of ill cases – by mode of transmission

No statistically significant difference exist in the slopes of the trends in the proportion of outbreak ill cases that are spread via food ($p>0.099$; 6.16a), but there is a statistically significant difference in the slopes of the trends in the proportions of ill cases that were waterborne ($F_{2,18}=12.5$, $p<0.001$; 6.16b). Again, the slope of the American trend, which decreased statistically significantly, was statistically significantly different from the non-significant trends in the other two countries ($p<0.013$). The trends in the proportion of ill cases from person to person outbreaks can not be compared because no adequate simple linear model could be constructed for the Scottish data.

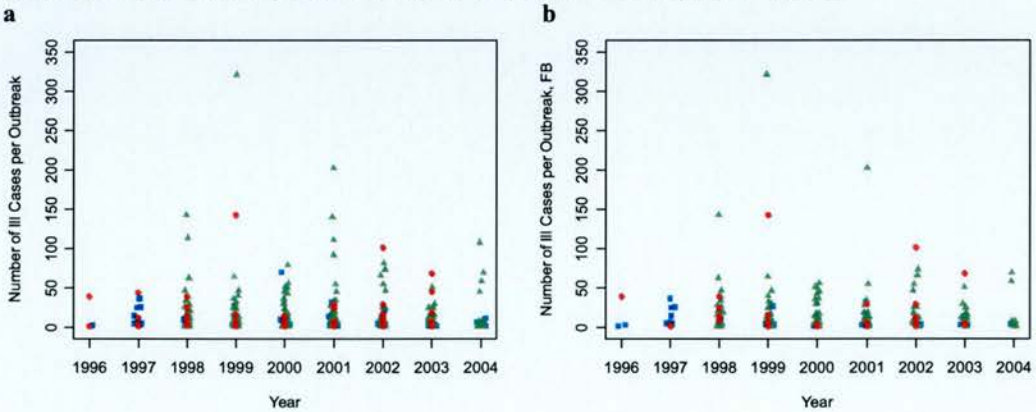
Figure 6.16 a-b: The proportion of ill cases from outbreaks by country and mode of transmission
The proportion of ill cases from outbreaks with a known mode of transmission by country where the mode of transmission is (a) foodborne, (b) waterborne.



6.3.4.6 Number of ill cases per outbreak

There is not a statistically significant difference between countries in the slopes of the trends in the number of ill cases per outbreak (p=0.488; Fig. 6.17a). The same result is seen when the trends in the number of ill cases per foodborne outbreak are examined: there is not an overall statistically significant difference between countries in the slopes (p=0.517; Fig. 6.17b).

Figure 6.17 a-b: The number of ill cases per outbreak by country, overall and foodborne
The number of ill cases per outbreak by country, (a) all cases and (b) foodborne cases. For ease of viewing, outbreaks with more than 350 cases (the large outbreaks) are not shown.



6.4 Discussion

In the previous three chapters of this thesis, the trends in *E. coli* O157 cases and outbreaks in three countries – Scotland, United States and Canada - have been

analysed. Now the trends between countries can be compared, which is the focus of the temporal trend analyses. Though there have been some descriptive analyses of outbreaks in these individual countries (Cowden 1997b; Rangel et al. 2005; Tinga et al. 2006; Woodward et al. 2002), no statistical comparisons of outbreak or case temporal trends between any countries, including Scotland, the United States and Canada have been published. The most comprehensive published comparison on *E. coli* O157 between countries has been a descriptive study of infection epidemiology between 1987 and 1991 in Scotland and Alberta, Canada (Waters et al. 1994), the regions with the then highest rates of *E. coli* O157 infection in the United Kingdom and North America, respectively. The lack of published comparisons in *E. coli* O157 epidemiology between countries may reflect the perceived difficulties in comparing data collected in countries with differing surveillance systems, a concern mentioned by Waters et al. Another factor may be the predominance of single outbreak studies in the published literature on outbreaks: of the 334 papers listed in a search of Medline using the key words "O157 and outbreak", 186 are reports on or research relating to either a single outbreak or a small number of outbreaks. While the establishment of EnterNet, a multinational reporting network, which "allows for trends in infections to be tracked and new and emerging health threats to be recognised in a multi-national setting" (European Commission 2007) (see Section 1.7.1), suggests that there is an increasing focus on examining *E. coli* O157 on an multinational level, the reports have to-date only listed reported numbers of cases and outbreaks without assessing trends statistically or offering any discussion or analysis of similarities between countries.

Comparison of trends between countries, however, has the potential to be a way of finding out more about *E. coli* O157 epidemiology. Evidence from descriptive studies suggests that geographically and demographically similar countries may have very different epidemiological trends regarding *E. coli* O157, and similarities may be found in countries with few such commonalities. For instance, Scotland and England & Wales share an extensive land border across which there is a continual flow of people, animals and goods, but these countries have not had the same rate of *E. coli* O157 infection. In 2004, the rate of infection in Scotland was more than three times that in England & Wales (European Commission 2006). Additionally, the countries

differ in the epidemiology of *E. coli* O157 characteristics such as the predominant phage types, proportions of isolates with both vtx_1 and vtx_2 detected (European Commission 2006; Locking et al. 2006a) and the common type of location for outbreaks between 1995 and 1998 (Cowden 1997b; Willshaw et al. 2001). Other countries which share borders and/or geographical similarities, but have had diverged epidemiologically with regards to *E. coli* O157 include Ireland and Northern Ireland, and Sweden and Denmark (Cheasty et al. 2000). Conversely, Waters and colleagues' report on Alberta, Canada and Scotland in the early 1990s (Waters et al. 1994) indicated that Alberta and Scotland had high rates of infection and similarities in the age distribution amongst infected persons despite Alberta having less than three-fourths the population of and approximately seven times the land mass of Scotland (Scottish Executive 2007a; Statistics Canada 2006).

Since the data above suggests that there is no one geographical or demographic factor which dictates *E. coli* O157 trends, a comparison of trends between countries may, in part, help to elucidate which, if any, factors do influence trends. In this chapter, for the first time, *E. coli* O157 trends over time in Scotland, the United States and Canada have been statistically compared, as it was possible to get outbreak data sets from these countries.

In keeping with the aims of this thesis, which are to analyse trends in *E. coli* O157 outbreaks and cases using simple linear models, the comparison has been conducted, where possible, using the same models and error structures as in the analyses of the individual countries. In order to compare the trends between countries, analysis of covariance was used, with country as the comparison variable. Use of the ANCOVA technique permitted comparison of the trends between countries for which data was available for different time periods, since 2004 data was not available for Canada. Additionally, as discussed in Chapter 4, for the United States, trends could only be analysed from 1998 to 2004. Omission of this data has the potential to reduce the power of the comparative analyses, but due to the reasons detailed in Chapters 4, a reduction in time period was epidemiologically preferable to inclusion of data that was not comparable.

6.4.1 Issues regarding the analyses

There are a number of issues that must be considered when comparing data between countries. These issues will be discussed with reference to the analyses of specific trends. The most prominent and potentially problematic issue is that of the differences between the three countries in the surveillance systems through which the data used in the analyses was reported. Data is collected on both the case (sporadic and outbreak) and outbreak level in each country (Scotland, United States and Canada). Individual cases – both sporadic and outbreak - have been reported on a national level in all three countries since at least 1994 with surveillance definitions clearly delineated and data published at least yearly and available online (see Section 1.7.1). Thus, while is none of the surveillance systems are expected to have detected all *E. coli* O157 cases and all have evolved since their inception. Thus, reporting is established and the potential short-comings and idiosyncrasies of each system are known. In each country, surveillance definitions for individual cases reporting are clearly delineated, with data published at least yearly and available online.

For outbreaks, the situation regarding surveillance is more complicated because of the differences between countries in the longevity of reporting systems, the amount and type of information collected and the reporting hierarchies. Scotland, with a population of approximately 5 million, is relatively small country, while in contrast Canada and the United States are very large countries, having populations of approximately 32 million and 291 million persons respectively. Scotland has a surveillance network that is run by part of the national health care system (Health Protection Scotland), and as a result, outbreak information transmission to HPS is either direct or via one of just 14 Health Board regions. In addition, the current outbreak reporting system has been operational since 1996 (Smith-Palmer & Cowden 2004), during which time outbreak data has been collected with a high degree of consistency in the way data is reported.

Responsibility for reporting of infectious diseases involves a multi-tiered surveillance hierarchy from the local, county and city levels up to the provincial/state and national levels, which can involve differing policies and definitions. In the United States, a national foodborne illness reporting form (52.13) has been available

in a number of print (and online) versions, but it was not clear from the information available for this thesis how non-foodborne outbreaks were reported prior to a specific section for that purpose being added to the form in 2004. A separate form (52.12) exists for waterborne outbreaks. However it appears that there are two parallel systems for reporting waterborne outbreaks because the outbreak listings in the annual waterborne outbreak reports (for example (Centers for Disease Control and Prevention 2006g)) do not always correspond with those listed in the *E. coli* O157 data set provided by the CDC. Finally, in Canada, reporting of outbreaks reporting is primarily carried out at the provincial or lower level; only confirmed outbreak (and sporadic) cases are recorded by the National Laboratory for Enteric Pathogens (National Laboratory for Enteric Pathogens 2006). National outbreak reports for 1994-1995 and 1997-1998 were compiled, however the data set made available for this thesis is the most complete pan-Canadian *E. coli* O157 outbreak data set available for this time period. Given the devolved nature of the Canadian outbreak surveillance system, there are a number of issues regarding this current data set, including "missing data, unknown values and sparse laboratory confirmation", as well as "different reporting standards, data and variables between provinces and territories." (Tinga et al. 2006). The most important factors, in terms of the degree of impact on the data set, are missing data and the lack of common variables. Information regarding the nature of missing data, specifically the years for which each province/territory provided data, and the completeness of the data was provided (see Chapter 5). Though the exact nature of the missing data could not be ascertained, the relative statistical significance of the data omissions to the results of the analyses in this chapter was assessed based on the results of deletion analyses. Of the provinces where outbreak data was missing, only three provinces contributed substantially to the overall number of cases. Analyses were run omitting the data from one or all three of these provinces (see Section 5.5.1). These analyses showed that in some instance, the omission of one or all provinces resulted in a statistically significant change in the temporal trend in certain instances. Such effects were discussed in Chapter 5, and will be discussed in this section where relevant. Lack of common variables across provinces is also a serious issue because the absence of a common reporting form or effective consensus on what information should be

recorded for outbreaks means that data from provinces might not be comparable. This issue is discussed in Chapter 5, with the detailed information from by PHAC on the data variables provided by each province or territory used to make the data relatively comparable.

The issues mentioned above – different time periods, surveillance systems and missing data – were addressed, either by adjusting the data sets or interpreting the results with respect to known issues. The variations across surveillance systems in the variable definitions, in particular, were minimised by establishing definitions, either from the existing surveillance definitions or created to accommodate the information available, for the variables within each country in this thesis. Exact details for each variable, both in terms of definition and the effect of other issues, will be discussed in Section 6.4.2.

6.4.2 Discussion of results

Since the analyses in Chapter 3 – 5 could not be performed when the large outbreak cases (Wishaw, Washington County Fair, Layton Avenue Sizzler and Walkerton) were included, and the data suggest that these outbreaks may be outliers, all comparison analyses involving case numbers are conducted excluding the large outbreaks.

6.4.2.1 Total cases and outbreaks

When the geometric mean numbers of the overall number of outbreak and sporadic cases combined are compared, statistically significant differences exist between countries for the geometric mean number and rates of cases per year, but not in the slopes of the temporal trends. The statistically significant differences in the geometric mean number of cases per year, with the United States having a higher mean than either Scotland or Canada, are not unexpected because the United States population is approximately 58 times that of Scotland and nine times that of Canada. Thus even if the United States or Canada had a lower rate of infection, the raw numbers of cases would still be higher. In addition, whilst the figure for the total number of confirmed cases in Scotland includes only confirmed *E. coli* O157 cases, the figure for Canada also includes non-O157 VTEC cases. The American number includes both confirmed and probable cases: the definition of a probable case extends

to include cases which are "epidemiologically linked to a confirmed or probable case" or where there is an elevated serum titre or identification of a shiga toxin in a clinically compatible case (Centers for Disease Control and Prevention 2006a; Centers for Disease Control and Prevention 2006b). Studies that indicate that less than 5% of reported cases in Canada were of serotypes other than O157 (Sockett et al. 2006; Woodward et al. 2002) suggest that the inclusion of the non-O157 cases is not likely to significantly affect the magnitude of the trends.

In contrast, the reporting of both probable and confirmed cases in the United States may result in a final figure that is a considerable overestimate of the number of confirmed cases. The degree of overestimation may be from one third to 10 times, as suggested by data from outbreak report forms during the study period. These data indicate that in the two outbreaks where specific information was available, there were 65 confirmed cases, but an additional 601 cases that were probable (Proctor 2000a; Proctor 2000b). Also, in the 310 outbreaks between 1998 and 2004 analysed in this thesis, the total number of ill cases was more than twice the total number of confirmed cases. In addition, totals for confirmed cases only were available via the PHLIS system until the end of 2001. While 15,487 confirmed and probable cases were reported between 1998-2001 via the NNDSS, just 11,186 confirmed cases were reported via PHLIS during the same time period (Centers for Disease Control and Prevention 1998; Centers for Disease Control and Prevention 2001b; Centers for Disease Control and Prevention 2002b; Centers for Disease Control and Prevention 2003b), a decrease of more than 25%. Yet, while this overestimate clearly affects the analyses of the mean number of confirmed cases per year, if the magnitude of overestimation is consistent from year to year, the slope of the trend may not be statistically significantly affected. It is intriguing, however, that even though the number of cases in the United States is likely to be over-estimated, the rate of total confirmed cases per year is statistically significantly lower in the United States than in the other two countries. The reason for the lower rate of confirmed cases in the United States is not known, but it may be related to percentage of ill persons that are tested (while still shedding) for the bacteria. The percentage of ill persons that are tested depends upon the practices of physicians and laboratories. For instance, if

only people with bloody diarrhoea are tested, fewer infections will be detected than if all persons with diarrhoea are tested.

A significant difference exists between countries in the geometric mean number and rates of outbreaks per year and in the slopes of the trends in the log transformed number of outbreaks. It is not unexpected that Scotland would have a significantly higher rate of outbreaks because, for the reasons discussed in 5.4.1, outbreak reporting in Canada and the United States is probably much less comprehensive than that in Scotland. In Canada, particularly, outbreak data from up to half the provinces/territories is missing each year (see Chapter 5). The slopes of the American and Canadian trends differ statistically significantly with the trend in Canada decreasing while the American trend does not change statistically significantly. The United States has a significantly higher geometric mean number of outbreaks than either Scotland or Canada.

However, these results should be viewed with caution because the Canadian trend may be statistically significantly affected by missing data. Specifically, data from three provinces which reported more than 100 overall cases per year was missing or incomplete in the first half of the time period being analysed (see 5.2.1.1). If, when the Canadian trend was analysed individually, the data from all three of these provinces were omitted from the analyses, the trend was no longer statistically significant. If the significance was in fact an artefact due to missing data, there may not be a statistically significant difference between countries because the Canadian trend was the only statistically significant trend. The difference in the geometric mean number of cases per year may also have been overestimated, because the numbers of outbreaks – according to the definition used in this thesis – may have been undercounted, particularly in Canada and Scotland. In Scotland, unlike in the other two countries, outbreaks must involve two or more cases out with the same household (Locking et al. 2004), so the number of Scottish outbreaks may be undercounted as per the American and Canadian definitions in which outbreaks are not restricted to being multi-household (Tinga et al. 2006)(Personal communication, Thai An Nguyen, CDC). The data on single-household clusters, was however, not available for this thesis. As previously mentioned, three or more years worth of data

from a number of provinces/territories was missing from the Canadian data set, thus the results presented here may not be applicable to the provinces for which no or only a few years data is available.

6.4.2.2 Analyses by mode of transmission - issues

A more detailed view of the trends is revealed when the outbreaks are broken down by the primary mode of transmission. In order to compare trends between countries by mode of transmission, the categories in the mode of transmission variable must be comparable across the three countries. Standardising the variable definitions requires both adjustments to category definitions and an understanding of the issues involved. Only in Scotland are outbreaks publicly reported using a specific set of categories for modes of transmission: foodborne, waterborne, animal contact, environmental, multiple including foodborne, multiple excluding foodborne and unknown (see for example (Smith-Palmer & Cowden 2004)). Since these categories have been used consistently and little other background information relating to the modes of transmission with which to reassign outbreaks to alternative categories was provided in the Scottish data set, the Scottish categories are used as the basis for defining the categories used in the comparison analyses. Standardised categories have not been used in either the Canadian or American data sets, but the updated 52.13 reporting form introduced in the United States in 2004 (Centers for Disease Control and Prevention 2004a) does provide a clear list of options for mode of transmission. It is also possible that a set of categories does exist in the computerized reporting system which is now used to report outbreaks to the CDC. Neither the updated form nor the computer system was in use during the majority of the time period analysed in this study.

However, for both the United States and Canada, more extensive information was provided in relation to the nature of the transmission, for example listings of food items or comments about suspected means of contamination in a farm setting. This information was then used to split the outbreaks into the categories used in Scotland with a few exceptions described below. Firstly, there are no definitive criteria in Scotland for categorising an outbreak as being transmitted by multiple modes including foodborne or multiple modes excluding foodborne, and additional data on

mode of transmission is not available for all outbreaks in the Canadian and United States data. Thus there is no accurate way of applying these categories to additional data. Several outbreaks in the United States are listed foodborne, with mention of an additional mode of transmission; however without information on how many cases were ascribed to the second mode of transmission, it was not possible to determine whether the foodborne transmission was dominant or co-dominant. Thus for the purposes of the comparisons in this chapter, the foodborne category has been broadened. For Scotland the category includes all outbreaks classified as foodborne or multiple including foodborne, and for Canada and the United States it includes all outbreaks where the main mode of transmission was listed as foodborne. As only three outbreaks in Scotland were categorised as multiple excluding foodborne, and a clear definition which could be applied to non-Scottish outbreaks is not available, these outbreaks were omitted in mode of transmission analyses. All outbreaks which related to animal or environmental contact were categorised as animal/environmental, while person to person, waterborne and unknown mode of transmission outbreaks were indicated as such in all three data sets. However, it is not known whether the definitions for waterborne, unknown and person to person are equivalent in (or even within) the three countries, so these categories may include outbreaks that are spread by multiple modes of transmission. However, there should be less bias in the categorisation of the waterborne and person to person outbreaks since the categorisations were made directly from the notations in the data sets.

Another major issue regarding the comparison of these modes of transmission categories between countries was the variability in data recording. Firstly, the nature of the persons making the determination of mode of transmission for each outbreak could vary widely. Depending on circumstance and country, the person making the choice might be made by a public health consultant or trained epidemiologist involved in the outbreak investigation, a staff member paid to input data into a database or an epidemiologist collating data from a number of disparate sources. Secondly, there was no standard between or within countries regarding the way in which the data was recorded – i.e. whether a checklist of options was provided for mode of transmission, a space for an answer, or some combination of the two methods. Such variations in the forms can strongly influence the choice of mode of

transmission. For instance, the set of categories used to record mode of transmission information in one Canadian province included neither foodborne or waterborne as an option. As a result outbreaks which were spread by food or water might be listed as 'other' or fit into an existing category (e.g. faecal-oral if food was contaminated by someone who didn't wash their hands properly), thus obscuring the true transmission nature of the outbreak. These issues are most applicable to Canada and the United States where outbreaks are apparently often recorded – at least below the national level - on various non-O157 specific forms, and a much greater number of reporting bodies are involved (state/provincial, city, county etc. public health agencies) in surveillance. While the experiences and biases of individual public health consultants can influence how outbreaks are reported in Scotland, their selection is guided by a list of defined modes of transmission. It would be anticipated that above discussed issues will result in a number of inbuilt biases in the comparison data set, and these biases will be discussed as relevant in the next section.

6.4.2.3 Analyses by mode of transmission - outbreaks

For all three modes of transmission which are compared – foodborne, person to person and waterborne, statistically significant differences exist in the mean numbers of outbreaks per year, with the United States having a higher mean number of outbreaks than either Canada or Scotland, and Canada a higher number than Scotland. The rates of outbreaks are not statistically significantly different, but the lack of apparent differences may be a result of the large 95% confidence intervals which resulted from having a low number of outbreaks per year, but high variation (0 to 12 in one instance) between years. The rates may also be different. There is a statistically significant difference between Canada and both the United States and Scotland countries in the slopes of the trends in the number of foodborne outbreaks, the slope in Canada being significantly different from the slope in either Scotland or the United States though none of the slopes were individually statistically significant. It would appear though, that the slope of the Canadian data is decreasing, whilst the American and Scottish slopes are increasing. However, these results should be interpreted with caution because the slope of the Canadian trend may be significantly skewed by the large quantity of missing data in the first few years of the data set. In

addition, particularly in Canada, there may be an overestimation of the number of foodborne outbreaks earlier in the data set. Firstly, since the initial *E. coli* O157 outbreaks in the United States (Riley et al. 1983; Wells et al. 1983) and Canada (Lior 1983) were associated with hamburger meat, and a large number of outbreaks in Canada have been associated with food (Todd 2000), outbreak investigations in the earlier years of the data sets may have placed a greater emphasis on food-related sources of infection (personal communication, Carol Tinga, PHAC). In Canada there were no outbreaks categorised as any mode other than food or unknown until 1998, and then no waterborne or animal/environmental outbreaks until 2000. Additionally, prevention and interventions programs aimed at foodborne transmission such as recalls, improvement in the processing practices in food plants and consumer education campaigns (Sockett et al. 2006) could potentially have contributed to a decrease in foodborne outbreaks in Canada.

There is not a statistically significant difference between countries in the slopes of the trends of the number of person to person outbreaks. The United States does have a statistically significantly higher mean number of person to person outbreaks per year, but a possible explanation for this difference is the much larger population of the United States. Canada has a statistically significantly different slope in the trend in the number of waterborne outbreaks when compared to either the United States or Scotland. However, the statistically significant increase in the number of waterborne outbreaks in Canada is likely to be an artefact of the under-reporting of outbreaks in the first half of the Canadian data set, as no waterborne outbreaks were listed on the data set until 2000.

Surveillance disparities such as missing mode of transmission values are likely to have had an impact on the results discussed above. In Canadian outbreaks, data is likely to have been lost in the process of trying to enter and compile data into/from databases with different mode of transmission categories. In the United States, particularly in the earlier years of the time period, values may not have been recorded accurately because of a reporting system in which *E. coli* O157 outbreaks were reported on forms designed for foodborne outbreaks. Additionally, the pre-computerised system depended on information reported from 50 state public health

departments and countless county and city health departments. However, these results could also reflect a true difference in the dominance over time of foodborne and waterborne outbreaks, suggesting that in the United States there is a shift away from water as a method of spread, while food is becoming less dominant in Scotland and Canada.

After outbreaks with an unknown mode of transmission were excluded, only the mean proportion of outbreaks that were foodborne per year and over the entire study period varied statistically significantly between countries. In both instances, Scotland had a statistically significantly lower mean proportion of foodborne outbreaks per year. When the temporal trends were compared, statistically significant differences existed between countries in the slopes of the trends in the proportions of known outbreaks that were foodborne and waterborne. For foodborne outbreaks, the slope of the American trend is statistically significantly different from that in the other two countries, which would be expected because the trend in the United States was non-significant, whilst there was a statistically significant decline in both Canada and Scotland. It has been suggested that the decline in Scotland may be in part due to an increasing awareness of non-foodborne modes of transmission, as highlighted by Strachan and colleagues paper on recent outbreaks linked to environmental contamination and animal contact (Strachan et al. 2006). Likewise the slope of the trend in the proportion of outbreaks that were waterborne was statistically significantly different in the United States as compared to the other countries, which is not unexpected because the American trend is the only one of the three to be statistically significant. Person to person outbreaks were not compared because no appropriate simple model could be found for the Canadian data

6.4.2.4 – Ill cases from outbreaks - analyses

Though there are statistically significant differences between all three countries in the mean number of ill cases from outbreaks per year, there are no statistically significant differences between mean rates. Additionally, the slopes of the temporal trends vary statistically significantly only between the United States and Canada: the trend in the United States was statistically significantly decreasing, while in Canada there was no statistically significant trend.

When ill cases from outbreaks are analysed by mode of transmission, no statistically significant differences are found between countries in the mean rates per year. As discussed with reference to rates for the number of outbreaks, however, the lack of statistically significant differences is likely due to a lack of power due to small case numbers and large variations in the number of cases from year to year. If foodborne outbreaks are analysed separately, the differences between the mean numbers of ill cases are still statistically significant, but the slopes are not. There is a statistically significant difference in the mean numbers of and slopes of the trends in the number of ill cases from waterborne outbreaks, with the United States having a statistically significantly decreasing slope as opposed to the non-significant slopes in the other two countries and the highest mean number of ill cases. No difference in trend slopes or intercepts exists between countries in the number of ill cases from person to person outbreaks. These last two results should be interpreted with caution due to the fact that the American trend is based only on data from 1998 onwards and there were no waterborne outbreaks reported in Canada until 2000. More complete outbreak data is needed to confirm the findings on trends in waterborne and person to person outbreaks.

When the mean proportions of ill cases were examined, both overall and by year, there were significant overall differences between countries in the proportion of ill cases from foodborne, waterborne and person to person outbreaks. Scotland has a significantly higher overall and yearly mean proportion of ill cases from person to person and waterborne outbreaks, and a lower mean proportion from foodborne cases. This disparity could reflect a true difference in outbreak transmission epidemiology. However, it is more likely that the higher proportion of non-foodborne outbreaks in Scotland is due to an under-reporting of non-foodborne outbreaks in the United States and Canada (personal communication, Carole Tinga, PHAC) and a greater awareness of such outbreaks in Scotland, at least in recent years (Strachan et al. 2006).

6.4.2.5 Number of ill cases per outbreaks - analyses

The average size of an outbreak – in ill cases – was statistically significantly larger in the United States than in either Scotland or Canada, but there were no statistically

significant differences in the slopes of the trends in size of outbreaks. There were also no statistically significant differences in the trend slopes when only foodborne outbreaks were considered. It is possible that outbreaks in the United States were significantly larger because a bias in the size of outbreaks reported to the CDC or that outbreak investigators in the United States use broader case definitions when reporting data to the CDC. The fact that the slopes did not differ significantly between countries suggest however that the large the size of outbreaks was the result of an epidemiological factor unique to the United States.

6.4.2.6 Correlation analyses

The correlation between the number of ill and ill and positive cases per outbreak varies very little between countries regardless of the inclusion of the large outbreaks, as evidenced by a large overlap in the 95% confidence intervals. The number of cases associated (ill) with an outbreak depends on the established definition of an ill case as well as the thoroughness of the outbreak investigation. For instance, if an person who is epidemiologically linked to an outbreak, but not laboratory confirmed, must have bloody diarrhoea to be counted as an ill case, fewer people are likely to be reported as ill cases than if a person only had to have diarrhoea. The number of cases confirmed (ill and positive) depends on factors such as the practices regarding taking of faecal samples and testing of samples (Deneen et al. 1998a), all of which vary within and between countries. As a result, it was anticipated that correlation would be quite different between countries. The similarity of the correlation between ill and ill and positive sizes is a finding of note because it suggests that the relationship between the number of ill and ill and positive cases may be closely linked to underlying microbiological epidemiological factors that are independent of the country in which the outbreak takes place. These factors could include the average dose of bacteria an infected person is likely to receive, the relationship between the dose and both the distribution of ages of infected persons and severity of symptoms (Haas et al. 2000; Strachan et al. 2005). Other factors include the relationship between verotoxin type and symptoms and patient age (Friedrich et al. 2002), and the likelihood of having symptoms severe enough to require medical intervention, faecal sampling and testing for *E. coli* O157 (Hedberg et al. 1997; Imhoff et al. 2004; Michel et al. 2000).

While Canada, the United States and Scotland are dramatically different in terms of demographic qualities which have been suggested to have an effect on infectious disease epidemiology like population, population density, land use and cattle population (see Table 2.1), the analyses in this chapter provide the first statistical evidence that there are relatively few statistically significant differences between the countries in case or outbreak temporal trends from 1996 to 2004.

Intriguingly, the data suggest that there may have been a decrease in the importance of water as a transmission source in the United States, as the United States alone had statistically significant trends – all decreasing – in the number and proportion of ill cases that were waterborne and the proportion of outbreaks that were waterborne. Further clarification about how waterborne outbreaks are reported in the United States could help in confirming or refuting these results as not all waterborne outbreaks included in yearly reports on waterborne outbreaks are also in the *E. coli* O157 outbreak data set. Thus it could be that progress has been made in reducing contamination of water supplies and bathing areas in the United States. However, another possibility is that there appears to be a decrease in waterborne outbreaks because the reports of some waterborne outbreaks are not being included in the *E. coli* O157 outbreak data set. Differences in geometric mean or mean numbers did exist, as expected, given the large variation in population sizes. The vast majority of the differences in trends could, in part, be explained by missing data and inter-country variation in reporting and surveillance. Therefore, the analyses suggest that Scotland, Canada and United States are quite similar in terms of temporal trends in *E. coli* O157 outbreaks and cases. This would suggest that the factors which are influential in causing changes in the trends in number of cases and outbreaks do not vary significantly between countries. It would be of great interested to follow up on these findings in future years as more data become available, both in terms of new data and more complete data from past years.

This chapter concludes the analysis of temporal trends. In Chapter 7, the focus will shift specifically to cases within outbreaks, with data taken from published reports rather than national data sets. Secondary and primary cases in outbreaks detailed in the published literature will be described and the overall proportion of secondary

cases calculated. In addition, analyses will be run to determine which factors are statistically significantly associated with higher rates of secondary cases.

Chapter 7 -- An analysis of primary and secondary cases in *E. coli* O157 outbreaks

7.1 Introduction

E. coli O157 infection can be acquired directly from the initial or point source of the bacteria, whether it be an infected animal, animal faeces, contaminated food or water (Coia 1999). However, secondary transmission from an infected person, directly or indirectly [e.g. through water] to another person is known to also occur (Armstrong et al. 1996; Coia 1998a) with between 10% and 20% of outbreak cases thought to have been acquired by secondary transmission (Armstrong et al. 1996; Locking et al. 2003a; Locking et al. 2004; Parry & Palmer 2005). These rates of secondary infection may reflect both the low infectious dose (Parry & Salmon 1998) and the potential for further spread from contact with asymptomatic carriers (Armstrong et al. 1996). These asymptomatic carriers present a special risk, as they may be unaware of their infective status (Armstrong et al. 1996).

The potential for secondary cases, alone, highlights the benefit of research into secondary spread. However, the potential for severe outcomes in the age-groups which are at the highest risk for infection is also of concern. Transmission between members of a household is a type of secondary spread, and age appears to be an important factor in determining individuals likely to be infected (Parry & Salmon 1998). Young children are the most likely to transmit infection to, and be infected by, other household members (Parry & Salmon 1998). This association between children and infection is reflected in other locations where outbreaks are likely to involve many secondary cases, including nurseries (Al-Jader et al. 1999; Belongia et al. 1993; Galanis et al. 2003; Spika et al. 1986; Sugiyama et al. 2005; Swerdlow & Griffin 1997), petting zoos (Blackmore & Ginzl 2005; Gage et al. 2001; Heuvelink et al. 2002) and swimming areas (Brewster et al. 1994; Friedman et al. 1999; Keene et al. 1994; Paunio et al. 1999). The apparent association of secondary cases with such locations, as well as with institutional settings such as hospitals and care homes, where individuals may have weakened immune systems and problems maintaining hygiene (Bolduc 2004; Coia 1998b; Pavia et al. 1990; Reiss et al. 2006), is of great importance because of the age-related morbidity and mortality (Coia 1998a). Furthermore, people who were infected via secondary transmission appear to have a

similar risk of severe outcomes such as HUS as people who most likely acquired infection by direct exposure (Locking et al. 2006a).

Though there are differences between the countries as far as potentially epidemiologically relevant characteristics like population density and land use are concerned, the fact that secondary case rates of 10 - 20% overall (Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a) and over 50% for individual outbreaks (Galanis et al. 2003; Public Health Laboratory Service 1998a) have been reported in multiple countries, seems to indicate that regardless of country, secondary cases are potentially important in terms of numbers and populations affected. This suggests that an examination of the relationship between primary and secondary cases, in particular any differences between modes of primary and secondary transmission and countries, is of interest. However, while some countries do report population based figures for secondary spread (Locking et al. 2006a) there has been little systematic quantification or characterisation of secondary outbreak cases. Generally descriptive information with no statistical analyses of the data has been presented (Armstrong et al. 1996; Belongia et al. 1993; Coia 1998b). Definitions of primary mode of transmission for surveillance purposes have generally been based upon the predominant transmission mode rather than the mode of transmission to the initial case or cases (Centers for Disease Control and Prevention 2004a; Health Protection Scotland 2004). Also, while nursery school-based outbreaks have been mentioned as a unique group with high rates of secondary infection (Armstrong et al. 1996), there have been otherwise almost no comparisons between primary or secondary modes of transmission or countries. Perhaps the only exception to this is the study of secondary household transmission in sporadic cases in Wales (Parry & Salmon 1998). However, this study was limited to a single country and because all secondary cases were the result of household contact, there was no comparison between modes of transmission. More comprehensive data and analyses on secondary cases, especially relating to any differences or similarities between countries, modes of transmission, and ages of secondary cases, would help greatly to better understand the observed heterogeneity in the size and severity of *E. coli* O157 outbreaks.

Therefore, this chapter will describe and characterise primary and secondary cases in *E. coli* O157 outbreaks and estimate the proportion of *E. coli* O157 outbreak cases that are the result of secondary transmission. It will analyse the relationships between primary and secondary cases in general *E. coli* O157 outbreaks by mode of transmission, country and age of cases.

7.2 Materials and methods

7.2.1 Literature search

A search was performed of online reference databases and the online archives of publications issued by national public health organizations for reports of *E. coli* O157 outbreaks occurring between January 1, 1982 and December 31, 2006 in Scotland, England and Wales, the United States, Canada, Japan, Ireland and Scandinavian nations (Sweden, Denmark and Finland). Medline and the Web of Knowledge were searched using one or more of the terms "O157", "outbreak", "STEC" and "VTEC". National and international organizational publications that were searched were Mortality and Morbidity Weekly (USA), Canadian Communicable Disease Report, (Canada), Communicable Disease Review Weekly and Communicable Disease and Public Health (England & Wales), SCIEH Reports (Scotland), Infectious Agents Surveillance Reports English version (Japan), Epi-Insight (Ireland), EpiNews (Denmark) and Eurosurveillance (European Union). References from papers or reports found in the above searches were also checked for relevant materials. While outbreak data sets from the United States and Canada were obtained for the studies in Chapters 2 -6 of this thesis, these data sets did not contain information on the number of secondary cases. As a result, the data sets could not be used for this chapter.

7.2.2 Definitions and inclusion criteria

7.2.2.1 Time period

This is selected to include outbreaks from 1982, the year in which the first *E. coli* O157 outbreak was reported, to mid 2006, 2005 being the last complete year for which outbreak data is available.

7.2.2.2 Definitions

7.2.2.2a Case definition

For the purposes of this chapter, a case is defined as a person-infection-episode of *E. coli* O157 infection that is serum or culture confirmed.

7.2.2.2b Outbreak definition and inclusion criteria

For all countries, outbreaks are defined as events with two or more epidemiologically related *E. coli* O157 cases. For an outbreak to be included in the data set used in this chapter, enough information had to be provided in the report to determine the total number of confirmed cases and the number of confirmed secondary cases (see 7.2.2.2d for definition).

7.2.2.2c Modes of infection transmission

Based on the information provided in the articles or reports, transmission of an infection to a person in the outbreaks was classified by the following routes:

1. exposure to a contaminated food, subdivided for this chapter into dairy products (*dairy product*) and non-dairy food items (*food*). Dairy products are defined here as milk, cheese, cheese curds, yogurt, cream and ice cream.

2. exposure to water from a contaminated water supply or recreational/outdoor activities water contaminated by a non-human source (*water*)

3. direct animal contact or exposure to infected animals (*animal contact*)

4. exposure to contaminated animal faeces or environments, but where direct animal contact has not been identified (*environmental*)

5. person to person transmission in a specific setting (*person to person...*)
    - *institution* – residential facility in which meals are generally prepared in a common kitchen or kitchens. For example: hospitals, prisons and nursing homes.

    - *nursery* – a nursery or day-care facility where attendance is generally limited to pupils of age five years or younger

- *home* – private home. Also includes any cases where the location of transmission is not specified as taking place in a commercial or public location.

- *other* – includes all other public or commercial locations including restaurants, camps and non-residential schools where contact took place outside residential settings

6. contact with recreational water contaminated by another human (*water*)

7. unknown pathways (*unknown*) – where no suspected mode of transmission was reported or the suspected mode of transmission was not from the above options

Primary and secondary cases

The definitions of secondary and primary cases were not consistent between outbreak reports. For instance in the 1993 minced-beef outbreak in Washington, people were defined as having secondary infection if they "became ill within 10 days of a household or other close contact with another patient and had not eaten at chain A during that time" (Bell et al. 1994). However, in an outbreak that same year in London (Hildebrand et al. 1996) linked to paddling pools, a secondary case was defined as person with "any gastrointestinal symptoms and/or HUS with onset date in 1993 between mid-June and end of July, who lived in south west London and in whom there was microbiological evidence of *E. coli* O157 infection…[and] who, in addition, had had household or close personal contact with a case in the previous two weeks" (Hildebrand et al. 1996). The second definition is more restrictive in terms of geographical region, but unlike the first definition, does not clearly rule out patients who had contact with both the initial source of infection (pool) and another infection person, and thus could have been infected either way. Also, in many papers secondary case definitions were not included (e.g. (Gammie et al. 1996; Hilborn et al. 1999; Marsh et al. 1992)). Thus, standard case definitions for primary and secondary cases have been established for this study.

A *primary case* is defined as a case in which the infected person was determined, (based on the information provided in the article or report) as having direct contact with the suspected initial or point source of *E. coli* O157 through transmission routes 1 – 4,7 as detailed above.

A case is considered to be *secondary* if there was no reported exposure to the suspected initial or point source of contamination, and the infection resulted from confirmed or suspected transmission from a primary case. Secondary transmission of an infection to a person can occur by transmission routes 5, 6 and 7. Where no secondary spread was reported in an outbreak, secondary mode of transmission was categorised as not applicable *(n/a)* and not included in any analyses of secondary mode of transmission. In outbreaks where cases were suspected to have had more than one mode of secondary transmission, mode of secondary transmission category was selected based on the transmission route of the majority of secondary cases.

7.2.3 Statistical analysis

All analyses have been performed using R (Version 2.4.0, R Foundation, 2006), with the exception of the Generalised Linear Model (GLM) analyses, which were performed using SAS® Version 9.1 (SAS Institute Inc. 2003). For the first section of the chapter, data from the outbreaks which met the above inclusion criteria are described in terms of median (due to non-normal distributions) number and range of ill, confirmed, primary and secondary cases, as well as by primary and secondary mode of transmission, country, year of outbreak, and age. The overall mean proportion of outbreak cases that are secondary and the overall rate of secondary cases per primary case have been calculated, as well as the mean proportion of cases secondary for each age category. Mean proportions are calculated out of the total number of confirmed cases with known transmission route (primary or secondary).

In all further analyses, mode of transmission and/or country categories with unknown values or containing fewer than five outbreaks are combined or omitted. For analyses involving mode of primary transmission, the categories 'animal contact' (n=7) and 'environmental' (n=2) are combined. Similarly for analyses involving mode of secondary transmission, the levels 'n/a' (n=21) and 'unknown' (n=2) are omitted and 'p-p institution' (n=4) and 'p-p other' (n=3) are combined. For analyses involving

country, categories Sweden (n=2), Ireland (n=1) and Finland (n=1) are omitted, and levels 'Wales' (n=1) and 'England' (n=23) are combined.

A categorical variable with four levels has been created for all analyses involving median age. The levels were: less than 6 years old (nursery-aged infants and toddlers), between 6 and 16 years old (school-aged children), greater than 16 years old to 59 years old (adults) and 60 or more years old (senior citizens). These median age groups are used to split the median ages into those including nursery-aged infants and toddlers, school-aged children, adults, and senior citizens.

Analysis of Variance (ANOVA) analyses are used to check for statistically significant differences in the log-transformed number of ill, confirmed and primary cases between modes of transmission, countries and age categories. When the ANOVA indicates a statistically significant overall difference, as discussed in Chapter 6 (see Section 6.2.3.1), contrasts are used to examine the differences between the individual variable categories. Log-transformation is applied to normalize the residuals. Fisher's Exact Tests are used to check for statistically significant differences in age categories between modes of transmission and countries. As there were many zero values in the secondary case variable, log transformation is not possible and so non-parametric analyses are used to compare numbers of secondary cases between modes of transmission and countries. Kruskal-Wallis Tests are performed to check for overall statistically significant differences, and Mann-Whitney Tests are used to test for differences between pairs of variable levels. The ANOVA and Kruskal-Wallis Tests are repeated excluding nursery outbreaks (see results). In all cases, values of $p < 0.05$ are considered to be statistically significant and subscripts are used to indicate degrees of freedom for F tests.

The relationships between the rate of secondary cases as compared to primary cases by mode of transmission, country and median age are explored using univariate GLMs with Poisson errors and the log link function.

In the models, mode of transmission, country or median age (categorical variable) is inserted as the explanatory variable, the number of secondary cases as the response variable and the number of primary cases, log transformed, as the offset variable.

This is because the expected number of secondary cases from the Poisson model will depend not just on the variables included in the model, but also directly on the number of primary cases in an outbreak. The offset variable (Crawley 2002) is included so that the chance of a given number of secondary cases occurring will reflect this relationship between the number of secondary and primary cases. The use of the offset variable thereby allows for the size of the population who could spread the infection (i.e. primary cases) in the analysis of the rate of secondary cases per primary case. The error distributions are not strictly Poisson and there is overdispersion, which is when an excess of unexplained variation (residual deviance) occurs, as compared to the number of data points or levels (degrees of freedom). Thus statistical significance is overestimated, and so to adjust for this, a scale parameter, estimated by the square root of deviance/degrees of freedom, is factored into the ANOVA calculations. This scale parameter in SAS® has a similar function as the quasi-function in R, used in early chapters (see2.3.4.2). F-values are used to assess the overall statistical significance between modes of transmission, countries and age categories (Crawley 2002), and where statistical significance is indicated ($p<0.05$), the statistically significant differences between different modes of transmission, countries and age categories are quantified by examining the differences of least square means.

Multivariate analyses including the above explanatory variables are run using the same basic GLM model as above to look for potential confounders and interactions between explanatory variables. F-values are used to assess overall statistical significance of explanatory variables. Model simplification is achieved by omitting model terms which are not statistically significant in three way interactions and running model with all three variables without interaction. Models are then run with two variables, and if there is no statistically significant interaction between the terms, the two variables are retained in the model, but without an interaction term.

7.3 Results

7.3.1 Descriptive summary – overall, country and age

There were 176 outbreaks, covering the time period 1982 to 2006, for which outbreak reports were available (Table 7.1). Of these outbreaks, 90 matched the

inclusion criteria for the study and were included in the descriptive portion of the study.

Table 7.1: Selection of outbreaks for the study
Table of the outbreaks included in the stages of the studies, and reasons for exclusion. The other rows show the data sets used in the univariate and multivariate analyses, when various omissions were made, including omission of nursery cases and omission of cases without a value for categorical median age. Outbreaks for which there were no secondary cases are classified as 'n/a' for the purposes of secondary transmission and person to person is abbreviated as 'p-p'.

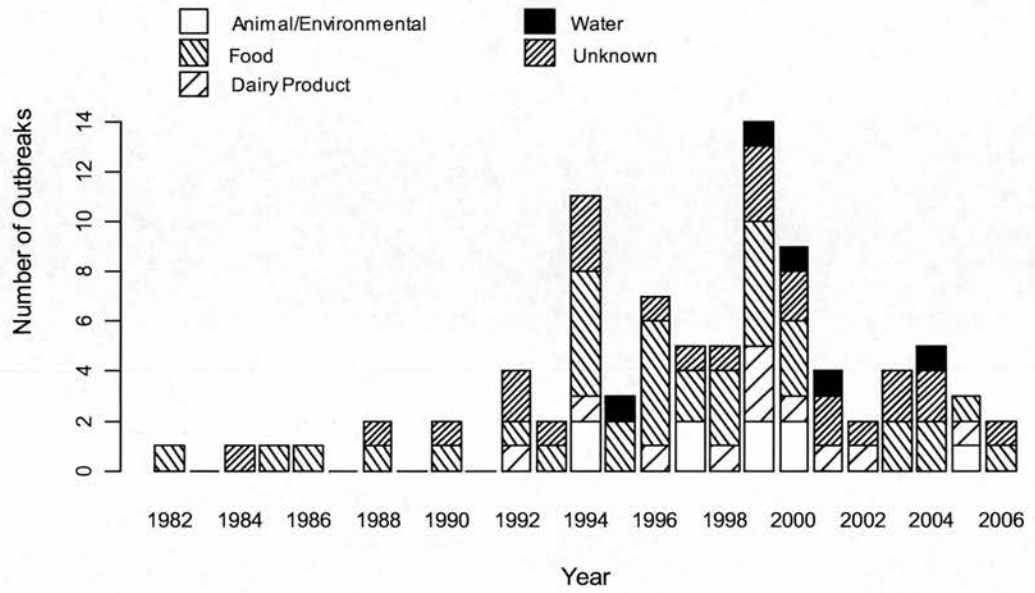
Selection of Outbreaks for the Studies	
Outbreaks + descriptive report(s) were found in the published literature	N = 176
- those outbreaks where the number confirmed could not be ascertained - those outbreaks where the number of secondary confirmed cases could not be ascertained	- 6 - 80
Outbreaks * met the case definition (sufficient information to determine the number of confirmed primary and secondary cases)	N = 90 (51%)
Outbreaks * after omission or combination of categories with < 5 outbreaks Combined: England & Wales, Omitted: Finland, Sweden, Ireland Combined: Animal & Environmental Omitted: Secondary Mode n/a, unknown, combined: p-p institution & other	 N = 86 N = 90 N = 67
* after omission of outbreaks without stated median age With all modes of transmission and countries With Sweden, Finland and Ireland omitted With Animal & Environmental combined With Secondary Mode n/a and unknown outbreaks omitted	- 15 N = 75 N = 71 N = 75 N = 57

The 90 outbreaks in the descriptive study occurred between 1982 and 2006, with the majority of outbreaks occurring between 1994 and 2000 (Figure 7.1, Table 7.2).

Table 7.2: References for outbreaks used in study (n=90), by country
References for the 90 outbreaks used in the study. Some references involve more than one outbreak and for some outbreaks, there are more than one reference

Canada	(Abbas et al. 2005; Bruneau et al. 2004; CDC 1983; Clark et al. 1997; Galanis et al. 2003; Health Canada 1999; Honish et al. 2005; Honish et al. 2007; Iebin et al. 2003; Lior 1983; MacDonald et al. 2000; MacDonald et al. 2004; McIntyre et al. 2002; Sutcliffe et al. 2004; Warshawsky et al. 2002)
England	(Allaby & Mayon-White 1995; Clark et al. 1997; Crampin et al. 1999; Gammie et al. 1996; Goh et al. 2002; Harrison & Kinra 2004; Health Protection Agency 1999; Health Protection Agency 2006a; Hildebrand et al. 1996; McDonnell et al. 1997; Morgan et al. 1988; Public Health Laboratory Service 1995; Public Health Laboratory Service 1997; Public Health Laboratory Service 1998a; Public Health Laboratory Service 1998b; Public Health Laboratory Service 1999a; Public Health Laboratory Service 1999b; Public Health Laboratory Service 1999c; Public Health Laboratory Service 1999d; Public Health Laboratory Service 2000a; Public Health Laboratory Service 2000b; Public Health Laboratory Service 2000c; Public Health Laboratory Service 2002; Public Health Laboratory Service 2006; Shukla et al. 1995; Verma et al. 2007)
Finland	(Paunio et al. 1999)
Ireland	(O'Donnell et al. 2002)
Japan	(Maruzumi et al. 2005; Michino et al. 1999; Michino et al. 1998; Sugiyama et al. 2005; Terajima et al. 1999; Yamamoto et al. 2001)
Scotland	(Brewster et al. 1994; Coia et al. 1996; Cowden et al. 2001; Davis & Brogan 1995; Howie et al. 2003; Jones & Roworth 1996; Kohli et al. 1994; Licence et al. 2001; Marsh et al. 1992; O'Brien et al. 2001c; SCIEH 1998a; SCIEH 1999; SCIEH 2001a; Upton & Coia 1994)
Sweden	(Wahl & Andersson 2004; Welinder-Olsson et al. 2004)
United States	(Ackman et al. 1997; Banatvala et al. 1996; Bell et al. 1994; Belongia et al. 1991; Belongia et al. 1993; Bhat et al. 2007; Breuer et al. 2001; Bruce et al. 2003; Centers for Disease Control and Prevention 1994; Centers for Disease Control and Prevention 1995a; Centers for Disease Control and Prevention 2005c; Cody et al. 1999; Crump et al. 2003; Feldman et al. 2002; Ferguson et al. 2005; Friedman et al. 1999; Gage et al. 2001; Gouveia et al. 1998; Hilborn et al. 1999; Keene et al. 1994; Keene et al. 1997; Ostroff et al. 1990; Palumbo et al. 2004; Payne et al. 2003; Proctor 2000a; Proctor 2000b; Reiss et al. 2006; Samadpour et al. 2002a; San Mateo County Health Services Agency 2004; Spika et al. 1986; Tuttle et al. 1999)
Wales	(Payne et al. 2003; Public Health Laboratory Service 1999e)

Figure 7.1: Number of outbreaks in the study, by year and mode of primary transmission
The outbreaks included in the descriptive study (n=90) split by year. For outbreaks where cases were reported in more than one year, the year noted in the outbreak report was considered to be the year of record. The bars are divided by mode of primary transmission.



These outbreaks took place in nine countries, with 30% of the outbreaks occurring in the United States (Table 7.3). One outbreak involved cases both in the United States and Canada, and since 86% of the cases took place in the United States was categorized as being in the United States for the purposes of the study.

Table 7.3: Outbreaks included in the study, by country
The outbreaks included in the descriptive study (n=90), split by country. Percentages shown are out of the total outbreaks in the descriptive study.

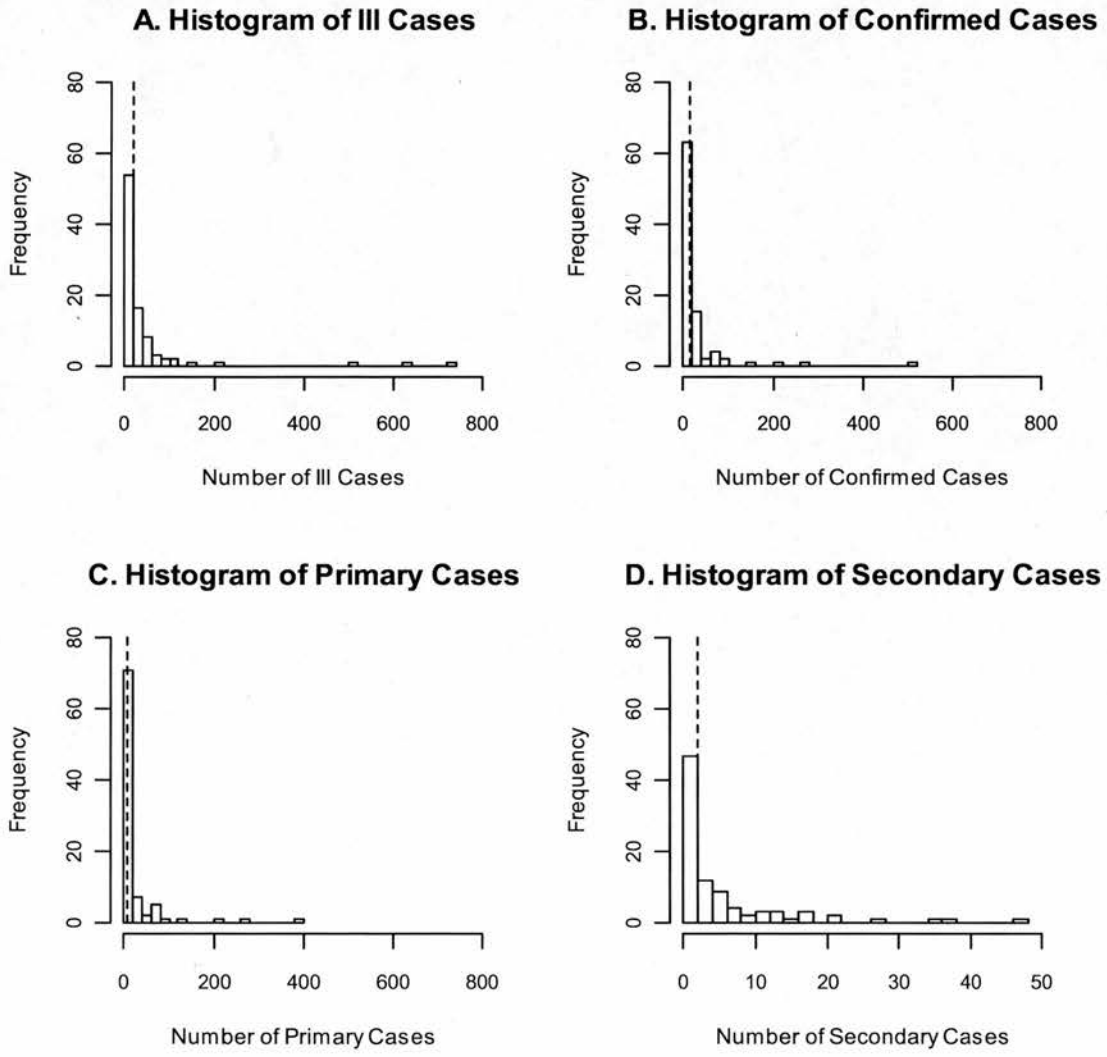
	Outbreaks (%)
United States	27 (30.0)
England	23 (25.6)
Scotland	16 (17.8)
Canada	13 (14.4)
Japan	6 (6.7)
Sweden	2 (2.1)
Ireland	1 (1.1)
Finland	1 (1.1)
Wales	1 (1.1)

In the 75 outbreaks with values for the categorical age variable, 20 (26%) outbreaks had cases with a median age less than 6 years, 27 (36%) a median age from 6 to 16 years, 23 (31%) a median age from 17 to 59 years and 5 (7%) a median age of 60 years or older.

The mean proportion of outbreak cases that were secondary was 0.195 (range 0 to 0.97) and the overall rate of secondary cases per primary case was 0.24. The highest mean proportion of secondary cases were found in outbreaks with a median age of cases that was less than six years (0.65: 95% CI=0.60-0.70), followed by outbreaks with a median age of sixty or more years (0.17: 0.10-0.24), outbreaks with a median age between 6 and 16 (0.12: 0.10-0.14 and those with a median age between 17 and 59 (0.09: 0.07-0.11).

Figure 7.2: Histogram of the number of ill, confirmed, primary and secondary cases per outbreak

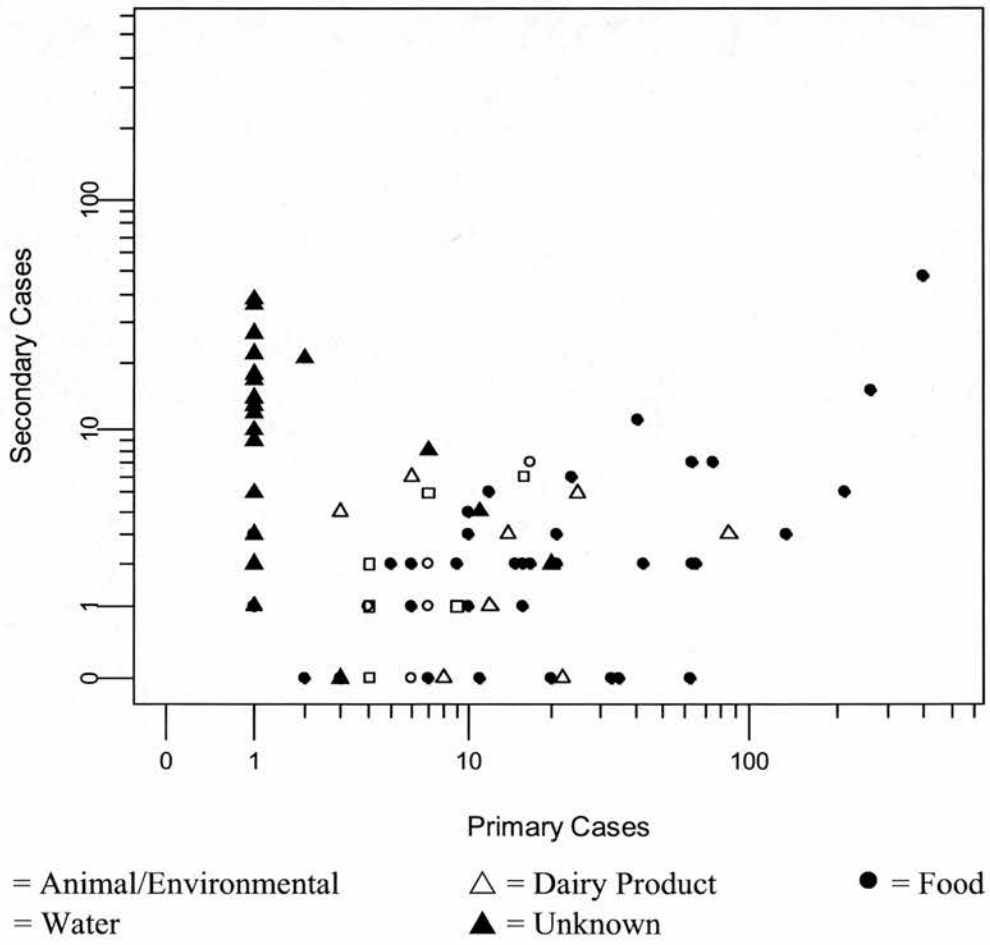
The histograms of the number of (a)ill, (b) confirmed, (c) primary and (d) secondary cases per outbreak in the descriptive study (n=90). The vertical dotted line indicates the median.



The frequency distributions of ill, confirmed, primary and secondary cases are all strongly positively skewed (Figure 7.2 a – d). The median outbreak size for the study is 17 ill (range 2 to 738) cases and 13.5 confirmed (range 2 to 501) cases. The outbreaks have a median of 7 primary cases (range 1 to 398) and a median of two secondary cases (range 0 to 48).

When the numbers of primary cases in each outbreak were plotted against the number of secondary cases (Figure 7.3), there was no linear relationship. The resulting pattern suggests that the majority of outbreaks (81%) have fewer than 50 primary cases and 15 secondary cases, and that the largest numbers of secondary cases (generally more than 10) are seen in outbreaks with 2 or fewer primary cases or those with more than 100 primary cases.

Figure 7.3: The numbers of secondary cases plotted against the number of primary cases
The numbers of secondary cases plotted against the number of primary cases on a log (1+x) scale, n=88. The outbreaks are classified by mode of primary transmission.



After variable categories were combined or omitted, as described above, to avoid categories with fewer than five outbreaks 86 outbreaks remained in the data set used for statistical analyses involving country and 75 for median age (Table 7.1). Between countries, there is a statistically significant difference in the log-transformed mean numbers of ill cases ($F_{4,81}$=2.8, p=0.033; Table 7.4a).

Table 7.4 a-b: Geometric mean number of ill, confirmed and primary cases and median number of secondary cases.
The geometric mean number of ill, confirmed and primary cases and median number of secondary cases by country and median age category are shown. For the geometric means, 95% confidence intervals are shown and for the medians, range is shown. The last row gives the F or χ^2 and p-values for the overall difference in the geometric means or medians.

a) by Country (n=86)

	Ill Cases	Confirmed Cases	Primary Cases	Secondary Cases
United States	25.2 (14.7, 43.1)	14.5 (8.6, 24.6)	6.2 (3.2, 12.2)	2 (2,6)
England & Wales	11.9 (7.9, 17.9)	10.8 (7.5, 15.6)	4.7 (2.8, 7.8)	3 (1,7)
Scotland	14.5 (7.5, 31.7)	14.4 (7.5, 27.6)	9.3 (3.9, 21.8)	2 (0,5)
Canada	13.4 (7.2, 24.9)	9.4 (4.8, 18.7)	5.5 (2.3, 13.4)	2 (1,5)
Japan	52.9 (17.2, 163.3)	43.3 (14.8, 126.4)	23.8 (3.4, 164.2)	3 (0,22)
	$F_{4,81}$=2.8, p=0.033	$F_{4,81}$=2.1, p=0.082	$F_{4,81}$=1.6, p=0.183	χ^2=1.3, p=0.859

b) by Median Age Category (n=75)

	Ill Cases	Confirmed Cases	Primary Cases	Secondary Cases
< 6 years	13.9 (9.0, 21.5)	10.4 (6.5, 16.7)	2.0 (1.2, 3.4)	6 (3,13)
6 -16 years	18.9 (10.2, 35.0)	14.5 (8.5, 25.0)	8.3 (4.2, 16.4)	2 (2,4)
17 - 59 years	20.2 (11.4, 35.7)	13.2 (7.5, 23.1)	10.2 (5.5, 19.0)	2 (0,2)
60+ years	29.5 (9.9, 88.4)	20.7 (11.1, 38.5)	10.9 (1.8, 65.1)	4 (0,10)
	$F_{3,71}$=0.6, p=0.639	$F_{3,71}$=0.5, p=0.665	$F_{3,71}$=5.4, p=0.002	χ^2=8.4, p=0.038

Japanese outbreaks have statistically significantly higher log-transformed mean numbers of ill cases than outbreaks in Canada, England & Wales and Scotland (t<-2.16, p<0.034; Table 7.4a) and outbreaks in the United States have statistically significantly higher log-transformed mean numbers of ill cases than those in England & Wales (t=2.25, p=0.027; Table 7.4a).

In addition, the mean number of primary cases and median number of secondary cases differ statistically significantly between age categories ($F_{3,71}$=5.4, p=0.002;

χ^2=8.4, p=0.038; Table 7.4b). Outbreaks with a median age of less than 6 years have a statistically significantly lower mean number of log transformed primary cases than outbreaks with a median age between 6 -16, 17-59 or 60+ (t<-2.2, p<0.025; Table 7.4b) and a statistically significantly higher median number of secondary cases than outbreaks with a median age of 17-59 (p=0.007; Table 7.4c). There is no statistically significant difference between countries in the proportion of total outbreaks in each median age category (Fisher's Exact p=0.737; Table 7.5).

Table 7.5: Number of outbreaks, by country and age (n=71)
The number of outbreaks by country (excluding category = other) and age.

	Canada	England & Wales	Japan	Scotland	United States
< 6 years	4	6	1	2	6
6 – 16 years	4	5	1	6	10
17 – 59 years	4	6	0	3	8
60+ years	1	0	1	2	1

7.3.2 Descriptive summary – mode of primary transmission

Food was the most often reported mode of primary transmission (as per the definitions in 7.2.2.2) (42%), with dairy products (12%) and animal contact (7%) the next most common modes of primary transmission (Table 7.6).

Table 7.6: Outbreaks included in the study, by mode of primary transmission
The outbreaks included in the descriptive study (n=90), split by country. Percentages shown are out of the total outbreaks in the descriptive study.

	Outbreaks (%)
Food	38 (42.2)
Dairy Product	11 (12.2)
Animal Contact	7 (7.8)
Water	6 (6.7)
Environmental	2 (2.2)
Unknown	26 (28.9)

Other identified primary modes of transmission were water (6%) and environmental contact (2%). In approximately 29% of outbreaks, the primary mode of transmission could not be determined from the information provided in the paper, and thus was classified as unknown.

The log-transformed mean numbers of ill ($F_{4,85}$=6.3, p<0.0001), confirmed ($F_{4,85}$=3.1, p=0.021) and primary cases ($F_{4,85}$=19.2, p<0.0001) and the median number of secondary cases (K-W χ^2=28.2, p<0.0001) differ statistically significantly between modes of primary transmission (Table 7.7).

Table 7.7: Geometric mean number of ill, confirmed and primary cases and median number of secondary cases.
The geometric mean number of ill, confirmed and primary cases and median number of secondary cases by mode of primary transmission are shown. For the geometric means, 95% confidence intervals are shown and for the medians, range is shown (n=90). The last row gives the F or χ^2 and p-values for the overall difference in the geometric means or medians.

	Ill Cases	Confirmed Cases	Primary Cases	Secondary Cases
Animal/Environmental	6.7 (3.6, 12.7)	6.6 (3.5, 12.5)	5.1 (2.4, 10.7)	1 (0,3)
Dairy products	12.0 (5.6, 25.7)	10.4 (4.9, 22.1)	8.1 (3.5, 18.8)	1 (0,5)
Food	33.8 (21.6, 52.8)	20.9 (13.4, 32.4)	18.1(11.5, 28.4)	2 (1,2)
Water	8.5 (4.5, 16.1)	8.3 (4.2, 16.1)	6.3 (3.5, 11.5)	1.5 (1,6)
Unknown	14.2 (10.5, 19.1)	11.3 (8.2, 15.5)	1.4 (1.0, 2.0)	11 (5,17)
	$F_{4,85}$=6.3, p<0.001	$F_{4,85}$=3.1, p<0.021	$F_{4,85}$=19.2, p<0.001	χ^2=28.2, p<0.001

Outbreaks where the primary mode of transmission was food have statistically significantly higher log-transformed mean numbers of ill and primary cases than those outbreaks with any other known or unknown primary mode (t>2.03, p<0.045; Table 7.7), and statistically significantly higher log-transformed mean numbers of confirmed cases than outbreaks with animal/environmental or unknown modes (t> 2.20, p<0.031; Table 7.7). Additionally, outbreaks where the primary mode is unknown have statistically significantly lower log-transformed mean numbers of primary cases (t<-1.28, p<0.005; Table 7.7) and statistically significantly higher median numbers of secondary cases (p<0.01) than those spread by all other (known) primary modes.

There are also statistically significant differences in age categories between modes of primary transmission when the outbreaks with an unknown mode of transmission are included (Fisher's Exact p=0.004). Outbreaks where food is the mode of primary transmission are statistically significantly different in terms of age categories from outbreaks where the primary mode of transmission is water or unknown (p=0.041, 0.001; Table 7.8). There is statistically significant difference in age groups between

waterborne outbreaks and those with an unknown mode of primary transmission (p=0.001, Table 7.8).

Table 7.8: Number of outbreaks, by mode of primary transmission and median age group
The number of outbreaks by mode of primary transmission and median age group (n=75).

	Animal/ Environmental	Dairy Product	Food	Water	Unknown
< 6 years	1	3	2	2	12
6 – 16 years	6	2	11	3	5
17 – 59 years	1	3	15	0	4
60+ years	0	0	3	0	2

7.3.3 Descriptive summary –mode of secondary transmission

For descriptive purposes, all 90 outbreaks were considered when considering mode of secondary transmission. The most common mode of secondary transmission is person to person within a home (46%), with transmission also commonly occurring person to person in nurseries (11%) and via water (10%) (Table 7.9). Approximately one quarter of outbreaks do not have secondary cases (23%).

Table 7.9: Outbreaks included in the study, by mode of secondary transmission (n=90)
The outbreaks included in the descriptive study, split by mode of secondary transmission. Percentages shown are out of the total outbreaks.

	Outbreaks (%)
Person to Person - home	41 (45.6)
Person to Person - nursery	10 (11.1)
Water	9 (10.0)
Person to Person - Institution	4 (4.5)
Person to Person - Other	3 (3.3)
Unknown	2 (2.2)
No secondary cases	21 (23.3)

For the statistical analyses of mode of transmission, only the 67 outbreaks with known mode of secondary transmission were included in the data set. The log-transformed mean numbers of ill and confirmed cases do not vary statistically significantly between modes of secondary transmission ($F_{3,63}$=1.1, p=0.353 and $F_{3,63}$=1.3, p=0.285; Table 7.10).

Table 7.10: Geometric mean number of ill, confirmed and primary cases and median number of secondary cases

The geometric mean number of ill, confirmed and primary cases and median number of secondary cases by mode of secondary transmission are shown. For the geometric means, 95% confidence intervals are shown and for the medians, range is shown (n=67). The last row gives the F or χ^2 and p-values for the overall difference in the geometric means or medians.

	Ill Cases	Confirmed Cases	Primary Cases	Secondary Cases
P-P home	23.8 (15.7, 36.2)	19.8 (13.4, 29.1)	11.9 (7.2, 19.6)	2 (2,5)
P-P nursery	25.8 (18.0, 37.1)	15.7 (10.2, 24.3)	1.4 (0.8, 2.6)	15.5 (9,22)
P-P other	15.5 (5.7, 41.9)	11.2 (5.3, 23.9)	6.0 (1.9, 19.0)	3 (1,10)
Water	11.9 (5.6, 25.5)	9.9 (4.3, 23.0)	1.6 (0.5, 4.9)	5 (3,17)
	$F_{3,63}$=1.1, p=0.353	$F_{3,63}$=1.3, p=0.285	$F_{3,63}$=8.9, p<0.001	χ^2=13.9, p<0.003

However there are statistically significant differences in the log-transformed mean numbers of primary ($F_{3,63}$=8.9, p<0.001) and the median numbers of secondary (p=0.003) cases. Outbreaks where the mode of secondary transmission was person to person spread in a nursery have statistically significantly lower log-transformed mean numbers of primary cases (t<-2.00, p<0.046), and higher median numbers of secondary cases (p<0.016) than all other outbreaks with person to person secondary modes of transmission. In addition, outbreaks with waterborne secondary transmission have statistically significantly lower log-transformed mean numbers of primary cases (t=-3.75, p<0.001) and higher median numbers of secondary cases (p=0.043) than those where secondary transmission was person to person in a home.

There are also statistically significant differences in age categories between modes of secondary transmission whether or not the outbreaks with an unknown mode of transmission are included (Fisher's Exact p<0.001; Table 7.11).

Table 7.11: Number of outbreaks, by mode of secondary transmission and median age group

The number of outbreaks by mode of secondary transmission and median age group (n=75). PTP = person to person

	PTP home	PTP nursery	PTP other	Water	Unknown
< 6 years	7	8	1	2	0
6 – 16 years	17	1	1	4	0
17 – 59 years	10	0	1	2	2
60+ years	0	0	3	0	0

Outbreaks in which the secondary mode of transmission was person to person in a nursery had statistically significant differences in median age categories as compared to outbreaks spread by all other secondary modes of transmission ($p<0.030$). Additionally, outbreaks in which the secondary mode of transmission was person to person in a home had statistically significantly different counts in median age categories than outbreaks where the secondary mode was person to person other ($p=0.003$).

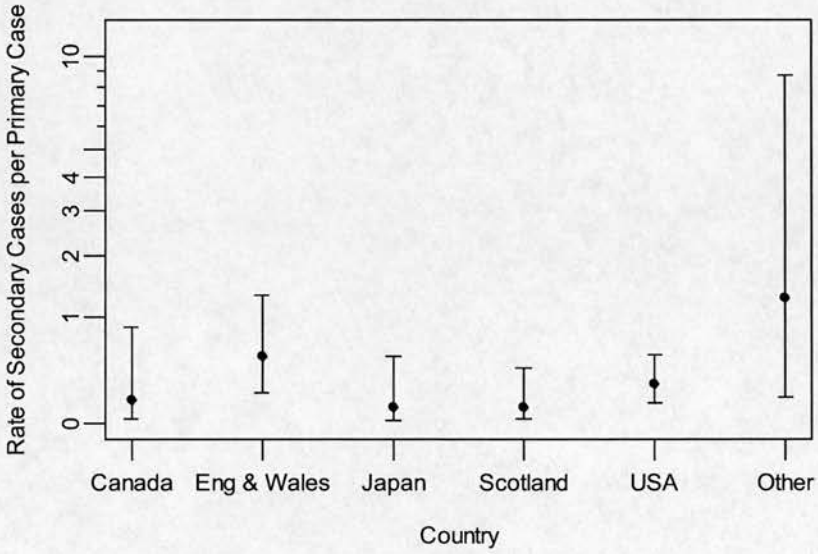
7.3.4 Rate of secondary cases in relation to primary cases.

7.3.4.1 Country and age –univariate models

Country

There is no statistically significant overall difference between countries in the rate of secondary cases to primary cases ($F_{4,81}=1.4$, $p=0.232$; Figure 7.4).

Figure 7.4: Rate of secondary outbreak cases per primary case, by country
The rate of secondary outbreak cases per primary case, by country, shown on the log (1+x) scale. The means and 95% confidence intervals – indicated by the bars - are calculated from the univariate GLM model using back transformation of the parameter and standard errors. For Canada n=13, England & Wales n=24, Japan n=6, Scotland n=16 and United States, n=27.

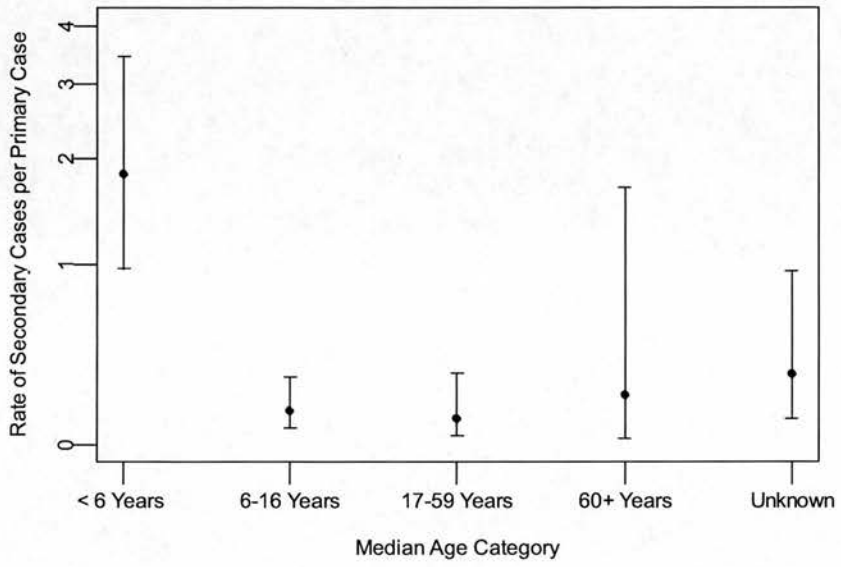


Age

A statistically significant overall difference in the rate of secondary cases to primary cases exists between age categories ($F_{3,71}=8.9$, $p<0.001$), with outbreaks of a median case age of less than six years having a significantly higher rate of secondary cases to

primary cases than outbreaks with a median case age of 6-16 years (χ^2=25.2, p<0.001) and 17-59 years (χ^2=18.8, p<0.001; Figure 7.5).

Figure 7.5: Rate of secondary outbreak cases per primary case, by median age category
The rate of secondary outbreak cases per primary case, by median age category, shown on the log (1+x) scale. The means and 95% confidence intervals – indicated by the bars - are calculated from the univariate GLM model using back transformation of the parameter and standard errors. For <6 years n= 20, 6 -16 years n=27, 17-59 years n=23, 60+ years n=5 and unknown, n=15.



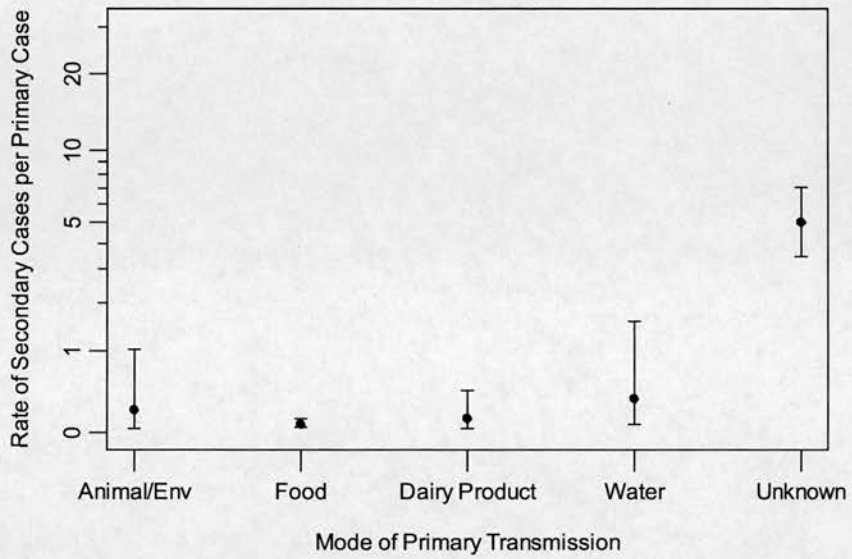
7.3.4.2 Mode of primary transmission –univariate models

Mode of primary transmission

There is a statistically significant difference between modes of primary transmission in the rate of secondary cases to primary cases ($F_{4,85}$= 37.9, p<0.001; Figure 7.6). Outbreaks in which the mode of primary transmission is unknown have statistically significantly higher rates of secondary cases than outbreaks with all other modes (χ^2 > 2.6, p<0.002). When the outbreaks with unknown modes of primary transmission are omitted, there is still a statistically significant difference in secondary case rates ($F_{3,60}$= 6.7, p=0.030) with waterborne outbreaks having statistically significantly higher rates of secondary cases than foodborne outbreaks (χ^2=8.6, p=0.003).

Figure 7.6: Rate of secondary outbreak cases per primary case, by mode of primary transmission

The rate of secondary outbreak cases per primary case, by mode of primary transmission shown on the log (1+x) scale. The means and 95% confidence intervals – indicated by the bars - are calculated from the univariate GLM model using back transformation of the parameter and standard errors. For Animal/Environmental n=9, Food n=38, Dairy Product n=11, Water n=6 and Unknown n=26.



7.3.4.3 Mode of primary transmission - multivariate models

No analysis was possible for the model with the three way interaction between Mode of Primary Transmission, Country and Median Age because there was a lack of counts for some mode-country-median age categories. When the variables Mode of Primary Transmission, Country and Median Age Category are all included in the model without an interaction term (Table 7.12), there are statistically significant differences in the rate of secondary cases between modes of primary transmission($p<0.001$) and between median age categories ($p<0.001$; Table 7.12). However, if the outbreaks for which the mode of primary transmission is unknown are excluded, the differences are no longer statistically significant ($p>0.165$; Table 7.12).

There are no statistically significant interactions between pairs of the variables (e.g. Primary Mode & Country, Primary Mode & Median Age and Country & Median Age) ($p>0.120$; Table 7.12), whether or not the outbreaks with an unknown mode of primary transmission are included. However, in the two analyses with two way interactions including Mode of Primary Transmission (Primary Mode & Country and

Primary Mode & Median Age), the overall differences in rates of secondary to primary cases between modes of primary transmission are statistically significant ($p<0.001$; Table 7.12).

Table 7.12: P-values for multivariate analyses including Mode of Primary Transmission
P-values for the multivariate analyses including Mode of Primary Transmission (ModeP). The column for 'With Unknown' includes outbreaks for which the mode of primary transmission was unknown N indicates the number of outbreaks included in the analyses, listed for analyses including and excluding outbreaks with an unknown mode of primary transmission. *** indicates an analysis that could not be performed because of a lack of particular interaction categories, and --- indicates analyses that were not applicable. A "*" indicates interaction between terms, a "+" indicates terms in the model without any interaction.

	N	With Unknown	Without Unknown
ModeP * County * Median Age	---	***	***
ModeP + County + Median Age	71, 50	ModeP p<0.001 Country p=0.194 Age p<0.001	p=0.165 p=0.760 p=0.488
ModeP * Country	86, 62	Interaction p=0.720 ModeP p<0.001 Country p=0.757	p=0.999 p=0.356 p=0.827
ModeP * Median Age	75, 52	Interaction p=0.120 ModeP p<0.001 Age p=0.458	***
Country * Median Age	71	***	---
ModeP + Country	86, 62	ModeP p<0.001 Country p=0.011	p=0.132 p=0.540
ModeP + Median Age	75, 52	ModeP p<0.001 Age p<0.001	p=0.380 p=0.746
Country + Median Age	71	Country p=0.479 Age p<0.001	---

In both analyses, when the interaction terms are removed, all the terms were statistically significant ($p<0.012$; Table 7.12). When Mode of Primary Transmission and Median Age only are included in the analysis, outbreaks in which the median age was less than 6 had statistically significantly higher rates than those of any other age ($p<0.040$) and those with a median age from 6 to 16 have a statistically significantly higher rate than those where the median age was 17 to 59 ($p<0.026$). When Mode of Primary Transmission and Median Age or Mode of Transmission and Country were included in the analysis, outbreaks where the mode of primary transmission was

unknown have a statistically significantly higher rate of secondary to primary cases than outbreaks with all other modes of primary transmission (all $p<0.001$). When the outbreaks for which the mode of primary transmission is unknown were excluded, none of the terms are statistically significant ($p>0.131$; Table 7.12).

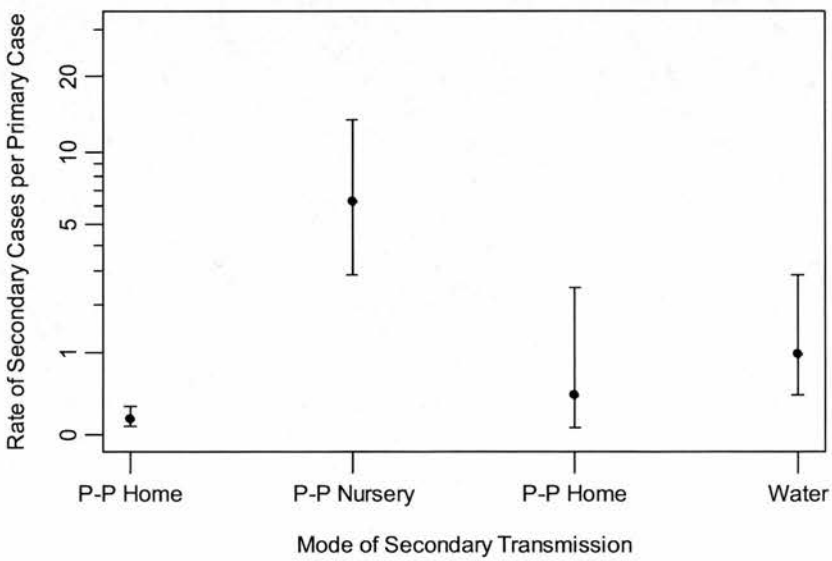
In addition, when only Country and Median Age are included the model, there are statistically significant differences in the ratio of secondary to primary cases between median age categories: the rate of secondary cases in outbreaks with a median age < 6 is statistically significantly higher than in those outbreaks with median case ages from 6 to 16 and from 17 to 59 ($p<0.01$). However, there is no overall statistically significant difference in the rate of secondary to primary cases between countries ($p=0.479$).

7.3.4.4 Mode of secondary transmission - univariate models

There is a statistically significant difference between modes of secondary transmission in the rate of secondary cases to primary cases ($F_{3,63}= 13.0$, $p<0.001$; Figure 7.7).

Figure 7.7: Rate of secondary outbreak cases per primary case, by mode of secondary transmission

The rate of secondary outbreak cases per primary case, by mode of secondary transmission, shown on the log (1+x) scale. The means and 95% confidence intervals – indicated by the bars - are calculated from the univariate GLM model using back transformation of the parameter and standard errors. For P-P home=41, P-P nursery n=10, P-P other n=7, and Water n=9.



Outbreaks where the mode of secondary transmission is person to person spread in a nursery setting have statistically significantly higher rates of secondary cases to primary cases than those where secondary transmission was by person to person spread in a home or other setting, or by water ($p<0.007$). Outbreaks where the mode of secondary transmission is water have a statistically significantly higher rate of secondary cases than those where secondary transmission is via person to person spread in a home ($p=0.002$).

7.3.4.5 Mode of secondary transmission - multivariate models

Again, the analysis could not be conducted when all three variables (Mode of Secondary Transmission, Country and Median Age) are included in the model with interaction terms. If the interaction terms are removed, there are statistically significant differences in the rate of secondary cases between modes of secondary transmission ($p<0.001$; Table 7.13) and between median age categories ($p<0.001$; Table 7.13).

Table 7.13: P-values for multivariate analyses including Mode of Secondary Transmission
P-values for the multivariate analyses including Mode of Secondary Transmission (ModeS). The column for 'With Unknown' includes outbreaks for which the mode of primary transmission was unknown N indicates the number of outbreaks included in the analyses, listed for analyses including and excluding outbreaks with an unknown mode of primary transmission. *** indicates an analysis that could not be performed because of a lack of particular interaction categories, and --- indicates analyses that were not applicable. A "*" indicates interaction between terms, a "+" indicates terms in the model without any interaction.

	N	With Unknown
ModeS * County * Median Age	---	***
ModeS + County + Median Age	57	ModeS p<0.001 Country p<0.001 Age p=0.756
ModeS * Country	64	Interaction p=0.720 ModeS p=0.004 Country p=0.695
ModeS * Median Age	57	Interaction p=0.158 ModeS p=0.114 Age p=0.210
ModeS + Country	64	ModeS p<0.001 Country p=0.468
ModeS + Median Age	57	ModeS p<0.001 Age p<0.001

When Mode of Secondary Transmission and Country are included in the model, there is no statistically significant interaction (p=0.711; Table 7.13), but whether or not the interaction terms is kept in the model, there is a statistically significant differences in secondary case between modes of secondary transmission (p<0.001; Table 7.13). As compared to outbreaks where secondary transmission was through person to person spread in nurseries or through water, outbreaks where secondary transmission was due to person to person spread in the home have statistically significantly lower secondary case ratios(p<0.002).

When Mode of Secondary Transmission and Median Age are included in the model with an interaction term, none of the terms were statistically significant (p>0.114), but if the interaction term is removed, there are statistically significant differences in the rate of secondary to primary cases between both modes of secondary transmission and median age categories (p<0.001). There is a statistically significantly higher secondary case ratio in outbreaks with a median age of less than 6 as compared to those with median ages of 6-16 and 17- 59 (p<0.001). In addition, outbreaks where the mode of secondary transmission was person to person in a home had statistically significantly lower rates of secondary to primary cases than outbreaks where secondary transmission was person to person in a nursery or through water (p<0.001).

7.4 Discussion

The results presented here provide what is believed to be the first large-scale systematic descriptive and statistical analysis of primary and particularly secondary cases in *E.* coli O157 outbreaks detailed in the published literature. These analyses suggest that there is a statistically significant relationship between the rate of secondary cases and modes of primary transmission, modes of secondary transmission and median age. However, there appears to be no such relationship between the rate of secondary cases and country. The results suggest that mode of secondary transmission and median age are the most important predictors of secondary case numbers and ratio of secondary to primary cases.

7.4.1 Issues with the data

However, some caution must be taken in interpreting these results. Firstly data availability was limited; for example, out of more than 350 outbreaks in the United States between 1982 and 2006 (Rangel et al. 2005)(official United States outbreak data set, Elizabeth Blanton, CDC), fewer than thirty are included in our analyses. As a result, the distribution of the outbreaks in our study between countries and modes of transmission may not be the same as the distributions in the population which includes all outbreaks reported to national surveillance organizations of countries included in this study. The actual number of outbreaks that have been reported to national surveillance organizations since 1982 varies considerably between countries, from fewer than five in Sweden (Soderstrom et al. 2005; Wahl & Andersson 2004; Welinder-Olsson et al. 2004; Ziese et al. 1996), to more than 90 in Scotland (Cowden 1997b; Locking et al. 2003c; Locking et al. 2004; Locking et al. 2006a) and more than 350 in the United States (Rangel et al. 2005). Thus while 40% of Swedish outbreaks are included in the study, fewer than 10% of United States outbreaks and less than 20% of Scottish outbreaks are included. Further, the study population (n=90) is biased towards outbreaks from the United States and England & Wales and towards outbreaks that are detailed in the literature published in English. Thus it is known that many Japanese outbreaks are not included. Additionally, the study population was limited to outbreaks in countries with developed surveillance programs. However, the distribution of outbreaks by mode of transmission in this study is similar to the proportion in reported outbreaks in the United States, Scotland and England & Wales. For instance, food was the mode of transmission in approximately 42% of outbreaks reported in Scotland from 1990-2004, 47% in the United States from 1982-2002, 38% in England & Wales from 1996-2005 (out of those with a known mode of transmission) and 45% in this study (Cowden 1997b; Rangel et al. 2005; SCIEH 1997a; SCIEH 1998b; SCIEH 2000b; SCIEH 2001b; SCIEH 2002d; SCIEH 2003c; Smith-Palmer et al. 2005; Smith-Palmer & Cowden 2004; Willshaw et al. 2006). Thus, except for the geometric mean ill results discussed below, the differences between the outbreak dataset for this study and the population of all reported outbreaks appear to be minimal. Therefore the data set is considered appropriate for the study. However the results from the study should only

be extrapolated to the population of reported outbreaks since the complete population of all outbreaks may be very different from that of those reported.

However, other factors should be considered in discussion of the results. Firstly because the decision whether to classify a case as primary or secondary was made based upon the information provided in each publication or report, the definition of a secondary case in this thesis may differ from that reported in some outbreak investigations (Cowden et al. 2001; Friedman et al. 1999). In many of these investigations, secondary cases were generally defined as being close and/or family contacts of primary cases (Bell et al. 1994; Friedman et al. 1999). In this chapter, all nursery and waterborne cases other than the index case(s) whom introduced the case into the nursery or contaminated the swimming water are considered secondary cases. Thus, the proportion of secondary cases in these outbreaks may be much higher than the proportion calculated when other definitions are used.

The 90 outbreaks included in the descriptive and statistical sections of our study, with a geometric mean of 13.5 (95% CI = 10.6, 17.1) confirmed cases, were on average, statistically significantly larger than the populations of reported outbreaks in the included countries, such as Scotland (4.1: 3.3-5.1) (1996 – 2003, Health Protection Scotland official outbreak data set) and United States (5.5: 4.9-6.2) (1996-2003, CDC official outbreak data set). This difference was anticipated because larger outbreaks are more likely to be reported in the literature. Large outbreaks have an increased likelihood of being reported because they are likely to be more thoroughly investigated. In addition, outbreaks with large numbers of cases are suited to case-control studies, which are carried out in many investigations detailed in published papers on *E. coli* O157 outbreaks (Allaby & Mayon-White 1995; Ostroff et al. 1990). Due to the bias, the results of this study are more likely to be representative of larger outbreaks. However, it is not known how the inclusion of smaller outbreaks would influence the rate of secondary cases. If smaller outbreaks are more likely to have secondary cases, the analyses would likely underestimate the rate of secondary cases because when the number of secondary cases is constant, the smaller the outbreak, the higher rate of secondary cases to primary cases. For the

same reason, if smaller outbreaks are less likely to have secondary cases, the analyses would overestimate the rate of secondary cases.

Since all of the outbreak reports on Japanese outbreaks mention asymptomatic cases, while many reports from other countries included in this chapter do not, it would appear that investigations that included broad-based screening of potential case patients and their contacts, such as those in Japan, are more likely to pick up asymptomatic cases. Studies have indicated that as many as 40% of secondary cases may be asymptomatic (Parry & Salmon 1998), and that secondary cases are, in general, less likely to have severe symptoms (Bell et al. 1994). As a result, the number of secondary cases reported may depend greatly on the level of testing and epidemiological study. Thus investigation into the rates and characteristics of asymptomatic cases in outbreaks would help in better understanding the true numbers and rates of primary and secondary cases.

Additionally, since many statistical tests have been performed in this chapter, the issue of multiple-testing must be considered. As discussed in Section 2.3.5.1, when a large number of tests of significance are performed, there is an increased risk of false positives. For this chapter, many of the same techniques to adjust and account for multiple testing presented in Section 2.3.5.1 were used. These include the use of F-tests for more conservative estimates of statistical significance, testing for overall statistical significance between categorical variables before examining contrasts and the presentation of actual significance values.

7.4.2 Exploration of the results

One of the most interesting findings in this chapter is the lack of statistically significant differences between countries in variables such as secondary case rate. Despite the fact that the countries in this study vary widely in factors which may affect infection transmission and frequency, such as population, population density and size, there were no overall statistically significant differences between countries in the mean number of confirmed, secondary and primary cases, in the median case age or in the rate of secondary cases. The statistically significantly higher mean number of ill cases in Japanese outbreaks could result from the investigative

practices in Japan (Sakuma et al. 2006). Japanese outbreak investigations can involve interviews and testing of all work, school and/or family contacts of known cases, thus increasing the number of ill cases recorded as part of the outbreak (Akashi et al. 1994; Sugiyama et al. 2005). This lack of statistically significant differences between countries suggest that outbreak size and rate of secondary cases are likely to be determined by more epidemiological factors or risks that are relatively independent of country, such as age and mode of transmission.

Indeed, the results of our analyses indicate that both mode of secondary transmission and median age are a statistically significant determinant in the rate of secondary cases. In particular, outbreaks where the mode of secondary transmission was person to person spread in nurseries are unique from those spread by other means in having statistically significantly higher median numbers of secondary cases and statistically significantly higher rates of secondary cases. The association between outbreaks in nurseries and secondary cases has been noted (Armstrong et al. 1996; Coia 1998b; Coia 1999), but this study is the first to statistically test and quantify this association, and to note statistical differences in secondary case rates between other modes of secondary transmission.

These findings can be explained, in part, by the ease of transmission between infants and toddlers in nursery settings, where close contact of persons with poor personal hygiene skills, and a higher likelihood of shedding the bacteria for extended periods of time (Swerdlow & Griffin 1997) provides many opportunities for direct or indirect transmission of contaminated faecal materials (Coia 1998b). Further transmission then occurs through contact with the nursery pupils by teachers, parents and siblings, though the majority of cases in nursery outbreaks are the nursery pupils themselves (Al-Jader et al. 1999; Allaby & Mayon-White 1995; Cheasty et al. 1998; Incident Control Team NHS Fife 2007; Public Health Laboratory Service 2000a; Sugiyama et al. 2005).

While secondary cases have been noted in outbreaks linked to primary and secondary schools (Belongia et al. 1991; Michino et al. 1998; SCIEH 1999), secondary transmission in these outbreaks seems limited to family members (Belongia et al. 1991; Michino et al. 1998; SCIEH 1999). This suggests that that older pupils may

have good enough hygiene skills (Eves et al. 2006) to prevent transmission outside settings of close family contact (i.e. primary school pupils are highly likely not to be wearing diapers, to be toilet trained, to be able to reach sinks and soap dispensers etc.). Also, all of the primary or secondary school outbreaks included in this study were thought to be the result of contaminated food consumed by students at the school (Belongia et al. 1991; Michino et al. 1998; SCIEH 1999). In contrast, all the nursery outbreaks in this study were suspected to have been triggered by the presence of one or more pupils infected outside the nursery (Allaby & Mayon-White 1995; Belongia et al. 1991; Belongia et al. 1993; Galanis et al. 2003). Residential facilities such as nursing and care homes are also are common locations for outbreaks involving secondary transmission (Coia 1998b; Griffin & Tauxe 1991). However, as with the primary/secondary school outbreaks, the majority of cases in the care home outbreaks in this study resulted from ingestion of contaminated food with subsequent secondary transmission (Lior 1983; Reiss et al. 2006).

The analyses in this study suggest that the mechanism(s) behind these differences in secondary case rates between secondary modes of transmission is likely to be linked to the age of infected persons. The results in this chapter provide evidence for the link to age. For instance, if median age is added to the model containing mode of secondary transmission, outbreaks where the mode of secondary transmission was person to person in a nursery no longer have statistically significantly higher rates of secondary case than outbreaks where secondary transmission was via water. However, outbreaks with secondary transmission through person to person contact in the home still have statistically significantly lower secondary case ratios that those outbreaks spread by water or person to person contact in the nursery. In addition, lower median age was associated with a higher rate of secondary cases, with outbreaks with a median age of less than 6 years having statistically significantly higher rates of secondary cases that those outbreaks where median age was 6-16 and 17 – 59. Association between young age and high rates of secondary transmission has been noted in a study where index cases under the age of 15 were mostly likely to transmit infection to a household contact, with household contacts between 1 and 4 years old the most likely to become infected (Parry & Salmon 1998).

Initially, the mode of primary transmission appeared to have a statistically significant role in determining the rate of secondary transmission. However, comparison of results from analyses including and excluding outbreaks where the primary mode of transmission was unknown revealed that all the statistically significant differences in secondary case rates were between the unknown outbreaks and outbreaks in one of the known mode categories. The statistically significantly higher ratio of secondary to primary cases in outbreaks with unknown modes of primary transmission is most likely caused by the secondary transmission related characteristics of these outbreaks. More than two-thirds of outbreaks where the primary mode of transmission was not known involved secondary spread through nursery person to person contact or recreational water ((Ackman et al. 1997; Allaby & Mayon-White 1995; Friedman et al. 1999; Paunio et al. 1999; Sugiyama et al. 2005)), the two modes of secondary transmission categories with the highest rates of secondary cases. Therefore it is not unexpected there would be a high rate of secondary cases amongst outbreaks of unknown primary mode of transmission

When the unknowns were excluded, outbreaks where water was the mode of primary transmission had a statistically significantly higher rate of secondary cases than those spread by food. The fact that all of the outbreaks in which water was the primary mode of transmission had a median age of 16 or under, suggests that the difference in secondary rates between waterborne and foodborne outbreaks is related to the age of the infected persons. Thus, it would appear that age is likely to be one of the predominant factors in determining the rate of secondary transmission.

7.4.3 Implications for public health practice

The results of the study in this chapter indicate that the median age of infected persons and the mode of secondary transmission are statistically significant factors in determining the rate of secondary case transmission. Statistically significantly higher rates of secondary transmission were associated with person to person spread within nurseries and via water, as well with outbreaks where the median age was less than 6 years. It should be noted that the outbreaks included in the study appeared to be statistically significantly larger than those in the total population of reported outbreaks, thus the results may not encompass all outbreaks. In particular, the

secondary spread within outbreaks solely within private homes, which by nature tend to be small and not published in the literature due to privacy concerns, may not be properly represented in these analyses. Nonetheless, the results as summarised above, suggest that prevention measures aimed at reducing person to person transmission in nurseries and swimming areas (natural or pool) as well as other locations where young children are likely to be involved would help to reduce the size of outbreaks and reduce the overall number of infections.

Such measures, as suggested by the recommendations of the 2001 Task Force on *E. coli* O157 (Task Force on E.coli O157 2001), include promoting good personal hygiene, especially hand-washing in settings such as nurseries and playgroups. This has been done in Scotland through the Food Standard Agency's "Food Hygiene" campaign (Food Standards Agency 2007) with similar programs in other countries such as Canada and the United States. Additionally, with a specific focus on nurseries, recommendations which have been put into action are the exclusion of all infected or probably infected children as well as contact siblings under age 5 from nurseries. Prompt recognition of symptoms and exclusion of suspected cases is vital, which requires education of both nursery staff and parents. Given the risks to such young children, firm regulation of nurseries in terms of hygiene standards and education of parents as to the risk involved with non-licensed nurseries, would seem advisable.

Whilst transmission via animal contact was not specifically addressed in this study due to a lack of reports on such outbreaks that met the conditions for inclusion, the results suggest that the measures to reduce animal to human transmission, such as the "Shedding Light on *E. coli*" program (Scottish Executive 2007b) and those set out in the United States (Centers for Disease Control and Prevention 2007b) should be continued and promoted. The support for such measures comes from the findings that outbreaks in which the median age is less than 5, young children frequently being part of trips to open farms and other locations where animal contact is prevalent and less likely to be able to maintain good hygiene, have a statistically significantly higher rate of secondary transmission. Another important measure, as recommended by the task force, is educating parents who work with animals to

remove work clothes prior to contact with children and keep such garments where they cannot be accessed by children.

Finally, the demonstration of higher rates of secondary transmission via water in this study also suggests that measures to promote proper hygiene in pools and natural swimming areas (beaches, lakes and rivers) should continue to be promoted. Again outbreaks in such facilities are likely to involve children, and it is important to reduce the risk of human faecal material from entering the water. Strict measures should be in place to prevent un-toilet trained babies and toddlers from entering pools and public bathing facilities, including signage and withdrawal of usage privileges from patrons who abuse the regulations. Since it is not feasible or desirable to prevent families with young children to visit swimming facilities, alternatives should be provided for young children such as specific toddler pools which are carefully monitored and cleaned regularly and/or staffed crèches for babies. Additionally, and vitally, parents and pool staff must be educated as to the risk faecal contamination presents in swimming facilities. This includes for the staff and managers the need for chlorine levels to be carefully monitored, for both parents and staff, the need for contamination promptly reported to staff and the need to (and reasons why) children with current of very recent bouts of diarrhoea should be excluded from public bathing facilities. Many of these measures are included in current programs like the CDC's "Healthy Swimming" (Centers for Disease Control and Prevention 2007c), and should be continued and enhanced as appropriate for the needs and existing regulations of each region or country.

7.4.4 Conclusion

This chapter provides the first multi-country quantitative description and analysis of secondary cases in *E. coli* O157 outbreaks in Scotland, England & Wales, Canada, the United States, Japan and Scandinavia during the last two decades. The results of this study indicate that approximately 19% of outbreak cases are secondary, supporting earlier reports that "secondary transmission after point-source outbreaks of *E. coli* O157:H7 is common" (Armstrong et al. 1996), but suggests that modes of secondary transmission and median age, rather than country are important in determining the secondary case characteristics of *E. coli* O157 outbreaks.

Statistically significant differences in rates of secondary cases were found between modes of secondary transmission, specifically between outbreaks spread through person to person contact in nurseries and all other modes, as well as between outbreaks where the median age was less than 6 and those with a median age of 6 – 16 or 17 – 59. The statistically significantly higher rates of secondary transmission in outbreaks with low median ages and secondary transmission through person to person spread in nurseries confirms the importance of simple, but effective strategies to prevent person to person spread amongst children (Al-Jader et al. 1999; Scottish Executive 2007b). The reasons behind these differences, however, are likely to be complex, and further research using more comprehensive outbreak reports, and including other variables such as location would help further elucidate the reasons for the differences that have been described.

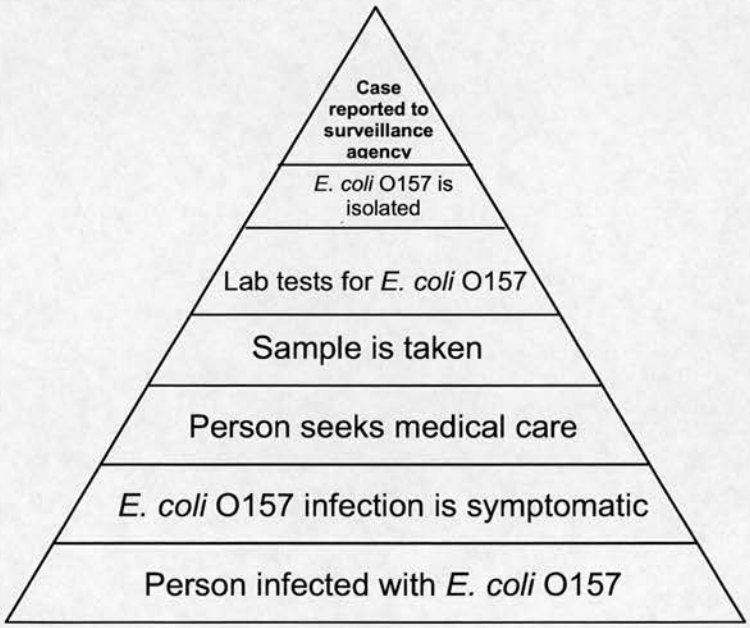
In the next chapter, Chapter 8, the issue of under-reporting of *E. coli* infections will be discussed and a framework for estimating the true prevalence of infections will be presented and tested using data from Scotland.

Chapter 8 -- A framework for estimating the true prevalence of *E. coli* O157 in Scotland

8.1 Introduction

Data from a number of studies, detailed in Section 1.9.1, suggests that the majority of cases of infectious intestinal disease are not reported (Cumberland et al. 2003; Wheeler et al. 1999). In particular, *E. coli* O157 cases in countries such as the United States, Canada and England & Wales are under-reported to surveillance agencies at rates of up to 98% (Adak et al. 2002; Bender et al. 2004; Mead et al. 1999; Michel et al. 2000). In order for a case of *E. coli* O157 to be reported via active or passive surveillance, it must pass through a number of steps which can be illustrated in 'reporting triangles'. Examples of triangles for general infectious intestinal illness were discussed in Section 1.9.2, but the steps for *E. coli* O157 in particular are illustrated in Figure 7.1.

Figure 8.1: Reporting triangle for *E. coli* O157
Generally reporting triangle for *E. coli* O157 showing the seven stages from infection to reporting of infection to a national surveillance agency.

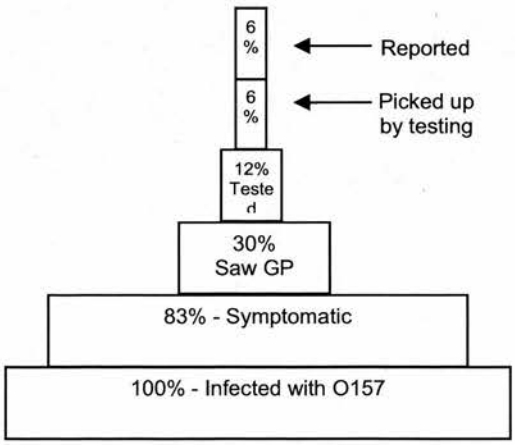


A person infected with *E. coli* O157 first has to have a symptomatic infection, and the symptoms must be serious enough to cause the person to be seen by a GP or other medical professional. Once an infected person seeks medical care, the medical practitioner has to take a sample (usually faecal), and the laboratory must test the sample for *E. coli* O157, and the test be sensitive enough to pick up the presence of the bacteria, toxin or antigen (see 1.4.5). Then to be reported, the test must be positive for *E. coli* O157 and the national surveillance agency notified by the

laboratory or other related medical personnel (Angulo et al. 1998; MacDougall et al. 2007).

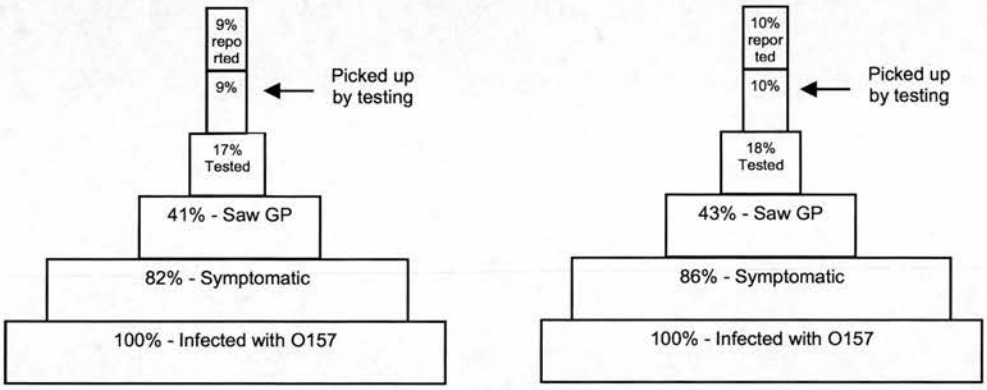
Estimates for the actual prevalence of *E. coli* O157 cases have been calculated in national and regional studies, introduced in Section 1.9.1 (Michel et al. 2000; Thomas et al. 2006; Wheeler et al. 1999). These studies used physician, laboratory and population surveys combined with other data to calculate the approximate number of cases that are lost to the surveillance system at each step of the triangle (Michel et al. 2000). The IID study (see Section 1.9.1) indicated that 50% of *E. coli* O157 cases in England & Wales were picked up by surveillance. In the United States, sporadic *E. coli* O157 infection rates in FoodNet surveillance sites were estimated through the use of series of patient, physician and laboratory surveys, and results suggested that 2-9% of cases were detected via FoodNet active surveillance (Bender et al. 2004; Hedberg et al. 1997). This calculation was later used to estimate that there are approximately 73,480 *E. coli* O157 cases per year in the United States (Mead et al. 1999). Another study of the FoodNet network indicated that stool samples were only collected for about one third of outbreaks where the etiology was not known, and just over half the time these samples were tested for *E. coli* O157 (Jones et al. 2004). Based up these studies and data from other papers (Angulo et al. 1998; Belongia et al. 1993; Deneen et al. 1998b; Imhoff et al. 2004; Michel et al. 2000; Voetsch et al. 2004a), a reporting triangle was constructed (as part of this chapter) for the United States (Fig. 7.2).

Figure 8.2: Reporting triangle for *E. coli* O157 in the Untied States
Reporting triangle for the United States



Using two models – one based upon adjusting for under-reporting at each level of the reporting triangle and the other based upon hospitalization data – Canadian researchers posited that for each case reported in Ontario, 4 – 8 cases were not reported (12-22% are reported)(Fig. 7.3) (Michel et al. 2000).

Figure 8.3: Reporting triangles for Canada
Reporting triangles for Canada based upon (a) all data from Michel et al 2000 and (b) data on Canada and Ontario from Michel et al and other sources.



A later study based on results from the National Studies of Acute Gastrointestinal Illness (NSAGI)(see Section 1.9.1), in which data was recorded from two provinces for the population study and nationwide for the laboratory study, resulted in higher estimates of under-reporting (Thomas et al. 2006). The estimates, ranging from 2% to 9%, are similar to those found in the above mentioned United States FoodNet study. No study estimating prevalence in Scotland has been published, thus it is not known whether the high reported rates (Locking et al. 2003b) are true or merely a reflection of a high rate of infection detection.

However, the above methods of calculating *E. coli* O157 prevalence provide only rough estimates of true prevalence and the techniques have a number of inadequacies. For instance, estimation of occurrence based upon stepwise calculation of the proportion of cases lost to the surveillance system is dependant on the level of uncertainty in the accuracy of the data and/formulas used (Michel et al. 2000). As an example, some formulas and entire estimates were based either on sporadic (Bender et al. 2004) or outbreak (Michel et al. 2000) cases (see Section 1.6.6 for information on outbreak and sporadic cases). Thus, the results may not be applicable to actual populations in which cases are distributed between outbreaks and

sporadic cases. Further, the IID calculations were based on just one case of *E. coli* O157 infection detected in the course of the study. Since the total study population was not more than 459,975, and the rate of *E. coli* O157 infection in England & Wales has been above 1 case per 100,000 since 1996 (Locking et al. 2003b), the study appears to have undercounted even the number of reported cases. It is thus unlikely to have provided an accurate representation of the true prevalence of confirmed cases in England & Wales. In addition, in terms of identifying the patient population which has the potential to seek medical care, there is no set definition as to what constitutes an *E. coli* O157 illness. In some studies, symptoms were defined as bloody diarrhoea (Bender et al. 2004; Hedberg et al. 1997), whilst some formulas took both bloody and non-bloody diarrhoea into account (Michel et al. 2000), and others involved all persons with any symptoms of infectious intestinal disease (Cumberland et al. 2003). Finally, all of the studies involved only a limited geopolitical region – Ontario in Canada, and the FoodNet regions in Oregon, California, Connecticut, Georgia and Minnesota in the United States – and thus the estimates based on data from these regions may not be applicable to regions where surveillance, diagnostic, medical practices or even underlying infection rates differ.

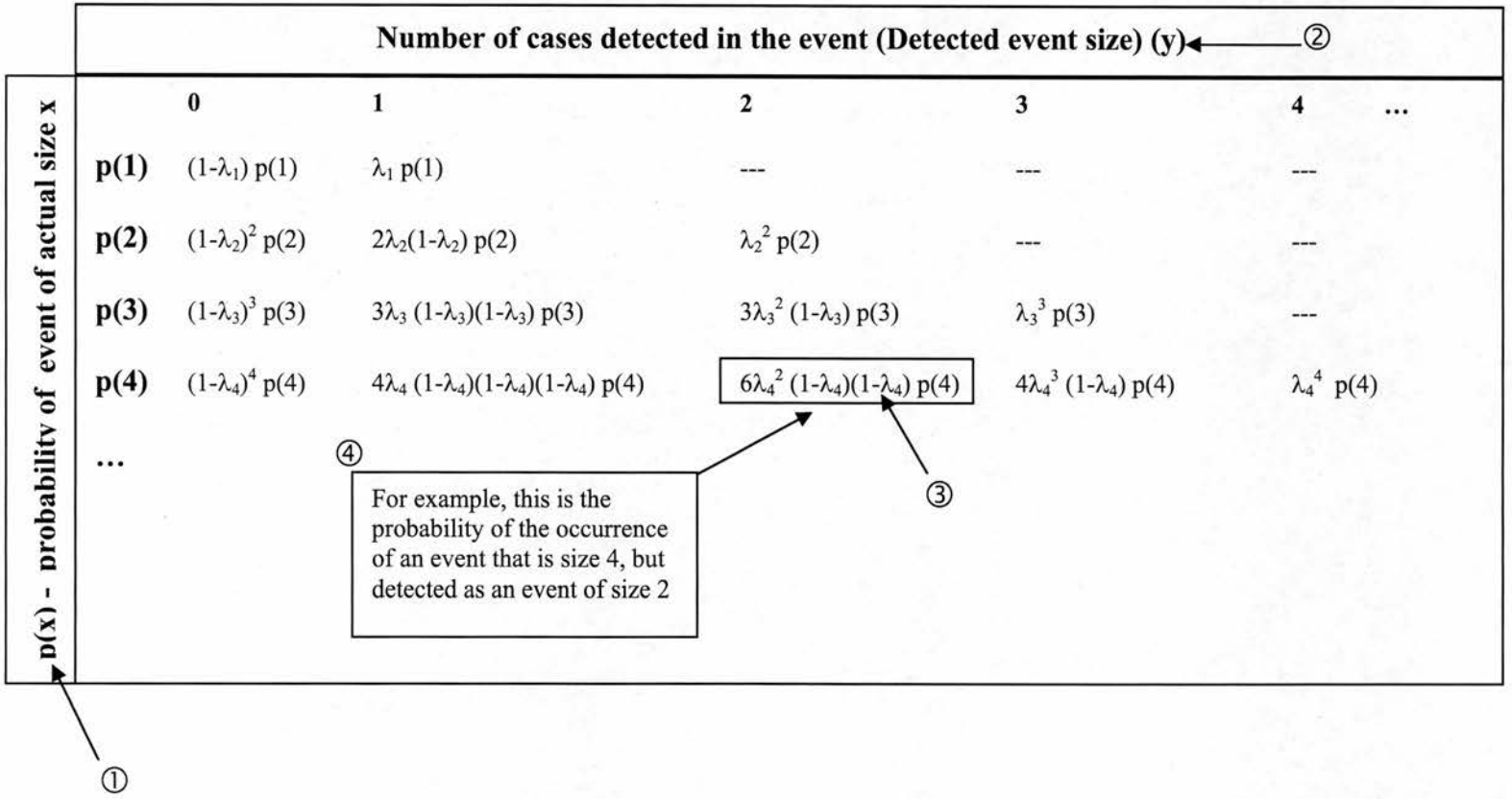
The development of a more accurate technique for estimating *E. coli* O157 prevalence is of considerable importance to public health epidemiologists and economists because incomplete knowledge about the true prevalence makes it difficult to allocate resources for prevention, treatment and research, and to gauge the effectiveness of and biases in surveillance efforts (Silk & Berkelman 2005; Thomas et al. 2006). More informed allocation of resources is particularly important in countries where there has been a historically high rate of *E. coli* O157 infection, and a significant financial impact from outbreaks (Roberts et al. 2000). As a result, estimating the true prevalence in Scotland, which has had some of the highest rates of infection in the world (Locking et al. 2003b; O'Brien et al. 2001b), and for which no studies estimating the degree of under-reporting have been published, would be of particular interest. Data from Scotland is particularly suited for prevalence estimation because outbreak and sporadic case data has been collected via a centralised, consistent and comprehensive surveillance program since 1996. As a result, the data on Scottish outbreaks and cases is less likely than that in Canada or

the United States to be affected by biases created by both changes in surveillance systems and variations in surveillance within a country. In analyses of temporal trends such as those in Chapters 2-6, these biases can be corrected for, or the effects of the biases considered in the interpretation of results. However in this chapter, the data is considered as a single distribution of outbreak sizes, and thus to correct for biases is more difficult. Additionally, single case data – the source of sporadic case counts – in the United States and Canada is captured by systems very distinct from outbreak data (CDC 2007), as compared to Scotland, where all *E. coli* O157 data is collected by closely linked systems. In this chapter, a new framework for estimating prevalence will be explored. The framework will be tested in order to determine whether it has potential for use with Scottish data, and thus to be further developed into a proper model. If the approach proves appropriate, it may lead to a model that can overcome many of the uncertainties involved with estimations based on sequential and hospital data and thus provide further insight into the extent of *E. coli* O157 detection in Scotland and in other countries where outbreak and sporadic data is available. In addition the framework may have applications in estimating prevalence for other infectious diseases including *Salmonella* and *Campylobacter*, for which as explained in Chapter 1, case numbers are underreported. Finally, the results of the framework exploration will be discussed with relation to the known issues regarding the Scottish data, and the biological distributions that underlie the best fits for the framework.

8.2 Materials and methods

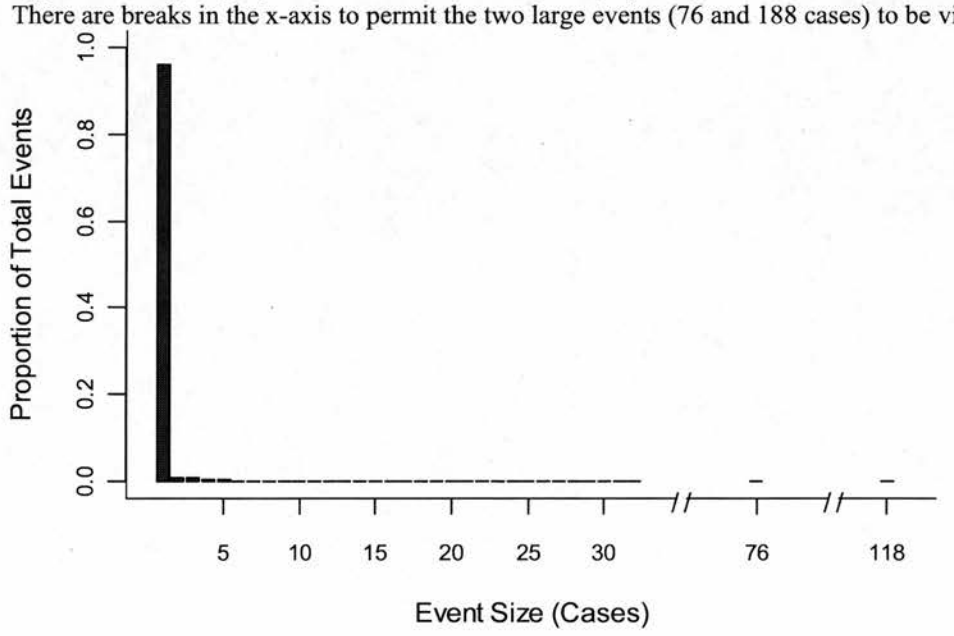
The framework for estimating prevalence of *E. coli* O157 in Scotland that is presented in this thesis is based on the fact that each *E. coli* O157 event (an incident of *E. coli* O157 infection, whether a single case or a group of epidemiologically linked cases) is either detected or not detected, but not every case is necessarily detected. Therefore, a proportion of events are only partially detected – i.e. not all cases in the event are detected, and thus each event has a detected size and an actual size, which may not be the same. The framework uses a matrix, shown in Figure 8.4, which allows the proportion of total cases comprised of each combination of actual size (x) and detected size (y, where $y \leq x$) to be calculated.

Figure 8.4: Matrix used to determine probabilities of actual event occurrence
The matrix used to calculate the probability of occurrence of an event of actual size x and detected size y, p(x, y)



The equation for each actual and detected size combination within the matrix is constructed through several steps. Firstly, the matrix is split into rows by the probability of actual event occurrence (p(x)), which is presented as a distribution, and into columns by the detected event size (y). The conditional probability for each detected event size is represented through an expansion of the binomial distribution since the probability of detection is binomial – an event is either detected or not detected at a certain size. These conditional probabilities multiplied by p(x) give the probabilities of particular observed numbers of cases from each event size. Within the binomial equations, the probability of detection of a case (λ) is also expressed as a function of the event size. The selection of distributions/functions will be discussed later in Methods and Materials (Section 8.3.1). These proportions (out of all detected events, e.g. detected size ≥ 1) for each detected event size (i.e. column) are then summed to produce a summary table and bar plot of the distribution of the proportions of sizes of detected events. In the bar plot, each bar is split to indicate for each detected event size, the proportion for each actual event size. This plot can then be compared to the plot of the proportion of event sizes in Scotland from 1996 – 2004 (Fig. 8.5).

Figure 8.5: Plot of the proportions of events, by size: Scotland 1996 – 2004
A plot showing the proportions of *E. coli* O157 events, by size, in Scotland from 1996 to 2004. There are breaks in the x-axis to permit the two large events (76 and 188 cases) to be visible.



Additionally, in the summary table, the proportions are multiplied by the number of observed events (1996 – 2004) to create tables of counts by event size (see Table 8.1) which can be compared with the real event counts for the same time period.

Distributions or functions (for both the probability of actual event size and event detection) and values for parameters in these distributions which are suggested both by actual biological mechanisms and by data from Scotland are used. The aim in using such distributions/functions is to see whether it is possible to create a detected size probability distribution for events that matches closely to the proportions observed in Scotland. Since the observed distribution in Scotland only shows events that are detected (size>0), the distribution created in the framework is adjusted so that the proportions are only out of those events that are at least detected size one (i.e. detected events). If a distribution constructed using the framework that matches the observed distribution, then the actual prevalence of events can be estimated. The actual number of events is estimated from the proportion of events that are detected as size zero (i.e. not detected) in the unadjusted data and the number of events that took place during the study period. Actual case prevalence is calculated by then multiplying the number of events that are not detected by the proportions of those events that are of each actual event size. For instance, if 100 events went undetected, and 50 of those were of actual size 1, and 50 were of actual size 2, there would be 50x1 + 50x2 or 150 cases that went undetected, so the true number of cases would be the number of detected cases plus 150.

8.2.1 Data sets

All data on confirmed (ill and positive) cases of *E. coli* O157 in Scotland for 1996 to 2004 was taken from reports issued by HPS, with outbreak data from the yearly Surveillance of Outbreaks of Infectious Intestinal Disease reports (SCIEH 1997a; SCIEH 2000b) and data on sporadic cases from the HPS website. Sporadic cases are those cases which are not epidemiologically linked to any cases other than those within the same household (see 1.6.6.1).

8.2.2 Definitions

Event

An event is considered to be a single incident of *E. coli* O157 infection, either a sporadic case or an outbreak.

Detected case/event

A detected case is defined as a laboratory confirmed case of *E. coli* O157 that is reported to Health Protection Scotland, and a detected event as an event involving at least one laboratory confirmed case of *E. coli* O157.

Actual Case/Event

An actual case or event is defined as an incidence of *E. coli* O157 infection – either a single case or outbreak – that could be detected by normal laboratory techniques, but does not necessarily end up being detected and/or reported to Health Protection Scotland. Thus the population of actual events includes both detected events and undetected events.

Observed case/event

A laboratory confirmed and reported case of *E. coli* O157 or an event involving at least one laboratory confirmed case of *E. coli* O157, as per HPS definitions and occurring during the time period from 1996 to 2004.

8.2.3 Framework construction

The process of creating a viable framework includes three steps:

- **Distribution selection.** First, potentially appropriate distribution(s) and/or functions for the distribution of the probability of actual event size occurrence and the probability of event detection must be selected for further consideration. For the probability of actual event occurrence, several discrete distributions, which are associated with biological mechanisms, are examined in detail. These distributions include Poisson, Logarithmic and Negative Binomial. The Logistic Growth Law function is considered for the probability of event detection because of the flexibility of the function shape which allows for a sigmoidal curve (see Section 8.3.1.2).

- **Exploration of the distribution for the probability of actual event occurrence.** Secondly, the parameter(s) for each of the selected probability of actual event occurrence distributions must be explored in terms of how the parameter values relate to the shape of the distribution and thus which, if any, distributions and parameter values or ranges are appropriate for the framework. The process includes relating the distributions and parameter values to biological mechanisms associated with infectious disease epidemiology. For instance, Lloyd-Smith (2007) indicates that the range of negative binomial parameter k is 0.032 to 5.2 in uncontrolled outbreak data sets, so when the negative binomial distribution is used, values of k within this range will be tried first.

- **Case detection function exploration.** Thirdly, the parameters for the function for the probability of event detection are examined in the same manner as discussed above

- **The complete framework.** Finally, the framework is completed by the addition of the probability of event detection function. The complete framework can then be used to calculate the distribution of the proportions of detected events. The fitting of the values to the framework will be an iterative process, with series of values for each parameter tried in order to find the values which produce the most appropriate distribution(s) of the proportions of detected event sizes. The process of finding or approaching an appropriate combination of distributions/functions is of importance, so examples will be used to illustrate why particular distributions and/or parameters are not appropriate. The exact construction of the matrix and how it is used to create the proportion of detected event size distribution will be detailed in Section 8.2.4.

8.2.4 Matrix construction and usage

The framework for the distribution of the proportions of detected event sizes is derived from the matrix in Figure 8.4. The construction of the matrix is based on two assumptions: (a) that an event is either detected or not detected but (b) that not all cases in each event are necessarily detected and so an event of size x may be detected as an event of a size y, where y is less than or equal to x. Therefore, each event has a detected size y and actual size x, and the equation for each actual and detected size combination (x,y) can be constructed through a series of steps:

(1) The rows in the matrix represent the probability of occurrence of an event of actual size x (p(x)), where x is an integer from 1 to 130. 130 has been chosen as the upper limit because the largest event size (when the Wishaw Outbreak is split into separate cohorts – see 3.2.3.2) is 118. This probability, which varies depending on the value of x, is expressed in terms of a distribution, the selection of which is discussed below in 8.3.1.1.

probability of an event of size x = p(x), where x= 1,2,3…130

(2) The columns represent the detected size of the event (y). The distribution of y is conditional because it is dependent on actual size (x). Thus there is a conditional probability (h) for each detected size dependent on actual size (h(y|x)) (e.g. there is a probability for each combination of actual and detected event size) and the conditional distribution is the binomial distribution where y is an integer ≥ 0 and x is an integer ≥ 1:

$h(y|x) = x!/\ y!(x-y)! * \lambda^{y} * (1-\lambda)^{x-y}$ x = 1,2,3…130 and y = 0,1,2…130

$$\sum_{y=1}^{x} h(y \mid x) = 1$$

(3) Within the binomial distribution, λ is the probability of an individual case being detected. The probability of a case being detected would appear to be likely to vary depending on the size of the event – a case in a large event is more likely to be detected than a single case, since an outbreak where a number of people are ill is less likely to go undetected and more likely to be intensively investigated than a single case. Therefore, while lambda may be a

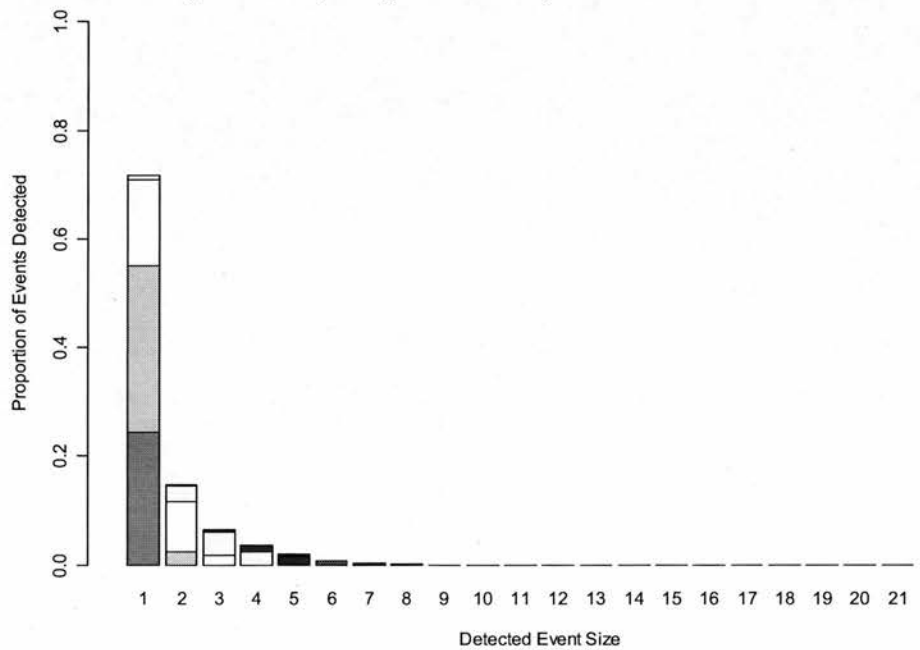
single value, it is probably a function of event size λ(x). The selection of lambda will be discussed in 8.3.1.2.

(4) Therefore, the final equation for the probability of each x and y combination is the binomial equation, which represents the conditional probability of detected event size, multiplied by p(x), the probability of actual event occurrence:

$$p(x, y) = p(x) * h(y|x) \qquad \sum_x \sum_y p(y, x) = 1$$

When parameter values are inserted and the framework is tested, a value for the probability (i.e. proportion out of the total) of each detected-actual event combination is produced. The distribution of the proportions of detected event sizes ≥1 is then obtained by calculating the sum of the probabilities for each y (detected event size), and dividing each sum by the overall proportion of events where y>0. The proportions for each detected event size are then illustrated in a bar plot within which each bar is split into bands indicating the proportions of actual event sizes for each detected event size.

Figure 8.6: Distributions of event size proportions
Distribution of detected event size proportions, with the coloured bands indicating proportions of actual event size (red = size 1, orange = size 2 etc.) within each detected event size.



This is illustrated in Figure 8.6, where red bands indicate actual outbreak size of 1, orange bands actual outbreak size of 2 etc. In this example, approximately 70% of detected events are detected as events of size 1, in other words, as sporadic cases, but only one third of the events are actually size 1 (red band). The remaining two thirds (orange band etc.) of the events are in fact size 2 or larger, but not all of the cases have been detected. This plot is compared visually with the plot of the observed distribution of the proportions of event sizes in Scotland, based on the data from 1996 to 2004. In addition, the proportions for each detected event size are multiplied by the number of total events between 1996 and 2004 (the study time period), to produce a summary table of event size counts (see Table 8.1). These counts are compared to the actual counts between 1996 and 2004, providing a numeric comparison between the model distribution and the observed distribution. If an accurate fit is found based on the plot shape and summary counts, the actual prevalence of events can be estimated from the proportion of events that are not detected (detected size=0). If, for instance, 90% of events are not detected, the actual event prevalence is ten times the number of events observed/reported. Case prevalence can then be calculated by multiplying the number of events of each size by the event size.

All analyses and plots are done using functions written for "R" (R Development Core Team 2007).

8.3 Framework construction

8.3.1 Distribution selection

In order to come up with non-arbitrary values for the distribution of the probabilities of actual event occurrence, p(x), and function for the probability of event detection, $\lambda(x)$, are assumed to be in the form of distributions or functions. Since the data for actual events is in the form of counts, the distribution is assumed to be discrete, and the related distributions Poisson, Logarithmic and Negative Binomial are all considered because they are frequently used to describe biological populations involving discrete data (Bliss & Fisher 1953; Cassie 1962; Johnson et al. 1992; Sampford 1955).

8.3.1.1 Probability of actual event occurrence

Regardless of the distribution selected, since an actual event must have at least one case (size = 1), the distribution of (the probability of) actual event occurrence (by size) must be truncated at 1.

Continuous distributions such as the gamma, exponential and Weibull were discounted because they are appropriate for continuous data and outbreak data by nature, since there cannot be a fraction of an outbreak, is in the form of non-zero integers. Neyman's A Distribution is also not considered because references to use in biological instances are with a non-truncated form (Beall & Rescia 1953; Evans 1953; Johnson et al. 1992; Martin & Katti 1965; Neyman 1939).

Truncated Poisson distribution

The first distribution considered is the Poisson distribution (Poisson 1837), which has been used earlier in this thesis in the context of an error structure for GLMs (2.2.6.2) and describes a situation where events occur independently and at random (Kirkwood & Sterne 2003). The Poisson distribution is frequently used in describing biological processes (Elliott 1971b; van Rest & Parkin 1933) because it is a discrete distribution in that it used in instances where there is integer data, such as counts (Crawley 2007). In infectious disease statistics, count data is very common (e.g. outbreaks, cases, number of bacteria), and, for example, the distribution of *E. coli* O157 in food items implicated in a Japanese outbreak has been assumed to have a Poisson distribution (Tennis et al. 2004).

The distribution has one parameter, lambda (λ) which is equivalent to the mean, and the mean is equal to the variance ($\mu = \sigma^2$) and (Crawley 2007; Johnson et al. 1992; Poisson 1837) the probability of a count of x is:

$$p(x) = \frac{e^{\lambda}\lambda^{x}}{x!}$$ where ! indicates factorial and e is the exponential constant ≈ 2.718

When the Poisson distribution is left truncated, in that the equation is adjusted to restrict values to a set minimum, it is referred to as the positive or conditional Poisson (Johnson et al. 1992). A positive or zero truncated Poisson distribution has been used to describe the distribution of counts of lymphatic filariasis microfilariae in human blood samples (Grenfell et al. 1990) and gall fly egg frequencies (Finney &

Varley 1955). For the positive Poisson (Johnson et al. 1992), the probability of a count is expressed as:

$$p(x) = \frac{\left(1 - e^{-\lambda}\right)^{-1} e^{-\lambda} \lambda^{x}}{x!} \quad \text{where } \lambda \text{ is the mean count per unit and is } \geq 0$$

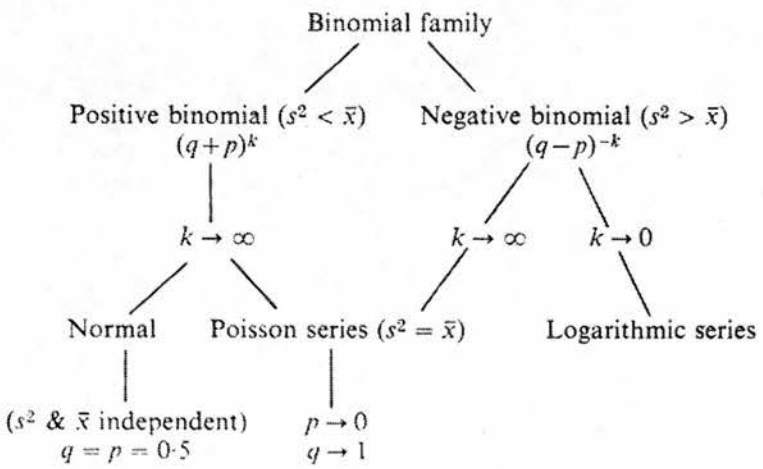
However, due to a number of issues, the truncated Poisson distribution is not considered to be a likely choice for the distribution of the probability of event occurrence. Firstly, the variance to mean ratio of the observed Scottish *E. coli* O157 data is approximately 37 to 1. Though the actual variance to mean ratio undoubtedly varies from the observed ratio, the actual range in event size should most likely be similar or larger to the observed range because events with small numbers of cases are the most likely to have gone undetected. Thus, there would be little difference between the actual and observed ranges, so the variance would still be larger than the mean. Since the mean to variance ratio for a Poisson distribution must be approximately one, and the difference between variance and mean still not large when the distribution is truncated, it would not appear to be appropriate for describing the *E. coli* O157 data. Also, one of the more rigid assumptions of the Poisson distribution is that the events occur at random – i.e. with equal probability of occurrence (Elliott 1971a). However, it has been shown that there are "significant clusters of sporadic *E. coli* O157 infections" in Scotland, with a suggested link to cattle to human population ratios or other rural factors (Innocent et al. 2006), and *E. coli* O157 cases are known to cluster temporally with infections occurring more often in the summer or early autumnal months (Douglas & Kurien 1997). Thus it is unlikely that *E. coli* O157 event occurrence is truly random, and so the use of a Poisson distribution is probably not appropriate.

Truncated negative binomial distribution

The second distribution considered for the distribution of the probability of actual event occurrence is the truncated negative binomial distribution. This distribution, which is used to describe aggregated populations where the variance is greater than the mean (Bliss & Fisher 1953; Elliott 1971b), is an increasingly widely used distribution, especially in biological contexts, because the assumptions are not as rigid as compared with the Poisson distribution, e.g. mean equal to the variance

(Elliott 1971b). Uses of the negative binomial distribution in biological contexts include describing the distribution of benthic invertebrates in bottom samples (Elliott 1971b), parasitic helminths in human hosts (Grafen & Woolhouse 1993), parasites in 48 parasite-host systems (Shaw et al. 1998), parasites in shrimp (Crofton 1971), ticks in mice and tapeworm in bream (Anderson & May 1978), mites on apple leaves (Bliss & Fisher 1953), plankton in water samples (Cassie 1962), number of lymph nodes involved in various cancers (Kendal 2007), parasites in dogs (Azlaf et al. 2007), green leafhoppers on bean plants (Moura et al. 2007) and nematodes in Spain (Campos-Herrera et al. 2007). It is a two parameter distribution – λ is the mean and k is the inverse measure of aggregation. Aggregation refers to the clumping or clustering of counts (insects, bacteria, humans etc.) in space or over time – e.g. outbreaks over time (seasonal trends) or bacterial counts in humans (most having low counts, with a few having high counts). High values of k indicate low aggregation – when k approaches infinity, the distribution becomes the Poisson distribution (see Fig. 8.7) – and low values of k indicate high aggregation – the distribution tends towards the logarithmic distribution when k approaches zero.

Figure 8.7: Relationship between the binomial family distributions
The relationships between the distributions in the Binomial Family, from Elliot (Elliott 1971b). The inverse measure of aggregation is k, $p = \mu/k$ and $q = 1 - p$, where μ is the arithmetic mean of the population. The variance is indicated as s^2 and $\bar{x}$ is the population mean.



The probability equation for the distribution is:

$$p(x) = \left(\frac{(k+x-1)!}{x!(k-1)!}\right)\left(1+\frac{\mu}{k}\right)^{-k}\left(\frac{\mu}{\mu+k}\right)^{x}$$

For the purposes of this chapter, the appropriate form of the distribution is the zero truncated version, in which the above equation is divided by 1-p(0) (Sampford 1955):

$$p(x) = \left(\frac{(k+x-1)!}{x!(k-1)!}\right)\left(1-\frac{\mu}{\mu+k}\right)^{k}\left(\frac{\mu}{\mu+k}\right)^{x}\frac{1}{1-\left(1+\frac{\mu}{k}\right)^{-k}}$$

Truncated negative binomial models have been used to describe family size (Hamdan 1975), the fish caught by recreational anglers (O'Neill & Faddy 2003) and microfilarial parasite loads in humans (Das et al. 1990).

There are several reasons underlying the selection of the truncated negative binomial distribution as one possible distribution for the probability of actual event occurrence. Firstly, it is known that the observed event occurrence distribution in Scotland is strongly aggregated: there are few large events/outbreaks and many events of size one or two (official data set, HPS). Thus the probability of actual event occurrence distribution is also likely to be aggregated, and so a distribution that allows for a high degree of aggregation is needed. Relatedly, with event sizes as least as high as 118, but as many as 87% of events of size 1, the variance is likely to be much larger than the mean – for reported outbreaks in Scotland from 1994 to 1999 the ratio was 120.4 to 1 (Woolhouse et al. 2001) and the approximate ratio for all detected cases between 1996 and 2004 was 37 to 1. In addition, the negative binomial, whether or not it is truncated, has been used to describe many biological distributions (see above), and the data does not have to adhere rigidly to the criteria of independent probability of occurrence for each event (Elliott 1971b). The relaxation of this criteria is important because HPS considers cases within household outbreaks to be sporadic cases (Locking et al. 2003a), whilst for this chapter household outbreaks are defined as events. Thus, since the data for this study comes from HPS, there are undoubtedly a number of household outbreaks within the events of size one. As a result, not all of the events in the distribution are independent in

terms of occurrence, nor is it truly accurate that the occurrence of cases and outbreaks of an infection such as *E. coli* O157 are truly independent. Studies have posited that factors such as the ratio of cattle to human population (Innocent et al. 2006) and contact with animals or animal faeces (Locking et al. 2001), and that there are geographic clusters of sporadic cases in Scotland (Innocent et al. 2006), thus unrelated outbreaks or sporadic cases may have increased probabilities of occurrence due sharing a risk factor such as geographic region.

Logarithmic distribution

The logarithmic distribution is the limiting case of the negative binomial distribution when k approaches zero (Fig. 8.7), and thus more highly aggregated than the negative binomial distribution is another potential choice for the distribution of actual event occurrence. In this distribution, the probability equation is:

$$p(x) = \frac{-[\ln(1-\lambda)]^{-1}\lambda^x}{x}$$ where x= 1,2,3… and ln is the natural logarithm

The distribution does not have to be truncated to be applicable to the study data, for which all values are greater than 0, because x must be a minimum of one (it is not possible to divide a number by zero) and x also must be an integer. The distribution was derived by Fisher as the limit from the truncated negative binomial distribution when $k \rightarrow 0$ (Johnson et al. 1992), and like the negative binomial does not require strict assumptions of mean equivalence to variance or independence of observations. In addition, Chatfield and colleagues have indicated that the logarithmic distribution "is likely to be a useful approximation to the negative binomial distribution" when the aggregation parameter (k) is less than 0.01 (Chatfield et al. 1966). The logarithmic distribution has been used to describe a number of biological patterns including head lice on hosts, numbers of individuals in British nesting birds (Johnson et al. 1992), as well as being commonly used to describe growth patterns. Truncated logarithmic distributions were used to describe the patterns of *Salmonella* serotypes in pigs (Izsak & Hunter 1992).

8.3.1.2 Probability of case detection

Little information is available about the probability of a case being detected, dependant on the size of the event. It seems intuitive that the probability of detection is a function of event size – e.g. the larger an event a case in part of, the more likely a case within that event is to be detected – the chance of a case in an outbreak the size of Wishaw Outbreak (see 2.4.1.1 for details) going undetected would seem to be much less than that for a single sporadic case. Since it seems unlikely that any single case would have a 100% chance of being detected (even if every infected person had a stool or serum sample tested, the tests do not have 100% sensitivity) or a 100% chance of not being detected, the most appropriate shape for the function would be one that is asymptotic near zero and one.

Since a sigmoidal function is necessary, the Logistic Growth Law function, which has been used to describe growth in populations such as the human population of the United States (Snedecor & Cochran 1971), fish (Chaudhuri 1988) and fungal colonies (Koch 1975), will be used. The equation for the distribution, which has three parameters is:

$$p(\text{detection}) = \frac{A}{1 + B\rho^{-x}}$$

where A defines the upper asymptote and B and ρ are shape parameters

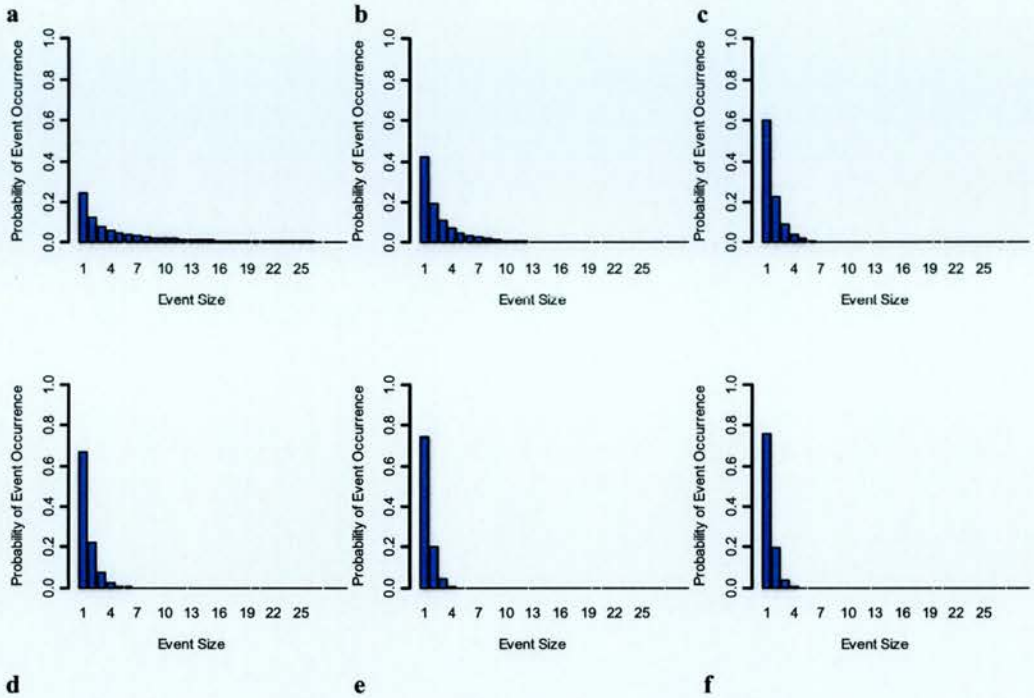
8.3.2 Exploration of the distribution of actual event occurrence

The parameter values for each actual event occurrence distribution must be explored to determine how the values affect distribution shapes, and which value(s) may be appropriate for the framework.

8.3.2.1 Truncated negative binomial distribution

With the truncated negative binomial distribution, there are two parameters which shape the curve, mu (μ) which is the arithmetic mean and k, which is the inverse measure of aggregation. When μ is held constant, low values of k produce a curve which is highly dispersed (Fig. 8.8a), meaning that there is a very long right tail. As the value for k increases, the curve becomes steeper, with an increasing p(x) at event size=1.

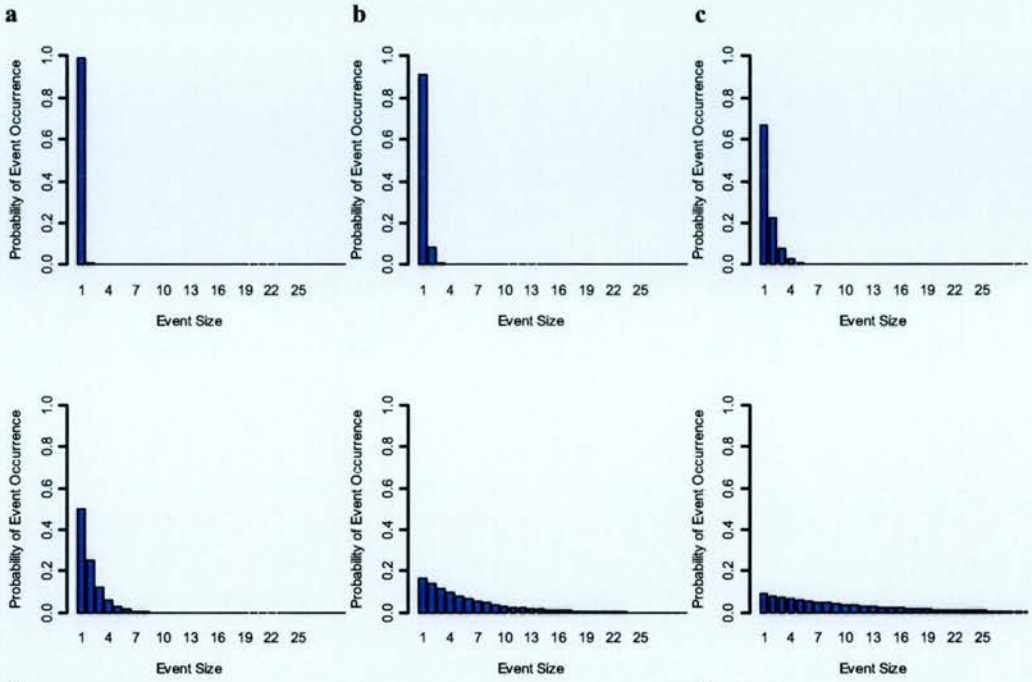
Figure 8.8 a-f: Truncated negative binomial distributions – varying k
A series of plots showing the effect of variation in the parameter k on the truncated negative binomial distribution for the probability of event occurrence. The value of μ is constrained at 0.5, with k increasing from a to f: 0.01, 0.1, 0.5, 1, 5 and 10. The table shows the counts for event sizes up to size 20 when the probabilities in the distribution are multiplied by 1927, the number of events in Scotland between 1996 and 2004.



	p(x)	Size of Event										
	μ =0.5, k	1	2	3	4	5	6	7	8	9	10	11-20
a	k=0.1	474	235	154	114	89	73	61	53	46	41	249
b	k=0.01	818	375	219	141	97	68	50	37	28	21	66
c	k=0.5	1163	436	182	80	36	16	8	4	2	<1	<1
d	k=1	1285	428	143	48	16	5	2	<1	<1	<1	<1
e	k=5	1435	391	83	15	2	<1	<1	<1	<1	<1	<1
f	k=10	1459	382	73	11	1	<1	<1	<1	<1	<1	<1

If k is held constant, skewness increases as the value of μ decreases. This can be observed in Figure 8.9 a-f, where k = 1, and the values of μ increase from 0.01 to 10. When μ=0.01, the probability of actual event occurrence is nearly 1 for events of size one, but is close to zero for events of two or more cases. However, at μ =10, the probability for events of size one is approximately 0.1, and the curve declines at a very slow rate.

Figure 8.9 a-f: Truncated negative binomial distribution – varying μ
Bar plots of a series of truncated negative binomial distributions demonstrating the effect of variations in the parameter μ. K is fixed at 1, with μ increasing from a to f (0.01, 0.1, 0.5, 1, 5, 10). The table shows the counts for event sizes up to size 20 when the probabilities in the distribution are multiplied by 1927, the number of events in Scotland between 1996 and 2004.

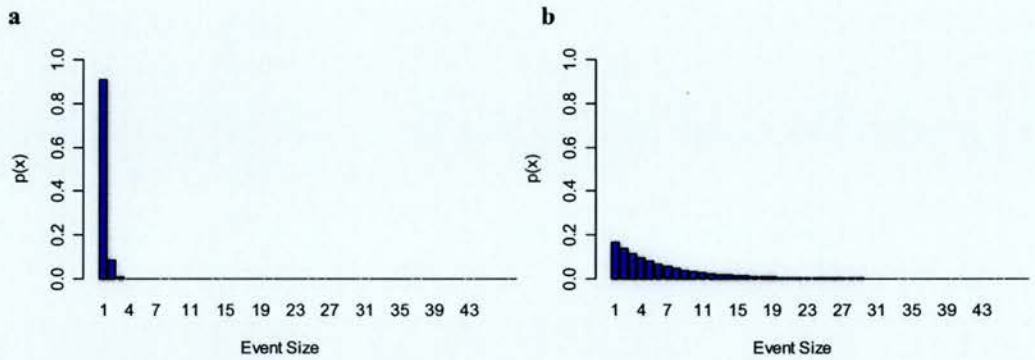


	p(x)	Size of Event										
	k=1	1	2	3	4	5	6	7	8	9	10	11-20
a	μ =0.01	1908	19	<1	<1	<1	<1	<1	<1	<1	<1	<1
b	μ =0.1	1752	159	14	1	<1	<1	<1	<1	<1	<1	<1
c	μ =0.5	1285	428	143	48	16	5	2	<1	<1	<1	<1
d	μ =1	964	482	241	120	60	30	15	8	4	2	<1
e	μ =5	321	268	223	186	155	129	108	90	75	62	260
f	μ =10	175	159	145	132	120	109	99	90	82	74	457

Since there have been events as large as 118 cases reported in Scotland, the assumption is that the probability of actual event occurrence would need to have a long right tail in order to allow for the occurrence of larger actual event sizes. Low values of μ and/or high values of k result in distributions which produce high enough counts (counts being used instead of probabilities in order to permit easier comparison to the data in Scotland from 1996 to 2004) for events of size 1, but these distributions underestimate event counts for large event sizes, particularly over size 10. For instance, when μ =0.1 and k=1, the framework predicts 1752 events of size

one, but then predicts almost no events larger than size 4 (Table 8.1 Line 1, Fig. 8.10a). In contrast, when there are high values of μ and/or low values of k, the tail of the distribution is sufficiently long, but the counts for events of size 1 are considerably underestimated and the counts for events for larger events are overestimated. When, for example, μ=5 and k=1, the framework results in only 321 events of size one, but more than 200 events of greater than size 11(Fig. 8.10b, Table 8.1 Line 3). If μ=1.2, which is the approximate mean event size in the observed population, values of k less than 0.2 are required to produce a long enough tail

Figure 8.10 a-b: Comparison of truncated negative binomial distributions for the probability of event occurrence
Bar plots showing the truncated negative binomial distributions for the probability of event occurrence when k=1 and (a)mu=0.1 and (b) mu=5. The plots are limited at event size=50.



These findings suggest that a negative binomial distribution is not likely to be an appropriate distribution for the distribution of actual event occurrence, but it is still possible that the addition of the function of event detection to the framework may sufficiently adjust the counts to permit a long tail with a high count for events of size one.

Table 8.1: Event counts, by size, derived from frameworks of detected event size proportions

Number of events, by event size, predicted by detected event size proportion frameworks. Counts are calculated by multiplying the proportion of events ≥ size 1 by the total number of events observed in Scotland between 1996 and 2004. Either the truncated negative binomial (μ and k) or logarithmic distribution (c) is used for the probability of actual event occurrence. Where 100% case detection was assumed, no parameter values are given for the Logistic Growth Law function of the probability of case detection. Otherwise Logistic Growth Law parameter values are as given (A, B, Rho). The observed count for each event size is shown in blue.

Line #	p(x)	p(detection)	Size of Event												
	μ, k or c	A, B and Rho	1	2	3	4	5	6	7	8	9	10	11-20	21-50	51-130
	Observed	----	1852	21	19	7	9	4	1	2	1	0	5	4	2
1	μ = .1, k=1	----	1752	159	14	1	<1	<1	<1	<1	<1	<1	<1	<1	<1
2	μ = .1, k=1	0.98, 10, 1.7	1855	67	~4	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1
3	μ = 5, k=1	----	321	268	223	186	155	129	108	90	75	62	260	49	<1
4	μ = 5, k=1	0.98, 10, 1.05	1081	421	189	94	51	30	18	12	8	5	15	2	1
5	μ = 5, k=1	0.98, 10000, 10	171	126	96	109	178	203	176	147	122	101	419	77	2
6	μ = 1.2, k=0.3	0.98, 500, 1.15	1723	133	34	14	7	4	3	2	1	1	~4	<1	<1
7	c = 0.1	----	1829	91	6	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1
8	c = 0.1	0.98, 10, 1.7	1887	38	2	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1
9	c = 0.9	----	753	339	203	137	99	74	57	45	36	29	121	32	2
10	c = 0.9	0.98, 10, 1.05	1259	337	136	68	38	24	16	11	8	6	19	4	<1
11	c = 0.82	0.98, 500, 1.15	1728	125	33	14	8	5	3	2	2	1	4	~1	<1
12	c= 0.66	0.98, 500, 1.15	1906	20	1	<1	<1	<1	<1	<1	<1	<1	<1	<1	<1
13	c= 0.8	0.98, 500, 1.15	1780	100	23	9	5	3	2	1	1	<1	2	<1	<1

8.3.2.2 Logarithmic distribution

In the logarithmic distribution, there is single parameter, c (also referred to as θ in some texts), which is limited to values between zero and one (Johnson et al. 1992).

In order to explore how variation in c affects the distribution of the probability of event occurrence (p(x), frameworks were created for distributions with values of c ranging from 0.01 to 0.99 (Fig. 8.11). As shown in Figure 8.11a-g, when there is perfect event detection, as c increases from 0.01 to 0.99, the distribution of event size proportions shifts to the right, with dispersion increasing and a greater proportion of events of larger sizes.

An estimate for c can be obtained from the population mean (μ), using the equation below (Johnson et al. 1992).

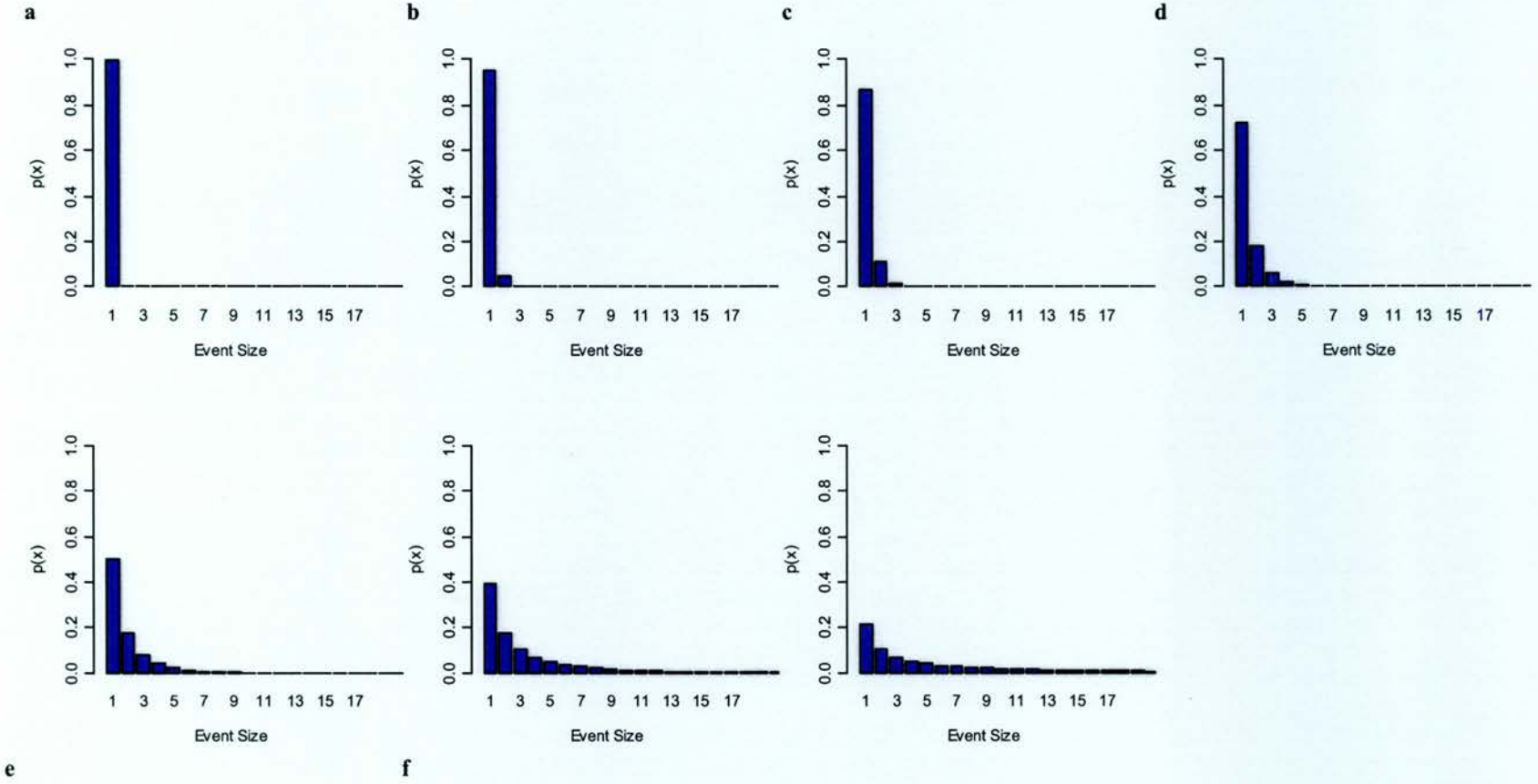
$$\mu = (-1 / \ln (1-c) * c) / 1 - c \text{ , where c is the shape parameter and } 0<c<1$$

The mean for observed Scottish *E. coli* O157 events has been estimated as approximately 1.2. Since a value of 0.3 for c gives a mean of 1.202, this would suggest a value of 0.3 as an approximate initial value for c.

Figure 8.11 continued (see next page for rest of figure and figure legend)

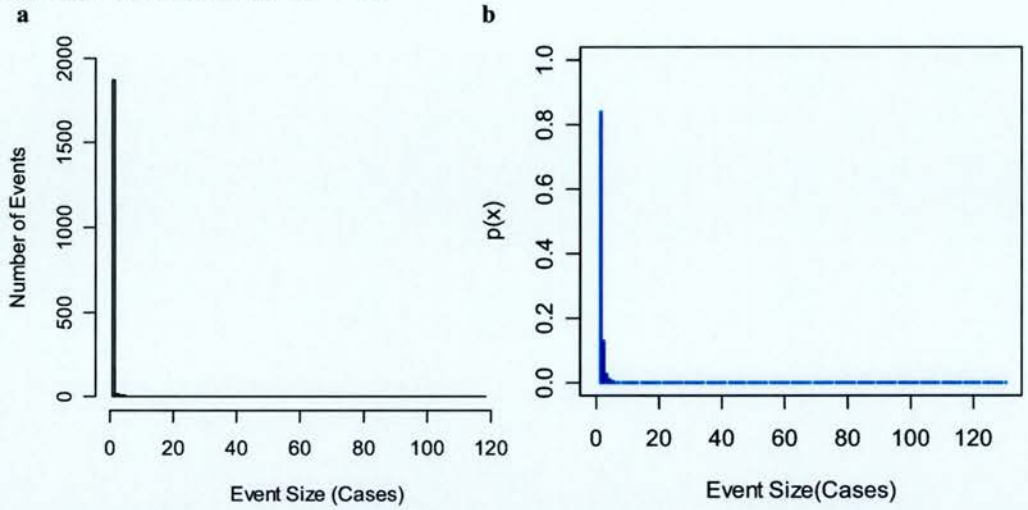
	p(x)	Size of Event										
		1	2	3	4	5	6	7	8	9	10	11-20
a	c=0.01	1917	~9	<1	<1	<1	<1	<1	<1	<1	<1	<1
b	c=0.1	1829	91	6	<1	<1	<1	<1	<1	<1	<1	<1
c	c=0.25	1675	209	35	6	1	<1	<1	<1	<1	<1	<1
d	c=0.5	1390	348	116	43	17	7	3	1	<1	<1	<1
e	c=0.75	1043	391	195	110	66	41	27	17	12	8	~16
f	c=0.9	753	339	203	137	99	74	57	45	36	29	~121
e	c=0.99	427	211	139	103	82	68	57	50	44	39	~249

Figure 8.11 a-b: Models showing the effect of variation in c on the logarithmic distribution of the probability of actual event occurrence
Plot showing the effect of a variation in c on the distribution of the probability of actual event occurrence, assuming 100% case detection. Values for c are (a) 0.01, (b) 0.1, (c) 0.25, (d) 0.5, (e) 0.75, (f) 0.9 and (g) 0.99. The table shows the counts for event sizes up to size 20 when the probabilities in the distribution are multiplied by 1927, the number of events in Scotland between 1996 and 2004.



However, as the logarithmic distribution is the limiting form of the negative binomial distribution when k approaches zero, similar issues are encountered as with the truncated negative binomial distribution when trying to approximate the distribution of actual event occurrence. When the value of c is small (less than 0.5, and so including 0.3; Table 8.1 Line 7), the distribution is highly skewed towards small event sizes. Thus, whilst the proportion/count for events of size=1 is close(r) to the observed proportion, for value of c up to 0.5, the count for size=2 is overestimated and the counts for events with more than 2 cases are all underestimated (Figure 8.12a vs. b).

Figure 8.12 a-b: Event size histogram: Scotland 1996 – 2004 and logarithmic distribution
(a) A histogram of the sizes of *E. coli* O157 events in Scotland from 1996 to 2004 and (b) logarithmic distribution when c = 0.3



Values of c equal to or above 0.9 (Table 8.1 Line 9 and Fig. 8.11g) predict high enough counts for events over size 10, but as with the truncated negative binomial distribution, models which predict sufficient counts for events with more than 20 cases underestimate the count for single case events and overestimate the counts for events between 2 and 20 cases. However, when models that predict sufficient counts for events > size 20 are considered, the counts predicted when the logarithmic distribution is used are closer (by values of one or two) to the observed counts than with the truncated negative binomial.

Thus it would appear both that the logarithmic distribution is probably a better fit for the framework than the truncated negative binomial distribution and that high value

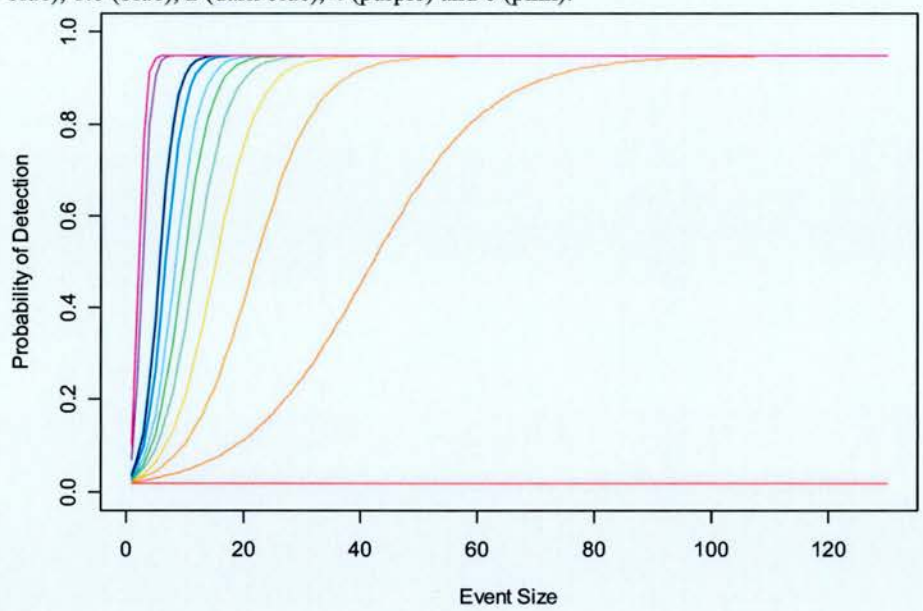
of c is most likely to be appropriate for the framework. The next step will be to determine the appropriateness of the distributions in the full framework.

8.3.3 Exploration of a function for the probability of case detection

8.3.3.1 Logistic Growth Law

For the logistic growth law, the parameter A indicates the value of the upper asymptote. B and Rho (ρ) are shape parameters, with B influencing the shape of the first part of the curve and ρ, the second part of the curve.

Figure 8.13: Logistic Growth Law functions for the probability of case detection– varying Rho
A series of Logistic Growth Law distributions for the probability of case detection demonstrating the effect of variation in the ρ parameter. For all curves, A is 0.95 and B=50. The values for ρ are 1 (red), 1.1 (dark orange), 1.2 (light orange), 1.3 (yellow), 1.4 (light green), 1.5 (green), 1.6 (light blue), 1.8 (blue), 2 (dark blue), 4 (purple) and 6 (pink).



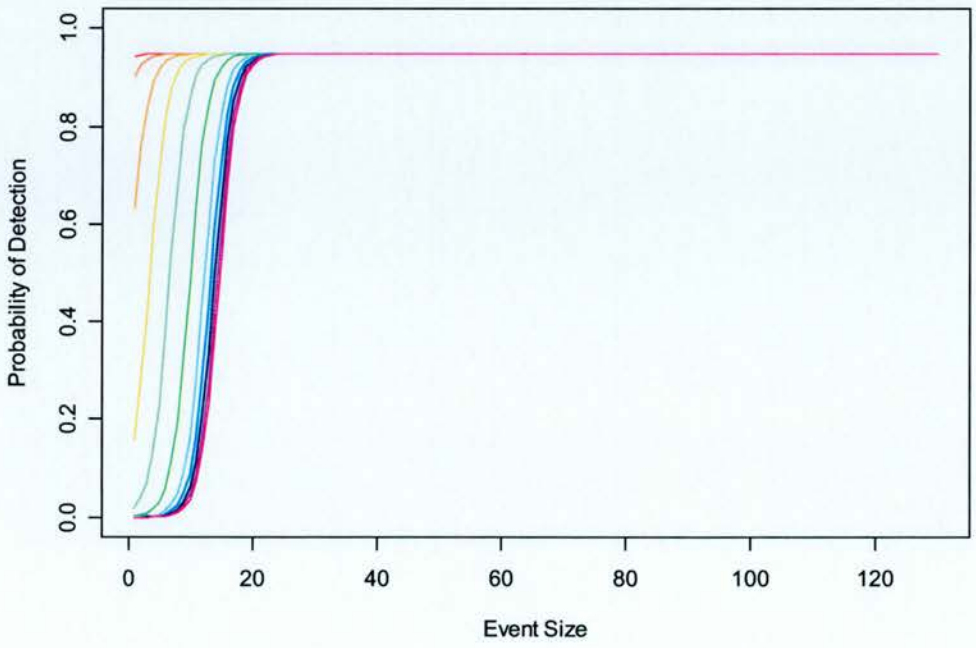
When ρ=1, the distribution is a straight line, with all values for event detection equivalent to 1/B. When ρ is less than one, the curve is decreasing and when Rho is greater the one, the curve is increasing. As the value of ρ increases above one, the curve becomes steeper (Fig. 8.13) and the values of y (probability of case detection) for each x (event size) increase. Thus, the higher the value of ρ, the lower the value of x at which the curve approaches the upper asymptote (at A), which in terms of this model means that the higher the value of ρ, the smaller the event size at which the maximum probability of case detection is reached. This trend can be visualised in Figure 8.13 where increasing values of ρ are represented as the colours proceed from

red, orange, yellow... through to pink. Whilst the red curve is flat (ρ =1), the pink curve is almost vertical and the asymptote is reached at a very low detected event size (approximately event size = 7, as compared to event size = 60, for the dark orange curve for ρ=1.1).

In contrast, as the value for parameter B increases, the value of y at each x decreases until the curve approaches the upper asymptote. For this framework, it means that as B increases, the probability of case detection for each event size decreases until the curve approaches the asymptote. Thus, the higher the value of B, the closer the probability of detection for events of size one is to zero. At very high values of B ($\geq$1000), the probabilities for small event sizes are similar, flattening out the curve as it approaches the lower asymptote. The influence of parameter B can be seen in Figure 8.14, where the values for the probability of case detection for events of size 1 do not begin to approach zero until B is 100 (first green curve).

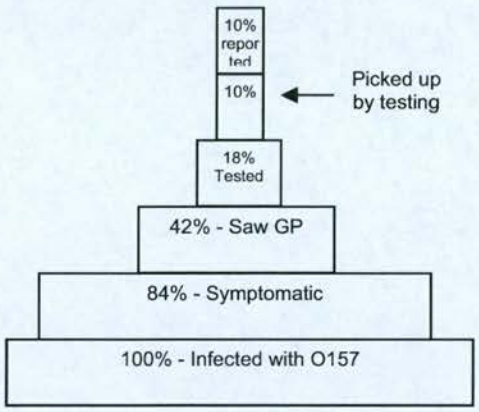
Figure 8.14: Logistic Growth Law distributions for the probability of case detection– varying B

Distributions for the probability of case detection demonstrating the effect of variation in the B parameter. For all curves, A=0.95 and Rho=2.0. B is 0.01 (red), 0.1 (dark orange), 1 (orange), 10 (yellow), 100 (light green), 1000 (green), 5000 (light blue), 10,000 (blue), 15,000 (dark blue), 20,000 (purple) and 25,000 (pink).



Estimates for *E. coli* O157 case detection have been calculated and discussed in a number of papers (Brewster et al. 1994; Feldman et al. 2002; Locking et al. 2004; Marsh et al. 1992; Michel et al. 2000; Sharp et al. 1994b), with reporting triangles frequently used to show why and where cases are not being detected. The use of reporting triangles and the published studies on prevalence estimation have been discussed in Chapter 1 and Chapter 8.1. Studies from Canada, the United Kingdom and the United States suggest that between 0.5% and 10% of cases are detected by current surveillance systems. However, no study on under-reporting has been published for data in Scotland. In order to estimate a possible overall p(detection) for Scotland, data from the above papers was used to construct a reporting triangle for Scotland (Figure 8.15), which suggests that the overall probability of detection might be around 10%.

Figure 8.15: Reporting triangle for *E. coli* O157 in Scotland
Reporting triangle for Scotland



However, neither this reporting triangle, nor any other study has addressed case detection as a function of event size. Therefore possible curve shapes for the function of the probability of case detection will be hypothesized based on what is know about methods of detection and rates of detection.

The parameter A indicates the upper limit of the distribution, and for initial models, A will be fixed at 0.98, a high value, but short of 1, which would indicate perfect detection. It is assumed that perfect detection would never be achieved due to a number of factors including the presence of asymptomatic cases (Griffin & Tauxe 1991), imperfect sensitivity of laboratory tests for infection (De Boer & Heuvelink

2000; Noël & Boedeker 1997) and the decreasing effectiveness of faecal testing of patients with increasing time after onset of symptoms (Tarr et al. 1990). Equally it is assumed that every case has at least a small probability of detection – for instance sporadic cases can be picked up as part of large outbreak investigations (Cowden et al. 2001) – and thus the curve for the case detection distribution should not ever reach zero.

Less, however, can be assumed about the behaviour of the curve between the minimal and maximal values. One possibility is that the probability of detection for sporadic cases or cases in very small events (size 2 – 3) is very low, but increases quickly after reaching events of size four of five cases, and then levels off at the upper asymptote. This scenario might occur if clusters with more than three cases are more likely to come to the attention of public health investigators because, for instance, the follow up of each confirmed case as a part of the Scottish enhanced surveillance program in Scotland (Locking et al. 2003a)) increases the chance that previously unrecognized or unlinked cases will be connected to events. However eventually, when events are large enough, the increase in detection probability with each additional case becomes negligible. However, there could be a much gentler curve, in which cases in very small events have a low probability of detection, and the probability of detection increases steadily as the event size increases. Both types of curves will be explored in the framework by using sigmoidal Logistic Growth Law functions for the probability of case detection.

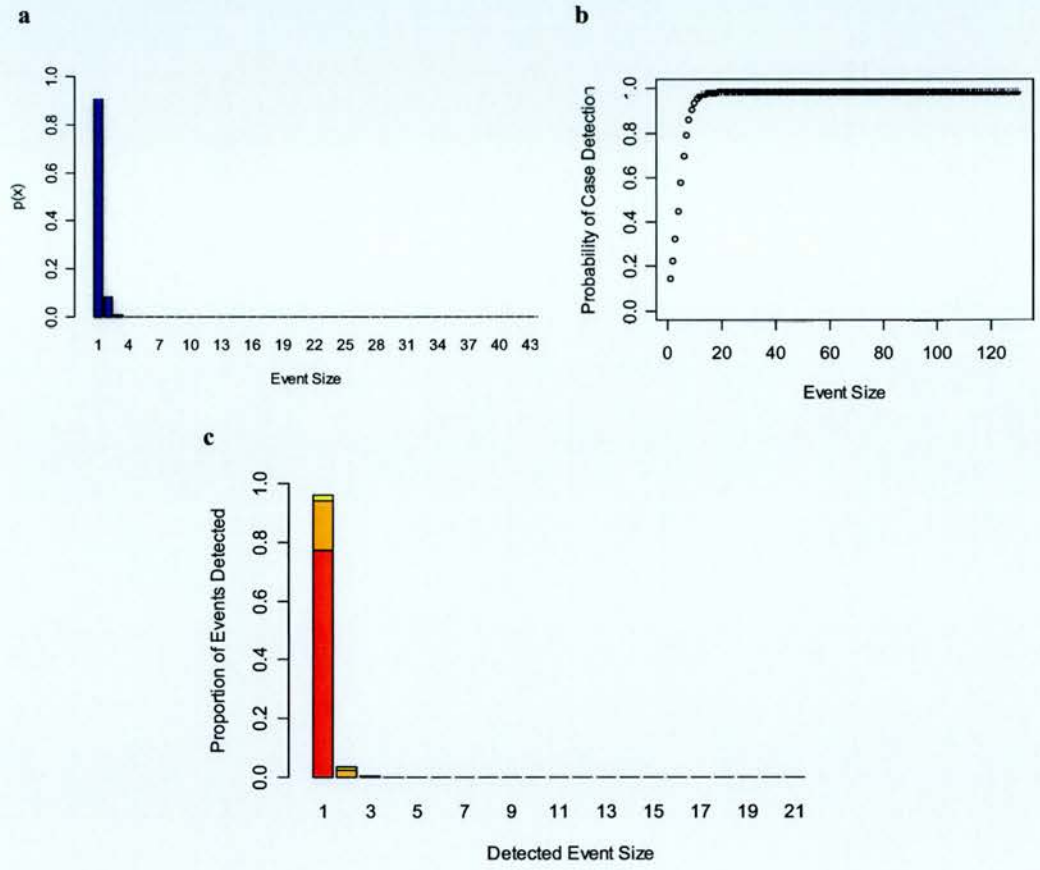
8.3.4 The complete framework

8.3.4.1 Truncated negative binomial distribution and Logistic Growth Law

When the Logistic Growth Law function for the distribution of case detection - irrespective of the parameters– is added to the framework, it is still not possible to find parameters for the framework for both the extremely large proportion of single case events and the much smaller, slowly decreasing proportions of events larger than size 1. When the right tail of the event occurrence distribution is very short as in Fig. 8.16a, events of large sizes (usually >10) cannot occur. Thus even if the distribution of case detection increases sharply (Fig. 8.16b) and is near perfect for larger event sizes, the proportions of large events are underestimated in the final

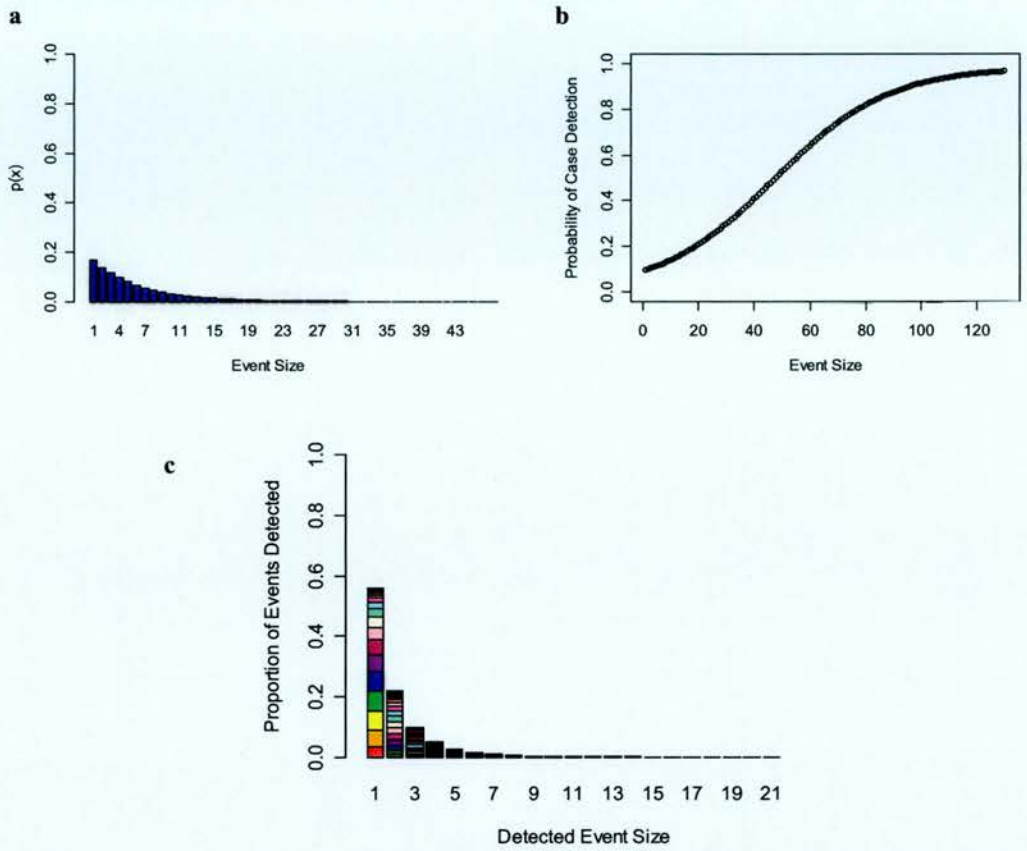
framework(Fig. 8.16c; Table 8.1 Line 2) (Counts for all frameworks are shown in Table 8.1).

Figure 8.16 a-c: Components of a framework that underestimates large event sizes
The (a) truncated negative binomial distribution of event occurrence, (b) logistic growth law distribution of case detection and (c) final framework of the distribution of detected case size proportions where mu=0.1, k=1, A=0.98, B=10 and ρ=1.7. The distribution of case detection decreases too sharply, and thus in the final framework, the proportions of event sizes greater than six are, in general, underestimated.



In contrast, if distribution of event occurrence is skewed (Fig. 8.17a) and thus the right tail is very long, even with a more gradual increase in detection probabilities (Fig. 8.17b), the proportions for events between approximately size 2 and 20 are overestimated (Figure 8.17c; Table 8.1 Line 4).

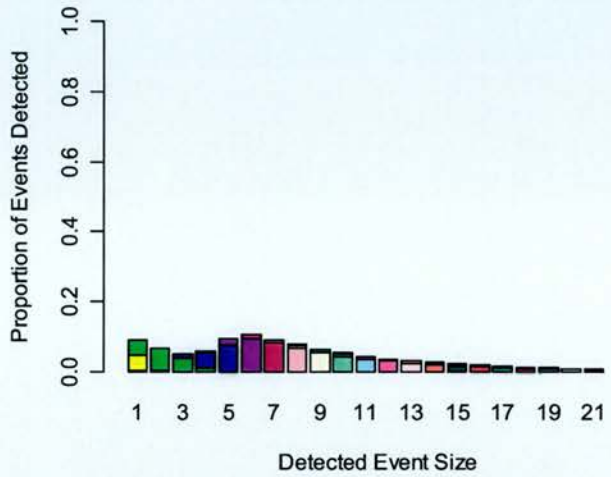
Figure 8.17 a-c: Components of a framework that overestimates small event sizes
The (a) truncated negative binomial distribution of event occurrence, (b) logistic growth law distribution of case detection and (c) final framework of the distribution of detected case size proportions where mu=5, k=1, A=0.98, B=10 and ρ=1.05. The distribution of case detection decreases too gradually, and thus in the final framework, the proportions of event sizes greater than one are overestimated.



The above frameworks using the framework either underestimate the proportions of large event sizes or overestimate the proportions of small event sizes. If it is accepted that for this reason the events of size one cannot be properly estimated by the framework, and so only the proportions for the events of size or 2 or larger are examined, an appropriate framework still cannot be found. The creation of a framework which reflects the very low proportions for events greater than size 1 is only possible by using a value for the case detection function parameter B which is so large that small events would have an extremely low probability of case detection (<0.001). Even the most extreme estimates of underreporting are only around 0.02, so the framework is highly unlikely to be appropriate. In addition, extreme values of

B produce peaks and valleys in the detected event size proportion distribution, in addition to low proportion values for events of small sizes (Fig. 8.18, Table 8.1 Line 5).

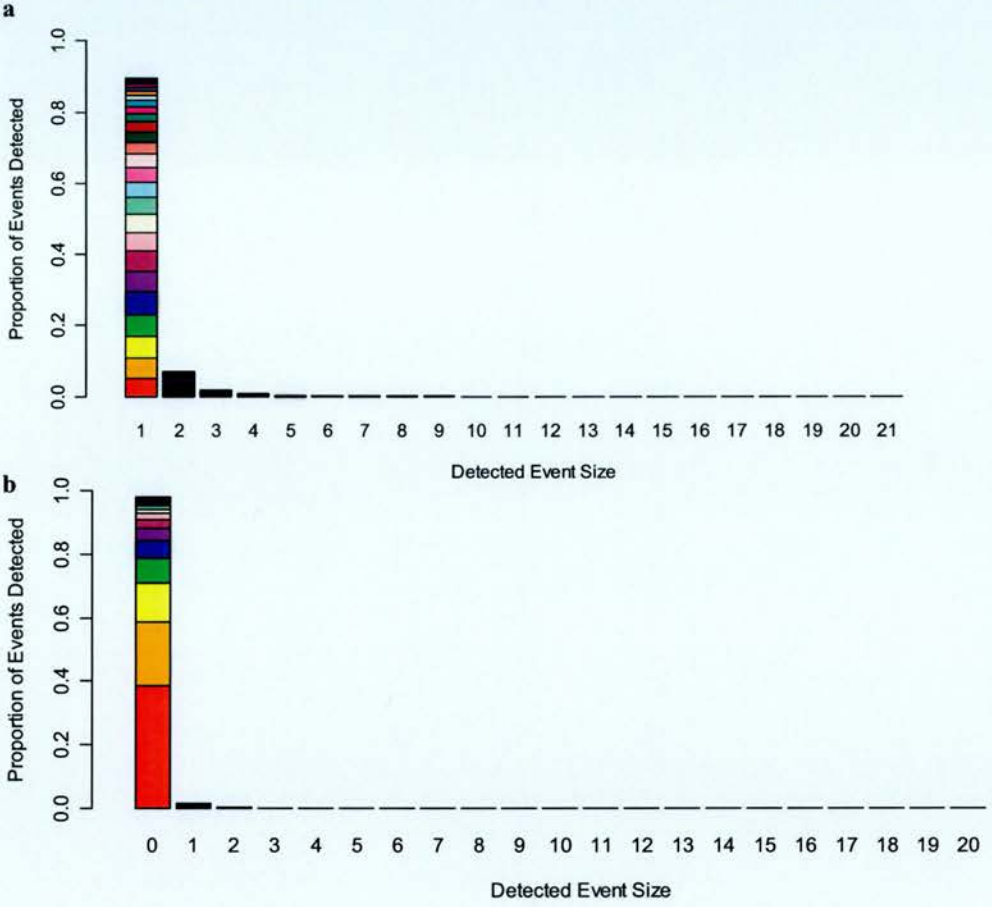
Figure 8.18: Inappropriate distribution of detected event size proportions
A distribution of detected event size proportion when mu=5, k=1, A=0.98, B=10,000 and ρ=10, demonstrating the inappropriateness of frameworks where extremely high values of B are used in an attempt to make the proportions of events sized 2 or greater sufficiently low.



Neither the peaks and valleys or small proportions values are consistent with the observed distribution (Fig. 8.5). Given the difficulties, as shown above, in constructing a framework that predicts the correct proportions of events for small event sizes and the issue regarding household outbreaks, the final approach is to construct a framework which allows for the fact that the detected distribution underestimates the proportion of events at size=1 and overestimates the proportions of events of size 2 to 5. The best fit is then found after mu is fixed at 1.2, which is the approximate mean of the observed event population, and the values for the other parameters are varied. When k=0.3 (a moderately overdispersed distribution for event occurrence), and A=0.98, B=500 and ρ=1.15, (a very gradual distribution for case detection), the proportions of detected events that are size 1 and size 3 to 20 are fairly accurately estimated by the framework (Fig. 8.19a, Table 8.1 Line 6). However, the framework highly overestimates the proportion of events that are size 2, and underestimates proportions of events greater than size 20. These over/underestimates may be plausible because a number of events of size 2 or greater are actually counted as sporadic cases. Therefore, the observed counts most likely overestimate the number of events of size one and underestimate the number of small

multi-case events. The underestimation of the number of very large outbreak (>20 cases) may reflect the fact that the large events are outliers or clusters of smaller events.

Figure 8.19: Distribution of detected event size proportions and predicted counts
(a) The distribution of detected event size proportions (event size ≥1) when mu=1.2, k=0.3, A=0.98, B=500 and Rho=1.15. (b) The complete distribution of detected event size proportions showing that 98.2% of events are detected at size zero – in other words, they are not detected.

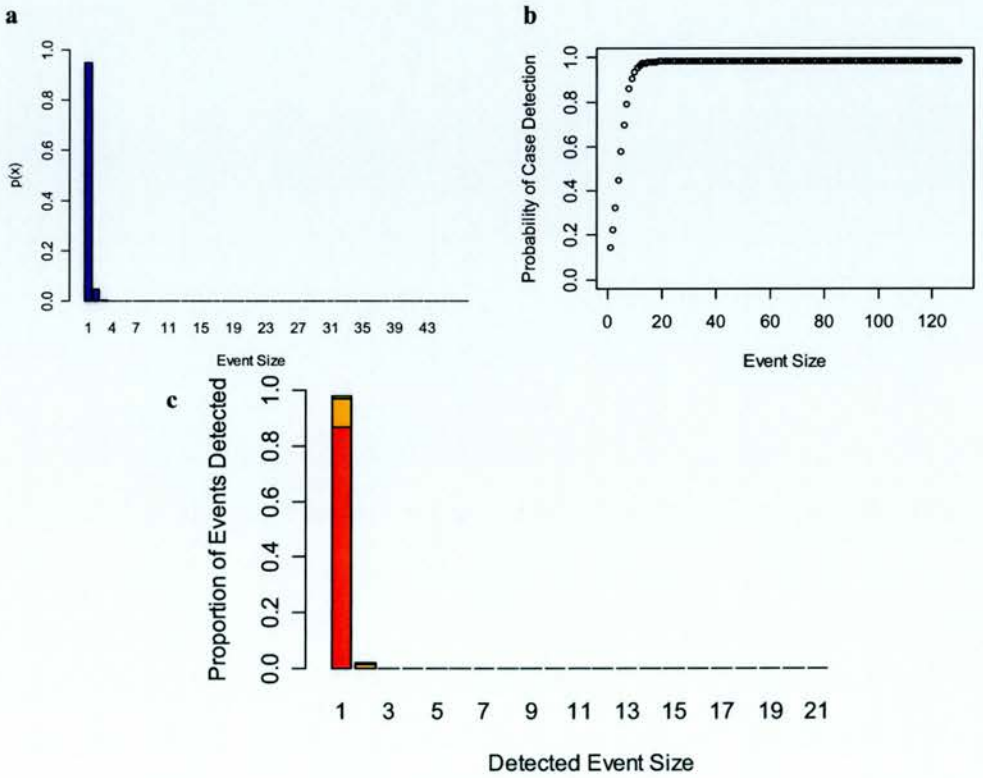


The framework suggest that 1.7% of events are detected (Figure 7.19b), which means than more than 98% of events are not detected. Since there were 1927 events in Scotland between 1996 and 2004, an underreporting rate of 98% suggests that there were approximately 107,000 events. Further calculation based on the proportions of the undetected events (size=0) indicated to be actual events of size 1 or greater, indicates that there were over 330,800 cases in Scotland. Thus the overall case detection rate was ~0.7%. The next step is to determine whether the logarithmic distribution for the probability of actual event occurrence produces a more accurate framework.

8.3.4.2 Logarithmic distribution and Logistic Growth Law

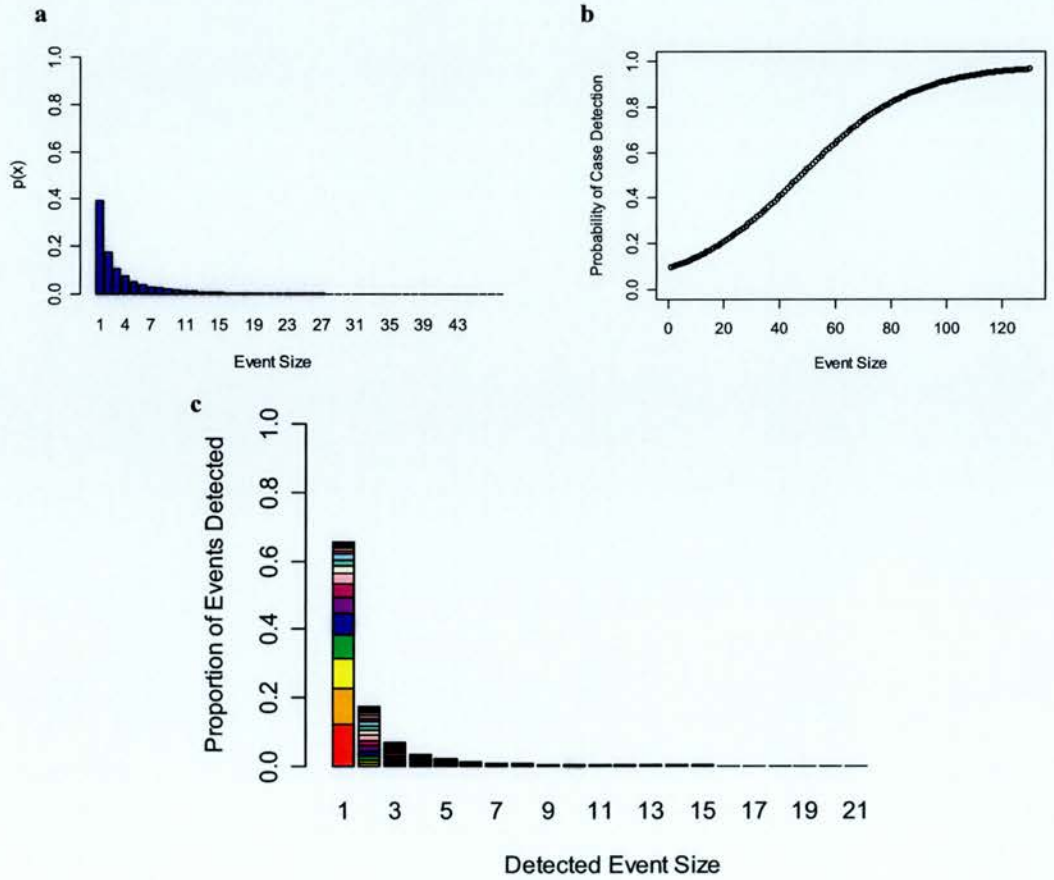
Similar issues as with the truncated negative binomial distribution arise when the Logistic Growth Law function for the probability of case detection is added to the framework. Where the value of c is low and thus the proportions of detected events are reasonably accurate for events of size one or two, the distribution of actual event occurrence declines too steeply and thus there is a very small probability of events over size three occurring. Even if the Logistic Growth Law function parameters are adjusted so that there is a very steep increase in the rate of case detection, the proportions of large events are still greatly underestimated. For instance, if the rate of case detection is initially assumed to be 100%, when c=0.1, there are 1829 events of size one and 91 of size two, but less than one event each for sizes greater than three (Table 8.1 Line 7, Fig. 8.20c). If a very steep function for the probability of case detection is added to the framework (Fig. 8.20b, Table 8.1 Line 8), there are 1887 events of size 1 and 38 of size two, but even smaller probabilities for events > size 2.

Figure 8.20: Components of a framework that underestimate large event sizes
The (a) logarithmic distribution of event occurrence, (b) logistic growth law distribution of case detection and (c) final framework of the distribution of detected case size proportions where c=0.1, A=0.98, B=10 and ρ=1.7. The distribution of case detection decreases too sharply, and thus in the final framework, the proportions of event sizes greater than three are, in general, underestimated.



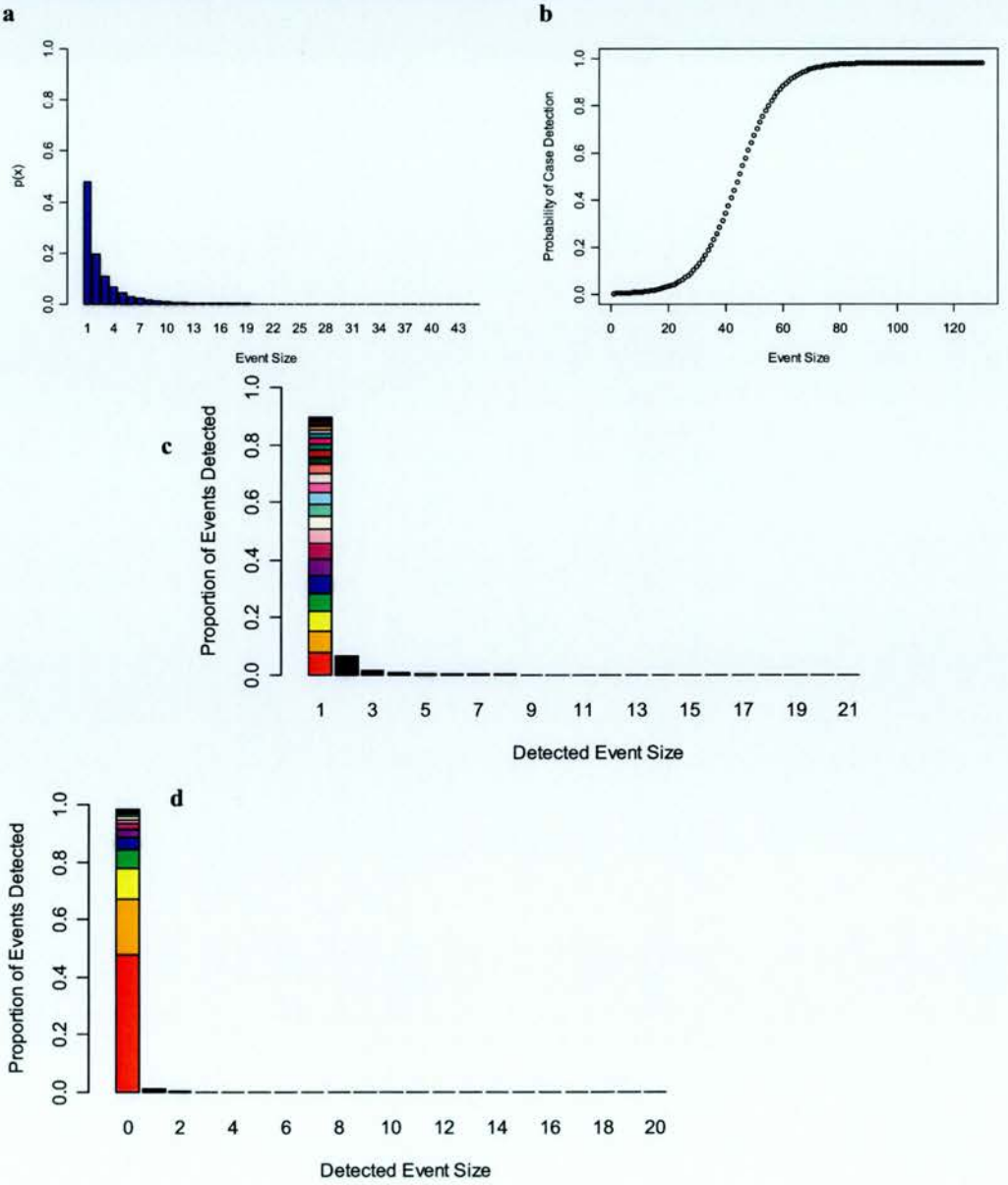
In contrast, where the value for c is high and perfect detection is assumed, sufficiently high counts/proportions for large event sizes can be estimated using the framework, but the counts for events size 2 – 49 are overestimated and the count for single case events is considerably underestimated (Table 8.1 Line 9). If a very gradual Logistic Growth Law Function is added to the framework to counterbalance the overdispersion, the underestimation of single cases events and overestimate of larger outbreaks remain. As an example, when c=0.9, the distribution of actual event occurrence is right skewed (Fig. 8.21a), but even with a very gradual function for the probability of case detection (Fig. 8.21b), the final framework still predicts too few size 1 events and too many events of size 2 or larger (Fig. 8.21c, Table 8.1 Line 10).

Figure 8.21a-c: Components of a framework that overestimates small event sizes
The (a) logarithmic distribution of event occurrence, (b) logistic growth law distribution of case detection and (c) final framework of the distribution of detected case size proportions where c=0.9, A=0.98, B=10 and ρ=1.05. The distribution of case detection decreases too gradually, and thus in the final framework, the proportions of event sizes greater than one are overestimated.



Again, considering the issue of household outbreaks, the most appropriate and final approach will be to construct a framework which has fewer single case events and more small (size 2 to 5) events than the observed distribution. The most accurate framework results from a high value of c(0.82) (Fig. 8.22a) and a steeply increasing logistic growth law function for case detection (Fig. 8.22b).

Figure 8.22 a-d: Distribution of detected event size proportions – best complete framework
The distributions of (a) actual event occurrence, (b) case detection and (c) detected event size proportions (event size ≥1) when mu=c=0.82, A=0.98, B=500 and Rho=1.15. Also, (d) the complete distribution of detected event size proportions showing that 98.6% of events are detected at size zero – in other words, they are not detected.



This framework underestimates the number or proportion of single case events, and overestimates the number of events with two to four cases (Fig. 8.22c, Table 7.1 Line 11), but this is expected because some of the events of size =1 in the observed distribution are likely to actually be events of size 2 to 5, given that most households do not have more than 5 persons). The framework indicates that less than 2% of events are detected (Fig. 8.22d). This framework appears to be an adequate fit, and thus may be a potential framework for the distribution of detected event size proportions. This is the final potential framework that will be presented, and the issues which must be considered in assessing the appropriateness of this framework will be addressed in the next section (Section 8.4).

8.4 Discussion

8.4.1 Introduction to the framework

A number of studies have shown that the prevalence of *E. coli* O157 infection is significantly underestimated by both active and passive surveillance systems (Handysides 1999; Hedberg et al. 1997; Thomas et al. 2006). Studies in Canada (Michel et al. 2000; Thomas et al. 2006), the United States (Hedberg et al. 1997; Mead et al. 1999) and England & Wales (Adak et al. 2002), based on surveys of patient, physician and laboratory practices (Bender et al. 2004; Hedberg et al. 1997; Thomas et al. 2006), data from previous studies and/or hospital records (Michel et al. 2000), have suggested that the degree of under-reporting may range from 50% to around 98%. However, these studies are limited by a number of factors including study areas which are not necessarily epidemiologically or demographically representative of the whole country (e.g. (Michel et al. 2000)), and lack of data about patient and physician practices. Furthermore, no study has been published on the magnitude of underreporting within Scotland. Accurate estimations of *E. coli* O157 prevalence in Scotland and other countries are of importance because they would set accurate benchmarks for the improvements needed to national or regional surveillance systems. They would also provide a measure of the magnitude of scientific efforts needed to lower infection rates. An accurate estimate of prevalence could also indicate whether the high reported rates of infection in Scotland are due to high prevalence of infection or high proportion of cases being reported.

Thus, in this chapter the aim has been to construct and test a different type of framework that, if functional, could be developed into a model to more accurately estimate prevalence. The framework depends only upon observed prevalence data both for sporadic cases and outbreaks. The framework is based upon the assumptions that all cases are part of an event – whether it is a single case event or large outbreak – and that each event is either detected or it is not detected. However, whilst an event is detected, not all of the cases in that event may be detected, resulting in the event being observed as smaller than true. As such, there is a distribution of the proportion of detected event sizes which includes events of size zero – events that are not detected. The concept of this framework is that this distribution is a function of the distribution of actual event sizes and the probability of event detection. Thus if the distribution of the proportion of detected event sizes, for which the distribution of events greater than size zero matches that of the observed distribution, the actual prevalence can be calculated from the proportion of cases that are 'detected' as size zero and the number of events seen

8.4.2 Advantages of the framework

The primary advantage of this approach is that it is not based on data from a series of population, laboratory and/or physician studies. Previous prevalence studies have had error attached to them because they have mainly been structured around calculations of how many cases are detected (or not detected) at each step of the reporting process. For instance in one Canadian study (Michel et al. 2000), there were five different variables – each based on a step between infection and detection – in one of the formulas used to estimate under-reporting. Whilst this type of model is epidemiologically and mathematically sound, the studies from which the data is drawn are often prone to bias (Michel et al. 2000; Thomas et al. 2006). This bias results from a number of different factors. For example, some studies or data sources involve only a specific province (Michel et al. 2000) or region (Adak et al. 2002), or only areas with high rates of infection (Hedberg et al. 1997), whilst other studies cover only symptomatic cases (Michel et al. 2000), and the definitions of clinical illness vary from study to study. The degree of inaccuracy that results is by no means small: in two studies the estimates ranged for percent of cases reported ranged from approximately 2% to 9% (Hedberg et al. 1997; Thomas et al. 2006). For

Scotland, where there about 200 cases per year, this is a not insignificant range because a reporting rate of 9% would suggest that there are about 2,000 cases per year, a 2% rate, 10,000 cases.

By using a framework which does not involve data from a plethora of sources, the approach in this chapter bypasses many of the biases in the step-wise estimations used in the above mentioned studies. Eliminating the steps in the framework also reduces the chance that any error in the calculation will be compounded as the rate of reporting is recalculated at each step. In addition, an appropriate framework (if found) could lead to the development of a model that can not only be used to estimate prevalence, but also be to estimate the relative rates of large and small outbreaks. Such a model, developed from the framework, can also provide insight into the relationship between outbreak size and detection. The latter information could be very helpful in that in can inform surveillance agencies as to the effectiveness of surveillance programs as a function of event/outbreak size. For instance, a probability of case detection curve that has a dramatic increase starting at events of size 5 might indicate that increased effort needs to be put into investigating sporadic cases and small outbreaks.

8.4.3 Issues with the framework

8.4.3.1 Uncertainty

However, the approach does involve a degree of uncertainty, and the sources of this uncertainty must be considered. Firstly, there appears to be no published literature about case detection as a function of event size or on event size distributions (with both sporadic cases and outbreaks considered as events). The parameters - and thus curve shapes - for these distributions, as a result, must be selected based on assumptions and what is known about current surveillance systems and investigational techniques. This introduces uncertainty into the framework, as the parameter values must be selected partially based on assumption. The paucity of information about the potential shape of a curve for case detection and actual cases size probability also means that it may be possible to construct a model for which the distribution of event size proportions matches the observed distribution, but for which the contributing distribution and function are not correct. This is a situation

which, much simplified, could be compared to finding the 'correct' pair of values that add up to four, but not having a firm understanding of what makes a value 'correct'. There are three possibilities: two and two, three and one and zero and four, but if being 'correct' means the values must be less than three, only one of the pairs would be correct.

The validity of the distributions/functions must then be assessed, in large part, by examining the composition of each detected size bar in terms of the proportion of events of each actual event size. Such an examination would reveal the magnitude of under-reporting, for example whether a high proportion of large events are being detected as very small events or the majority of under-reporting is in terms of only a one or two cases. There is data from previous studies in other countries about the proportions of cases being lost to detection at various stages (Adak et al. 2002; Bender et al. 2004; Hedberg et al. 1997; Majowicz et al. 2005; Michel et al. 2000; Thomas et al. 2006). Therefore, it is possible to determine whether the magnitude of under-reporting as indicated by a model is within the range reported in countries with similar surveillance programs (the final models presented in this chapter suggest a case detection rate of 0.5%, which is lower than the ranges (2%-9%) suggested in the studies mentioned above.

Additionally, one strength of the multi-step approach towards prevalence estimation is the insight into specific biases in the surveillance data (Michel et al. 2000). The approach in this chapter does not permit under-reporting to be apportioned by cause. However, data from the framework presented in this chapter - a more accurate estimate for total prevalence and information about the probability of detection by event size - could be used together with the data from previous studies to create a more accurate reporting triangle. The final issue, which will be discussed in greater depth in the following section, is that of differing event definitions. In Scotland, only clusters of cases outwith a single residence are considered to be general outbreaks (Locking et al. 2003a), while cases from single household clusters are reported as unrelated (e.g. sporadic). However, for this framework, every epidemiological cluster must be considered as a separate event, and since data on single case events for the study is based on counts of sporadic cases from the HPS website (overall

count – number of outbreak cases), study data for events of size one is likely to include cases from household clusters. It is therefore assumed that the distribution of event size proportions for the observed data – which the model is being matched to – will overestimate the number of single case events and underestimate the number of small events. As the average household in Scotland is around 2.3 persons (General Register Office for Scotland 2006), it is presumed that it is the estimates of events of 1 to 3 cases or less that will be most affected, with effects potentially extending up to events with approximately 5 cases. As shown by the counts of the frameworks (Table 8.1), the difficulties with fitting the frameworks were most acute in the smaller event sizes.

8.4.3.2 Uncertainty – methods of addressing the issue

When trying to estimate the prevalence (or degree of underreporting) of a disease or condition, one important issue is sources of uncertainty, including incomplete or non-representative coverage and recall bias in surveys of a population or medical practitioners, imperfect sensitivity of a diagnostic tests and variation in calculations for factors. These sources of uncertainty must be recognised and the model or framework adjusted to account for them.

The methods of accounting for uncertainty vary depending upon the way in which prevalence or under-reporting is being estimated. In studies where prevalence is being estimated by sequential adjustment at each level of the reporting triangle, uncertainty is often introduced in surveys or surveillance. For example, since is rarely feasible to include the entire population of interest in a study, survey/surveillance participants, whether they are patients, person who seek GP treatment or physicians, may not be representative of the complete population (of patients etc.). Published data may also suggest a range of figures for particular factors such as the percent of patients with *E. coli* O157 infection who have bloody diarrhoea (Hedberg et al. 1997; Michel et al. 2000) and a degree of uncertainty is often known to be involved in mortality and hospitalisation rates due to misclassification (Mead et al. 1999).

Such uncertainty may be accounted for by simple methods, such as multiplying surveillance data by a factor to account for cases or deaths not picked up (Mead et al. 1999), or where estimates of a factor vary, using high and low estimates to calculate a range of results (Hedberg et al. 1997; Thomas et al. 2006). Also, confidence or credibility intervals are given both for figures from prior studies and for estimations (Flint et al. 2005; Prattley et al. 2007; Weller & Stanberry 2007; Wheeler et al. 1999). Uncertainty is also reduced by including data from as many sources as possible, which both increases the number of persons involved and provides are more accurate representation of the range (Michel et al. 2000). The use of high and low estimates, is not however, feasible when the data – as for the framework – is in terms of a distribution, and there was only one source of data available. In addition, the framework was developed in order than methods such as these, which produce imprecise estimates, do not have to be used.

When a formal statistical model is developed (both in sequential studies and those simply from surveillance data), techniques with increased statistical complexity can be used. On such technique is that of inputting each coefficient, which accounts for one level of the reporting triangle, as a distribution (Flint et al. 2005; MacDougall et al. 2007; Prattley et al. 2007). Additionally, the final estimation is often presented as a distribution. By using a distribution instead of a single figure or high and low figures, the model allows for uncertainty within each calculation and in the final result. Distributions that are used include the triangular distribution (Michel et al. 2000; Voetsch et al. 2004b), the normal distribution (Michel et al. 2000) and the uniform distribution (Michel et al. 2000). Parameters in a model equation may also be represented as probabilities, accounting for uncertainty within the estimation of the parameter (de Vlas & Gryseels 1992). To further adjust for uncertainty, a model may be run using a Monte Carlo simulation (Bradshaw et al. 2007; Groenewald et al. 2007; Joubert et al. 2007; MacDougall et al. 2007; Michel et al. 2000; Norman et al. 2007; Schneider et al. 2007a; Schneider et al. 2007b). As many as 10,000 or more iterations may be done, with the large number of model runs providing a more detailed distribution. Such methods, however, are not feasible for a framework, such as the one presented in this chapter, because a formal model is required to run a simulation. As will be discussed in the next section, the ability to develop this

framework into a formal model is currently limited by the known issue with the data regarding the inclusion of household outbreaks as sporadic cases and the more complex statistical techniques required.

Finally, using capture-recapture studies with confidence intervals may help reduce uncertainty. The method involves using multiple data sets from the same population, with prevalence being calculated based on the number of person listed on all data sets and the number of persons listed on one or some of the data sets. By using independent (or partially independent) data sets, the method allows for greater power in the statistical analyses of prevalence (Rota et al. 2007). This method was not feasible for the framework because data was only available from one source for Scotland.

8.4.3.3 Uncertainty – methods used in the framework

The major limitation of the approach presented here is that it is a framework, rather than a model. As a framework, approximate parameter values have been estimated by inserting values suggested by models of the functions and distributions that are included. Additionally, the final frameworks presented are not fully accurate descriptions of the Scottish data, with further analyses and more information regarding the data set, needed to come up with a final framework. Thus, unlike a formal statistical model for which a specific equation has been determined and hundreds or thousands (10,000 were done in above mentioned studies) of simulation tests have been run to come up with parameter values in the form a distribution, there is no formal or appropriate way to estimate uncertainty bounds. The next step would be to develop the framework into a formal model so that a technique such as Monte Carlo simulation could be used to run large numbers of model simulations. However, for two reasons, it was decided not to further develop the framework at this time. First, was the inability to properly use the framework without further knowledge about the number and size of household events included within the sporadic cases. Additionally, the framework needs to be expanded to account for the variations of the probability of cases detection within an event, a step which requires a significant increase in the complexity of the statistical analyses. As such, while the sources of uncertainty have been discussed, a formal estimate of uncertainty bounds

is not appropriate until the framework is developed into a model, at which point a method such as Monte Carlo simulation can be used to estimate uncertainty bounds for the parameters.

8.4.4 Exploration of the results from the framework

Exploration of the various potential distributions and functions and parameter values revealed that there is no straightforward model for the distribution of event size proportions in Scotland.

With no apparent published data on infectious disease event detection as a function of event size, the Logistic Growth Law function was selected to describe the function of the probability of case detection for two reasons. Firstly, a curve that was asymptotic at both ends [e.g. sigmoidal] was considered to be necessary because the probability of detection for any particular case, regardless of event size, is highly unlikely ever to be zero or one. The fact that asymptomatic cases, which have the highest potential to go unreported, are nevertheless reported (Locking et al. 2003a; Locking et al. 2004; Locking et al. 2006a) suggests that every case has a non-zero probability of detection. At the opposite end, it is extremely unlikely that any one case would have certain probability of being detected. Even in an instance where symptoms are severe, testing sensitivity is often less than 100% (Michel et al. 2000). Additionally, as evidenced by Figures 8.13 and 8.14, the Logistic Growth Law has a great deal of flexibility in the shape of the curve, an important factor since the exact shape of the detection curve is not known.

A range of distributions were considered for the distribution of the probability of actual event occurrence. Since the actual distribution is clearly not known, some assumptions must be based on the observed distribution. The observed distribution is highly overdispersed and the actual distribution is likely to be overdispersed because the events which are not being detected are likely to be smaller events, so the range is not likely to vary much from that of the observed distributions and the proportion of events that have one case is unlikely to decrease significantly. Due to this overdispersion, the truncated negative binomial and logarithmic distributions were considered to be the most appropriate distributions. The logarithmic distribution is the limiting form of the negative binomial distribution when k approaches zero (and

thus overdispersion is highest) (Elliott 1971b), so can be used if the truncated negative binomial distribution, even at low values of k (when over distribution is highest) is not over-distributed enough for the purposes of the framework.

The first step in the frameworks was to look at the distribution of the probability of actual event occurrence alone, the equivalent of a framework with a perfect probability of detection for all event sizes. Selecting an appropriate distribution for the actual event occurrence was vital because regardless of the probability of detection, events will not be detected if the probability of occurrence is too low. Yet, it is clear that modification of the distribution by the function of event detection is necessary. When the household clusters are accounted for, the proportion of events that are size one far exceeds that of any other event size. Since the function of case detection increases with event size and the magnitude of the difference in proportions between outbreaks of size one and size 2 is great, it would appear unlikely that this difference would be only a function of the probability of case detection. Resultingly, the distribution of the probability of actual event occurrence is most probably a major influence in the difference between the proportions of events of size one and events of size two or larger.

When complete frameworks were constructed, regardless of the distribution of event occurrence selected, very similar issues arose. First, distribution parameter(s) were selected to reflect the steep drop in proportions of detected events between events of size one and two of ~97% to ~1%. Yet, even with a very steeply increasing function of event detection, the predictions for counts/proportions of large outbreaks in the final framework were not large enough. Contrastingly, a much more gradual distribution shape was selected for event occurrence in order to ensure the occurrence of large event. However, even with a very gradual, increasing function of event detection, there was not a large enough difference between the proportions of events of size one and larger events. The logarithmic distribution performed slightly better in describing the highly aggregated distribution of the probability of actual event sizes.

8.4.5 Final frameworks

The results in Section 8.3.4 suggest that the distribution of the proportions of detected event sizes cannot be completely accurately constructed using the distributions and function explored in this chapter. However, the reasonable fit of the final frameworks in 8.3.4 indicates that the approach and framework have potential and provide clues about the reasons that for the difficulties in constructing an appropriate framework. When the best fitting frameworks are examined, the distribution of actual event occurrence can be described equally as well by a truncated negative binomial or logarithmic distribution, though the logarithmic framework will be discussed here. In either case the distribution has a relatively gradual curve, with the probability of events of size one approximately twice that of events of size two. The function of case detection is also not extreme, with a gentle increase in case detection until about event size =20 followed by a steep increase until maximum detection probability is effectively reached around event size=60. The resulting framework predicts the proportions/counts for all events other than those of size 2 fairly accurately (statistical measures of the framework fits using Chi-square techniques were not possible because of the low counts). The only major discrepancy between the framework and the observed distribution is in the count/proportion for events of size 2. The number of single case events is under-estimated, and the number of events size 3 and 4 are moderately overestimated, however these observed discrepancies are what could be expected given the issues with the household events/outbreaks discussed below.

The issues with household events arise because HPS classifies household outbreak cases as sporadic cases (SCIEH 2000b). For the purposes of this framework, each epidemiologically linked cluster – sporadic case or outbreak - must be considered to be a separate event because one of the assumptions of the function of event detection is that the size of an event affects the probability that a case will be detected. Household transmission is a well documented phenomenon (Belongia et al. 1993; Marsh et al. 1992; Parry & Salmon 1998) and testing of household contacts is often carried out in cases of *E. coli* O157 infection where transmission through an item or person within the household is suspected (Parry & Salmon 1998). Therefore it would be expected that a case that is part of a household outbreak would have a

much greater chance of being detected than a true sporadic case with no other epidemiologic linkage.

With an average household size of 2.7 in Scotland (General Register Office for Scotland 2006), the decision to count household outbreak cases as separate events would be expected to manifest itself in an observed distribution where there is an over-reporting of single case events and an under-reporting of events from size two up to ~ size five. Where each cluster is counted as a separate event, the expectation is that fewer single case events would be reported and there would be an over-reporting of events from size two to five. In the framework with the best fit to the data, the predicted distribution matches the latter situation, which is what would be expected from the actual, rather than observed data. This suggests that, despite the issues discussed, the frameworks may be one possible fit for the data

When the data from the best fit framework is then used to estimate the overall prevalence of events, the framework predicts that only 1.4% of events are detected, with 98.6% of events going completely undetected (e.g. detected size zero). Based on this detection estimate, the framework predicts that there were actually ~138,000 events and approximately 361,600 cases from outbreaks with ≤20 cases between 1996 and 2004. These figures suggest that about 0.7% of cases are reported, which is near the 2% lower estimate of the Canadian and American studies (Hedberg et al. 1997; Thomas et al. 2006). This estimate for the case detection rate seems low since Scotland has a comprehensive enhanced surveillance program which attempts a follow up of every reported case (Locking et al. 2003a). Such a program would be anticipated to result in a higher detection rate as compared to larger countries such at the United States, where the greater magnitude of case numbers (~200 per year in Scotland as compared to ~2000 in the United States) (Centers for Disease Control and Prevention 2005e; Health Protection Scotland 2007b), the size of the covered population and the number of levels of bureaucracy involved (local, county/parish, state, federal etc.) make it difficult to follow up on all cases.

However, it is the predictions the framework makes about the magnitude of under-reporting in terms of event size that would appear to be less supported by existing data. The best fitting framework suggests that fewer than 10% of the events which

are detected as size one are actually size one, and that events larger than 20 cases are being detected as size one. Also, of the outbreaks which have gone completely undetected, fewer than half are predicted to be size one, with around ten percent of the undetected events actually involving at least five cases. With the detailed follow-up of the enhanced surveillance program (Locking et al. 2003b; Locking et al. 2003a) and the PFGE typing of every isolate submitted to the Scottish *E. coli* O157 reference laboratory (Locking et al. 2003b), it would at first seem highly unlikely that 19 of 20 cases in an event would be missed. However, the explanation for lower observed rates of infection may lie in low rates of health care uptake and low test sensitivity (Michel et al. 2000). These prevent cases from ever coming to the attention of the enhanced surveillance system. In addition, more than half of the study period occurred prior to the implementation of enhanced surveillance. However, if the final framework (see Figure 8.20) was applied to data before (1035 events) and after (892 events) the onset of enhanced surveillance, both detected event size distributions were highly skewed (Table 8.2, lines a and c).

Table 8.2: Counts of detected event sizes, frameworks before and after the onset of enhanced surveillance

The counts of event sizes when counts are calculated from the framework with c=0.82, A=0.98, B=500 and Rho=1.15 using event size data (line a, b) before and (line c,d) after the onset of enhanced surveillance (ES). Actual counts are in lines a and c, the counts based on the framework proportions in lines b and d.

		Size of Event										
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11-20**
a	**Before ES**	1006	5	4	3	3	2	1	1	1	0	3
b	**framework**	928	67	18	8	4	2	2	1	1	<1	2
c	**After ES**	846	16	15	4	6	2	0	1	0	0	2
d	**framework**	800	58	15	7	4	2	1	1	1	<1	2

Thus the framework fits were similar to that of the complete data set, with the framework distribution overestimating the number of events with two cases and underestimating the number of events with one case. The fit for the data after enhanced surveillance was slightly better for events larger than size 2 (Table 8.2). A slightly better fit (not shown) for the distribution prior to enhanced surveillance by increasing the value for parameter B and lowering the value for parameter Rho in the Log Growth Law function for case detection (see Section 8.3.3). As discussed in

Section 8.3.3 the effect both of lowering Rho and raising B is to lower the rate of case detection in events, which could suggest that case detection has improved with the onset of enhanced surveillance.

The rate of patients seeking medical treatment for *E. coli* O157 infections generally depends upon the severity of symptoms, with a higher rate for patients with bloody diarrhoea (Thomas et al. 2006). Studies suggest that approximately 45% (18 to 91%) of cases involve bloody diarrhoea (Hedberg et al. 1997; Michel et al. 2000), and that the rate of patients seeking care with bloody diarrhoea varies from 33% to 73% and non bloody diarrhoea from16% to 22%. Furthermore, rates of sample submission and test sensitivity (especially when a longer period of time has elapsed between onset of symptoms and the taking of a faecal sample) are as low at 18% (Michel et al. 2000). Therefore, it is not impossible that many events are completely or largely missed. The highest reported rates of infection tend to be amongst young children and the elderly (Coia 1999). Evidence of infection has been found amongst farm families with no history of diagnosed infection (Silvestro et al. 2004). Though these results cannot be directly extrapolated to the whole population, they could suggest that a large number of infections in healthy adults do not cause any or severe symptoms, and thus are never detected by surveillance systems. It is possible, then, that larger events may only surface when the susceptible population contains a large enough proportion of young children or elderly to result in an increase in physician visits.

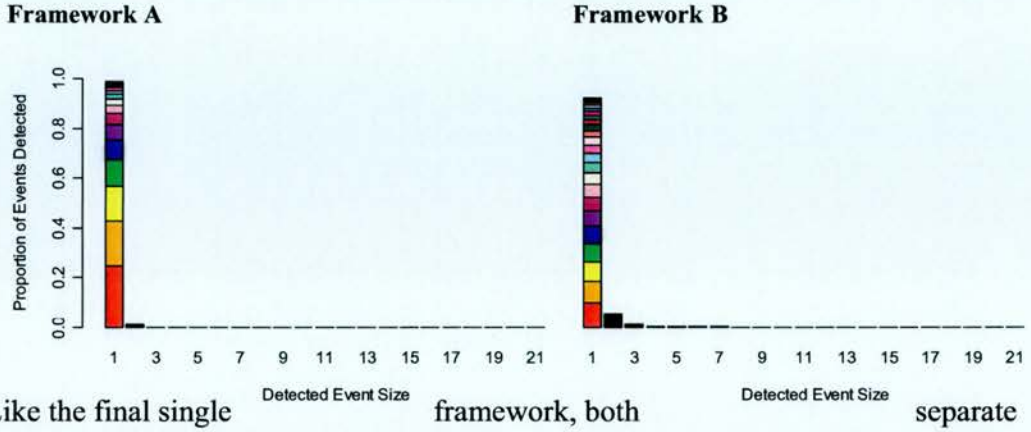
The best fit for the framework is still only one possible explanation for the distribution of detected event sizes and there are other factors which may affect the ability to find an adequate fit for the framework. First of all, the framework is not designed to function in the situation where a larger outbreak may be detected (completely or incompletely) as two or more smaller outbreaks. This is, of course, the situation with household outbreaks, but may also occur in instances where the 'outbreak fragments' are events of two or more cases. If indeed, some events are being detected as two or more separate events, the true distribution of detected events sizes would be expected to be much more gradual, and thus more easily fitted using the framework. Alternatively, it is possible that some larger outbreaks would be

better split into a number of smaller events. In fact, the Wishaw Outbreak, for the purposes of this chapter and other chapters (3 and 6), was considered to be five smaller outbreaks. If this is true, the actual distribution of detected events proportions would also be more gradual, but with a shorter right tail.

Also, in addressing the relationship between the probability of case detection and the size of an event, it is assumed that every case in an event would have the same probability of detection. In reality, this is probably not true because once a single case has been detected, it would seem that the chance of second or further case being detected is likely to increase. The probability of detection would increase because the confirmation of infection in one family member, classmate or other contact would seem to increase the chance that a person, especially one with symptoms would come to the attention of public health professionals or individually seek attention. The effect would be especially pronounced in event locations such as a nurseries or care facilities, where screening is more likely to be done on contacts in order to exclude infected persons until they are no longer shedding the bacteria (Al-Jader et al. 1999; Belongia et al. 1993; Cheasty et al. 1998; Pickering et al. 1986). If the above is true, the probability of the first case being detected would be linked to their symptoms, and chance of seeking treatment, but the probability of subsequent cases being detected is linked not only to the above factors, but to their connection to the first case.

An attempt was made to find more adequately fitting frameworks by splitting the detected event size distribution, allowing for different distributions of actual event occurrence depending on event size. In split frameworks, the assumption is that while there is only one function for the probability of case detection, there are two event occurrence distributions. For instance, when c=0.66 and A=0.98, B=10 and Rho=1.7, the framework predicted counts of 1906 and 20 for events of size one and 2, respectively (Table 8.1, Line 12, Fig. 8.23). If the value for c was adjusted to find the most accurate counts for events of greater than size 2, the most appropriate was c=0.8 (Table 8.1, Line 13). Whilst this split framework is still not a perfect fit for the observed data, the possible fit of the data it suggests that the distribution of actual event size may be best described by a mixture of distributions.

Figure 8.23: Split frameworks for the distribution of the proportions of detected events
Two frameworks which are combined to create one framework for the distribution of the proportion of detected events. Framework A, with c =0.66 is appropriate for events of size one and two, whilst Framework B with c=0.8 is appropriate for events of size three or greater. For both frameworks the parameters for the Logistic Growth Law are A=0.98, B=500, Rho=1.15.



Like the final single framework, both separate frameworks in this technique suggest at very high rates of under-reporting and many cases in large events going undetected. However, the greater accuracy of this framework, in particular the very accurate prediction for outbreaks of size two, should be considered with caution given the issues discussed above about the household outbreaks. If indeed the observed data overestimates the number of single case events and underestimates the number of small events (2-5 cases) due to the inclusion of household outbreak cases as sporadic cases, then an accurate prediction of the proportions of smaller detected events would not be expected. This approach does, however, suggest that the true distribution of actual event sizes may not as simple as a single distribution, but in fact a mixed distribution. Mixed distributions have been proposed as a method of describing biological patterns (Magurran & Henderson 2003). In this instance, the authors are describing areas where a few species are very common and most species are rare. They suggest that the common species are log normally distributed while rare species "follow a log series distribution". When these distributions are combined, the result is a skewed distribution like the one for the distribution of detected event sizes proportions in *E. coli* O157 in Scotland. Exploration of mixed distributions is beyond the scope of this thesis, but is a potential route of future exploration of the framework.

Finally, while seemingly adequate fits have been found for the distribution of the proportions (or counts) of detected event sizes using the presented framework, the

possibility that the framework is simply not appropriate must be considered. Since prior studies suggest that only a small fraction of cases are detected (Michel et al. 2000; Thomas et al. 2006), it may be that the population of detected events is too different from the actual population of events to extrapolate back from the detected distribution. In particular, since many factors such as geographic region, age, exposure to animals or animal faeces have been shown to affect the risk or chance of infection (Bryant et al. 1989; Coia et al. 1998; Locking et al. 2001; Mead et al. 1997; O'Brien et al. 2001a; Parry et al. 1998; Voetsch et al. 2006; Werber et al. 2007), there may be too many contributing factors to be able to describe case detection in terms of a simple function of event size.

It may also be that case detection is less strongly linked to event size, and more closely related to other factors. Firstly, mode of transmission could be a factor if, for instance, it was easier to track other potential cases through a link to a store or restaurant where a foodborne outbreak occurred, rather than a link to an infected animal or faeces (linked to an environmental outbreak). Another factor could be time period, e.g. whether the case occurred when there was great publicity due to a large ongoing event(s) – such as the Wishaw Outbreak (Cowden et al. 2001), Walkerton Outbreak (O'Connor 2002) or the Spinach Outbreak (Centers for Disease Control and Prevention 2006h) during which there was more awareness of *E. coli* O157 and thus potentially a higher rate of patients seeking medical care and being tested). A final factor could be location, for example if an event took place in location where more samples were tested for *E. coli* O157. If such factors are predominant in determining the probability of case detection, then a framework like the one explored here, which assumes case detection to be a function of event size and/or detection order of the case, would be unlikely to be able to appropriately framework case detection. In this instance, the best method of estimating prevalence would likely be the existing method of step wise estimation using data from population, physician and laboratory studies. However the estimate would have to be adjusted based on the factor or factors shown to be of influence. For instance, if cases were more likely to be detected in a certain region, then the estimation data would need to be waited to account for the disparity in detection. Additionally, the framework may not be appropriate because of the necessity to delineate cases by

event, a task that is complicated by the lack of an unequivocal way of identifying an event. For this chapter, all outbreaks except one (Wishaw Outbreak) were assumed to be individual events, but it is possible that some or many of the other 4 outbreaks of 5 or more cases, according to the definition in the chapter, are actually several events. However, since case by case information on outbreaks was not available due to reasons of maintaining patient anonymity, it is not possible at this time to determine which outbreaks might be composites of several events. If data on household outbreaks could be made available, the framework could be applied with more precision. Otherwise the degree of error – i.e. the number of sporadic cases that are in fact part of larger events – must be roughly estimated if the framework is to be used. An estimate of the error could be made if either a histogram of household outbreak sizes or the number of household outbreaks was available. Fitting the model to data from other countries where all the outbreak definition includes household clusters would help to validate the model. However, the model could not be fitted to the Canadian data because single case data was only available for all VTEC cases, rather than just *E. coli* O157, and also could not be fitted to the United States data because of missing data on confirmed outbreak size, particularly in 1998 and 1999 (see 4.2.1.2), the lack of single case data on only laboratory confirmed cases of *E. coli* O157 (see 4.2.1.1) after 2001 and the significant changes in outbreak numbers in 1998 due to a change in reporting practices (see Section 4.2.3.2 for further details).

8.4.6 Further use of the framework – discussion of issues

As discussed earlier, due to lack of data on sporadic cases, it was not practical to test the framework with data from Canada or the United States or to develop the framework into a formal model. However, if data were to be obtained from either of these countries or another country, when considering the applicability of the framework, there are a number of issues which must be considered. The primary consideration is that the mechanisms behind the distribution/functions of actual event sizes and detected event sizes may differ between countries.

The size distribution of events detected (and then listed in data sets) in each country could potentially depend on the type(s) of surveillance and detection methods being

used, the definition of an outbreak used in surveillance and the underlying distribution of actual event sizes. Firstly, in countries such as Scotland where there is extensive enhanced surveillance which includes an attempted follow-up of all persons with confirmed infection (Locking et al. 2003a), there may be a higher chance of linking together cases in an outbreak. Increased linkage of cases has the potential to result in more large events being reported and fewer single case events. Furthermore, depending on whether or not the definition of an outbreak includes events within a single household outbreak, the resulting distribution may be more or less skewed. The time period being considered when constructing a distribution of detected event sizes (which would vary depending on the data available from a country) might also affect outbreak sizes. For instance, one or a few large and/or well publicised outbreaks in a short-term data set could influence a distribution if the publicity and/or investigation surrounding an outbreak resulted in more sporadic cases being detected. For instance, the investigations surrounding the Wishaw Outbreak resulted in the detection of 7 outbreak cases for which no link to the source of contamination could be found (Cowden et al. 2001). This, however, is not likely to be a significant issue when using data sets that span 8 or 9 years, such as one from Scotland.

It may also be the case that the underlying distributions of actual event sizes differ between countries. Variation between distributions in actual event sizes could lead to differences in the sizes of events being detected. Hypothetical reasons for differing underlying distributions could include variations in control measures and in the predominant modes of transmission. For example, control measures which improve the speed of infection recognition and/or reduce the possibilities for secondary infection could result in a reduction in the average event size. Additionally, the size of events could be affected by the primary or secondary mode of transmission in an outbreak. As shown in the last chapter, it appears that outbreaks with particular primary or secondary modes of transmission have statistically significantly higher rates of secondary cases. Thus the predominant mode(s) of primary or secondary transmission in a country could potentially have an affect on the size of events. The frequency of modes of primary or secondary transmission could be affected by other factors including the age distribution of the

population and the urban/rural split of the population. For instance, a population where the majority of the population lives in very urban areas may be more likely to have outbreaks which are spread by food rather than by environmental contact.

Additionally, the applicability of the framework to countries other than Scotland could be influenced by the factors which affect case detection. These factors include the guidelines regarding the taking of faecal samples, and the accuracy of testing. A country where all samples from all patients with diarrhoea are cultured for the bacteria would likely have a higher detection rate than one where samples are only taken for culture from patients with bloody diarrhoea. Equally, where more accurate testing techniques are used, the likelihood of additional cases in an event being detected increases.

The level of public awareness about symptoms and the availability of affordable medical treatment could also affect the rate of detection by influencing the number of persons who seek medical treatment when they or a family member have symptoms. Investigational techniques also have the potential to influence the detection rate of infections. For instance, in a country such as Japan where widespread screening of case contacts is sometimes performed as part of an outbreak investigation (Akashi et al. 1994), the probability of subsequent cases being detected is much higher than in countries where such screening is not done for practical or economic reasons. Finally, the setting of an event could also play a role in the probability of case detection. For example, cases which are part of an outbreak at a nursery might have a higher chance of detection because young children are more likely to have severe symptoms. If the predominant settings for outbreaks vary between countries than the probability of event detection may be different.

If the framework presented in this chapter is to be developed into a model that can be used for other countries besides Scotland the factors discussed must be taken into account. The differences between countries might be adjusted for simply by changing the parameter values in the function or distributions involved or by using another function/distribution. However, if the influences regarding event detection are too complicated in a particular country to be explained by a single or mixture of functions/distributions, the framework would not be used. It may also be possible

that the combination of factors makes it impossible to find an appropriate distribution to adequately describe the distribution of actual event sizes.

8.4.7 Further development of the framework – validation

Another step that can be taken if the framework is developed into a proper model is for the model to be validated. At present, the lack of complete data about household outbreaks makes validation impractical as a final model cannot yet be determined. If a model(s) can be constructed for one or more countries, validation of the approach could potentially be accomplished through a number of methods. The only completely accurate method of validation would be to do a study of the entire population included in the model. Since the framework has been developed in order to provide an estimate because it is impossible to survey an entire country, this approach is clearly not practical.

However, there are two methods in which surveys could possibly be used to validate a model or models. Firstly, if accurate surveillance data can be acquired from two or three countries, the approach could be validated by seeing whether the same model (or models with different distributions according to the epidemiological factors in the individual countries) could be used in the countries. Since it may not be feasible to properly validate the approach if different models are required for each country, a more accurate method would likely be to focus on a specific region within a country. For instance, more detail about infection and outbreaks are available from the states and cities in the United States which are part of the Food-Net (active) surveillance network. If the data from one or more of the involved states/regions was modelled, and the modelling results could be compared to the estimates of prevalence based on the active surveillance studies that have been done. A similar method of validation could potentially be developed using the Canadian provinces involved in the NSAGI study, however the approach is not likely to be feasible in Scotland given the low outbreak and case counts overall.

However, a potential problem with validating a future model using data from past surveys or active surveillance is the difference in case definitions. For the framework, a case is considered to be an *E. coli* O157 infection that would be picked up by current detection methods. However, surveys such as NSAGI and IID,

primarily detect infections which are severe enough to cause symptoms of note to the patient. Even where surveys of the general community are done, they are usually by phone and thus based on general symptoms with no follow-up to determine the actual pathogen (if any) involved (Imhoff et al. 2004).

One other possible validation method is to the split the data set randomly and fit the framework to half the data. If the model is valid, then the parameters from this fit should be a good fit for the remainder of the data. This process could be repeated with multiple times using a computer program to randomly select a different half of the data each time.

8.4.8 Conclusion

In this chapter, a new framework for estimating the prevalence of *E. coli* O157 has been introduced, explored and discussed. Designed to avoid the inaccuracies in other estimation methods which rely on assessing the number of cases lost to surveillance at various stages between infection and reporting, the new method involves modelling the distribution of the proportions of detected events sizes. When the modelling was carried out, the logarithmic distribution appeared to be most accurate for the probabilities of event size occurrence and the Logistic Growth Law was used for the function of case detection. The ability to find an adequate model was hampered by differences in the way household events (or outbreaks) are defined by HPS and in this approach. The best fitting model using this distribution suggests that there is a very high (>99%) degree of under-reporting with a high magnitude of under-detection of cases within events, and an actual event size distribution which is much more gradual and less skewed towards smaller events than the observed distribution.

A number of concerns arise with this framework, including the issue with regards to household outbreaks and the assumption made that the probability of detection is identical for each case within a single event. However, the aim of this chapter was not necessarily to find a perfect fit for the Scottish data, but to determine whether a distribution based modelling approach might be suitable for estimation of case/event prevalence in Scotland. The construction of a model that is a reasonable fit to the data when the issues of household outbreaks were taken into account, suggests that

the approach is valid, though the possibility must always be considered that this framework is not appropriate even though a possibly adequate model was found. Furthermore, the work done in constructing and testing the model has highlighted a number of directions in which the framework can be taken in order to find a better fit for the Scottish data. These directions include the use of other distributions or functions, especially mixed distributions for the probabilities of actual event occurrence, and adjusting the function of case detection to allow for differing case detection probabilities within a single event.

Further exploration of this framework with these new directions could potentially lead to a more accurate estimation of the prevalence of *E. coli* O157 in Scotland. It is anticipated that a more accurate model, particular with regards to the household outbreak cases would predict a slightly higher rate of event detection with fewer large events going completely or mostly undetected. If this further exploration leads to a model that is a better fit and can be validated, the framework may be used to for estimation of prevalence in other countries, in particular to see whether the estimates agree with those calculated in prior studies using step-wise methods (Adak et al. 2002; Bender et al. 2004; Hedberg et al. 1997; Majowicz et al. 2005; Michel et al. 2000; Thomas et al. 2006). It is hoped that an accurate estimation based on models using the approach outlined in this chapter can compliment and improve estimations of case loss between infection and detection, with the end goal of providing epidemiologists and CsPHMs with a clearer picture of the magnitude of *E. coli* O157 infection in Scotland.

In summary, the study in this chapter has shown that true case prevalence can potentially be estimated using a framework based on the function of case detection and the distribution of actual event sizes. The models constructed using the framework suggest that the level of under-reporting in Scotland is higher than levels estimated for England & Wales, Canada and Scotland. However more exploration is needed to validate the approach, and to determine whether the framework is truly appropriate. Future pathways of exploration include the use of mixed distributions and an adjustment of the framework to make case detection of a function of event size and the order in which a case is detected within an outbreak.

In the following and final chapter, the conclusions of the studies in each chapter will be summarised, with a discussion of future research directions suggested by the study results.

Chapter 9 -- Conclusions and future directions

E. coli O157 has become a significant contributor to infectious intestinal illness in the developed world, and increasingly in the developing world (Nataro & Kaper 1998). Major outbreaks with more than 250 associated illnesses in Scotland, Canada and the United States, as well as hundreds or thousands of sporadic cases, have resulted in high costs (Frenzen et al. 2005; O'Connor 2002; Roberts & Upton 2000). These costs are both economic and in terms of mortality and morbidity – particularly of young children. As a result, the bacterium has been the subject of a substantial amount of research, focusing on microbiological (Friedrich et al. 2002), clinical (Tarr et al. 1990) and epidemiological (Rangel et al. 2005; Renter & Sargeant 2002) factors. However there have been several facets of *E. coli* O157 epidemiology that have not been addressed in the published literature. The research presented in this thesis has thus been conducted to fill in some of these gaps. The major focus of this thesis has been on statistical analysis of outbreaks and prevalence, with the three main foci of the research being temporal trends, secondary and primary cases in outbreaks and estimation of infection prevalence. Chapter 1 presented a brief literature review of *E. coli* O157 as well the issues of under-reporting and surveillance. The following five chapters, 2 – 6, provided statistical analyses of temporal trends in outbreak and cases in the three individual countries covered in this thesis: Scotland, the United States and Canada. These analyses were done in order to provide context for the comparison of temporal trends between these countries in Chapter 6, whilst Chapter 7 examined the characteristics of primary and secondary cases in *E. coli* O157 outbreaks detailed in the published literature. The chapter also included a statistical analysis of the relationship between secondary case rates and modes of primary and secondary transmission, median age and country. Chapter 8 was an exploration of a framework for estimating true case prevalence using Scottish data from 1996 to 2004. In this final chapter, the conclusions of these chapters will be briefly summarised, and the future directions suggested by these results will be discussed.

9.1 Temporal trends

The results in Chapters 3 (Scotland), 4 (United States) and 5 (Canada) indicated that the approach of using simple linear regression models to describe temporal trends in both outbreaks and cases was statistically appropriate in some, but not all cases.

Regardless of the country being analysed, two major factors were responsible for the inability to model certain trends. The first factor was low case or outbreak counts, due to a relatively short time period being analysed and/or a low number of outbreaks overall. The United States data set, in particular, had a shorter time period. This was because changes in the reporting practices, which resulted in a noticeable increase in the number of reported outbreaks (Lynch et al. 2006), meant that data could only be analysed for the period since 1998. Thus where counts for each year were highly variable (e.g. ranging from 2 to 252 ill cases in animal/environmental outbreaks per year) or very low (e.g. no more than 5 waterborne outbreaks per year), generalised linear models were not always able to accurately describe the trends. Low counts were also a problem, particular in Scotland where there were as few as three outbreaks in 1996 and 1998. Also, mode of transmission categories that tended to have low counts such as animal/environmental and waterborne were difficult to model. The second factor was large outbreaks which acted as potential outliers and thus reduced the ability of the trends involving case numbers to be adequately modelled using simple linear models. In each of the three countries, there were one or two outbreaks which involved more than 400 ill cases – the Wishaw Outbreak in Scotland, the Washington County Fair and Layton Avenue Sizzler Outbreaks in the United States and the Walkerton Outbreak in Canada (Cowden et al. 2001; New York State Department of Health 2000; O'Connor 2002; Proctor 2000a). In particular with the Scottish data, trends involving case numbers could not be appropriately modelled until these large outbreaks were removed from the data set. In order to allow the Scottish data to be modelled without completely excluding the Wishaw Outbreak cases, the trends were also modelled with the Wishaw Outbreak split into five outbreaks, the largest including 118 cases (as in the cohorts suggested by Cowden et. al, see Section 3.2.3.2). Clusters that shared an infection source, but were geographically distinct have been considered as separate outbreaks in other countries (Michino et al. 1999).

However, despite these difficulties with some of the variables, for the rest of the variables, simple linear models did fit the data. In Scotland the results suggested that there has been a decrease in sporadic cases and events, but not in the number of outbreaks or outbreak cases. However, outbreaks do appear to be involving fewer

cases, as there were statistically significant decreases in the number of ill and positive cases per outbreak and per foodborne outbreak. There were also statistically significant decreasing temporal trends in foodborne cases, which would appear to confirm that food is becoming a less significant factor in infection transmission (or non-foodborne outbreaks are becoming more significant factors). Finally, analysis of the Scottish data would appear to support statements regarding a shift in dominant outbreak phage types (Locking et al. 2003c), as outbreaks were statistically significantly less likely to be PT2 and more likely to be PT21/28 over the course of the time period. Additionally, the results suggest that the Wishaw Outbreak could be acting as a statistical outlier, and could possibly be considered as five (or more) separate outbreaks for the purposes of trend modelling. Contrastingly in the United States, analysed in Chapter 4, there was a statistically significant decline in the number of ill cases in outbreaks, both overall and for waterborne and foodborne outbreaks. Finally, the study of the Canadian trends in Chapter 5 suggested that outbreaks are increasing in size, both overall and for non-foodborne modes of transmission such as water and person to person contact.

The analyses of the three separate countries were brought together in Chapter 6, which is the first statistical comparison of temporal trends in the published literature. This comparison indicated that there were almost no statistically significant differences in temporal trends between countries. Though the statistical power of the comparison was limited by the short time period being compared (only 1998 to 2003 for all three countries), the few statistically significant differences in temporal trends could potentially be explained by the missing data from some provinces in the Canadian data set. The effect of this missing data was briefly explored by analysing data with and without the provinces/territories for which the largest amount of data was missing. While these analyses provided some perspective on how the missing data may have affected the results, this is one aspect of the study which could be further investigation. Such future explorations will be discussed in the next section.

9.2 Future directions for analyses of temporal trends

The analyses in Chapters 2 – 6 have established that some trends in *E. coli* O157 outbreak and cases can be modelled using simple linear regression models.

Furthermore there are some statistically significant trends in *E. coli* O157 outbreak and cases, but few statistically significant differences in these trends between countries. These findings are of interest because they provide the first evidence that some temporal trends can be statistically analysed using simple linear models and that the majority of these trends do not vary statistically significantly between countries. Yet they are only a first step towards investigation of these trends, and there are a number of future directions to build upon them. One of the major concerns regarding the study was the data because the time period was less than ten years and both in the United States and in Canada, the data sets were acknowledged to be incomplete. With two or three more years having elapsed since the receipt of the data sets, a new analysis including the data from 2004 (Canada) or 2005 through 2006 (or soon 2007) would permit a full ten years of data to be analysed. Allowing for delays in data release due to the need for the respective agencies to check and clean the data prior to allowing it to be accessed by researchers, such data should be available. Adding this data would provide more statistical power for detecting trends and suggest whether any of the detected trends in Chapters 2-6 are no longer statistically significant. In particular, the numbers of *E. coli* O157 infections in Scotland have increased in 2006, and are thus far, higher again in 2007 as compared to the same period in 2006. This increase may have resulted in a less (or non) statistically significantly decreasing trend in sporadic cases than noted in Chapter 3. Additionally, as the Canadian data set was recently collated, more complete information might be available. However, the official data sets made available for this thesis did not include data for years beyond 2004, so analyses of longer time periods could not be conducted. Also, it would be of interest to obtain more precise data from the United States and Canada on total and sporadic cases, as the current data was taken from web-based sources and the definitions for case inclusion not identical to that for the outbreak data sets. Official data for sporadic cases was not available for this thesis. If this data could be obtained for future studies, it would allow a more accurate analysis of trends in outbreaks in Scotland, and sporadic cases in all three countries.

There are issues concerning data protection with regards to sporadic case data, but only the aggregated month by month data is needed for analysis. Thus, if the data

could be stripped of patient identifiers and the data released only in terms of cases per month, it appears that the research needs and patient confidentiality needs could both be met. Even for months with only a handful of cases, the lack of any other information besides case numbers and the lack of press attention to single cases make it highly unlikely that data aggregated by month could ever be linked to a specific person.

Another point of future research is the large outbreaks – Wishaw, Walkerton, Washington County Fair and Layton Avenue Sizzler. As shown in Chapter 3, subdividing the Wishaw Outbreak into five smaller outbreaks (suggested by cohorts listed in a published report (Cowden et al. 2001)) allowed trends to be modelled without completely excluding the outbreak. It did not appear that the Walkerton Outbreak could be subdivided by cohorts, but there was not enough information on the latter two outbreaks (in the United States) to determine whether subdivisions would be possible. Additionally, more than 30 cases in the Wishaw Outbreak could not be assigned to a cohort based on the available information (Cowden et al. 2001). Thus it would be of great interest to get more detailed case information on the above mentioned large outbreaks as well as a number of large outbreaks which have occurred in 2005 and 2006 such as the spinach outbreak in the United States (Centers for Disease Control and Prevention 2006h). This would make it possible to examine in greater depth whether such outbreaks might be considered as clusters of smaller outbreaks when looking at long term trends or whether they are truly large outbreaks.

Finally, since some trends could not be described by simple linear regression models, future studies could explore techniques that might be able to describe these trends. In particular, if more specific time information – e.g. day by day or month by month – could be obtained for all of the data sets, the additional data points might enable other trends to be modelled. Using data subdivided by time periods less than a year would require additional analysis to remove the effects of seasonal trends. However, techniques –such as including year and month (or day) as variables, have been successfully used to model temporal trends (Guerin et al. 2005b). Additionally, as mentioned above, it would seem that monthly data on sporadic cases could be released with compromising patient confidentiality. Whilst there might be more

issues regarding month by month data for outbreaks in Scotland, outbreak data is already publicly released on a quarterly basis in the HPS Weekly Report and the month of outbreaks is not infrequently made known via published reports, news stories and peer reviewed articles. Also, it would be of interest to compare trends in other countries for which data was not available for this thesis with the trends in Scotland, Canada and the United States. Of particular would a comparison with England & Wales because despite the fact that the countries share a land border (between England & Scotland), rates of infection in England & Wales have consistently been lower than those in Scotland (Coia 1999). In addition, the surveillance systems in Scotland and England & Wales use the same outbreak definitions (Wall et al. 1996), so a comparison between these two countries could be made without the need for detailed sporadic case data. A final direction for future exploration would be to see if these simple linear regression models are adequate for describing trends in other infectious pathogens such as *Salmonella* and *Campylobacter*. As *Salmonella*, in particular, has been under surveillance for a longer than *E. coli* O157 (Centers for Disease Control and Prevention 2005e), data should be available for a longer time period, potentially enhancing the statistical power of the models.

9.3 Primary and secondary cases in outbreaks

In Chapter 7, the analyses of data from published reports on outbreaks in nine different countries, suggested that approximately 19% of cases in outbreaks are due to secondary transmission (as based on the definition used in this thesis). Additionally, the results indicated – that as with temporal trends – statistically significant differences in the numbers of primary and secondary cases and rates of secondary infection did not exist between countries. There were statistically significant differences in the geometric mean number of ill and positive cases outbreaks between countries, but this was not unexpected, given that the analyses in Chapter 6 indicated that there were statistically significant differences between the three countries analysed in Chapters 2-6 in the mean number of cases per outbreak. However, the results indicate that median age and secondary mode of transmission, but not primary mode of transmission, were statistically significant factors in determining the rate of secondary cases. In particular, outbreaks where the median

age was less than 6 and outbreaks where the mode of secondary transmission was person to person spread in a nursery school setting.

9.4 Future directions for analyses of primary and secondary cases in outbreaks

One of the primary limitations of this study on secondary cases was the biases inherent in taking data only from published studies. The analyses indicated that the outbreaks included from several countries were statistically significantly larger than the general population of reported outbreaks in countries where outbreak data sets were available. In addition, the data set for the study was, for example, biased towards outbreaks from the United States and England & Wales since these were the countries with the largest number of published papers in English on outbreak. In addition, it was clear that the lack of reports in English limited the number of Japanese outbreaks included, and may have affected the inclusion of outbreaks from Scandinavia. (The analysis was limited to countries with established *E. coli* O157 outbreak reporting systems). Furthermore, because published reports were used for the data, secondary cases were not always explicitly defined (e.g. (Gammie et al. 1996; Hilborn et al. 1999; Marsh et al. 1992)) and when they were defined, these definitions varied between outbreaks (Bell et al. 1994; Ostroff et al. 1990)). Thus the designation of secondary cases had to be made based only on the information provided in the paper. This information was not always complete and may have reflected the biases of the organisations and researchers involved (personal communication, Mary Locking, HPS).

However, the results obtained appear to suggest risk factors for higher incidences of secondary cases – e.g. secondary transmission in a nursery and young age, and so confirming the findings using more comprehensive data should be the first step in any future studies. In particular, a primary focus of any future explorations would be to obtain more complete outbreak information directly from national data sets rather than using data from published papers. (The data sets provided for Chapters 2-6 did not include information on secondary cases). This would not only permit a larger number of outbreaks to be analysed, but would also allow secondary cases to identified and defined with more precision. Equally as important, detailed information of outbreaks might help to clarify the primary mode of transmission,

which was missing for some of outbreaks (~29% of the 90 in the study) where the published report concentrated on the secondary mode of transmission. Additionally, data sets would likely provide information on more recent outbreaks (2003 and onwards) which are less likely to have so-far been detailed in published papers. Including these recent outbreaks would help because non-foodborne modes of transmission have been more commonly reported in recent years in some countries, including Scotland (Strachan et al. 2006), in recent years. Thus a more inclusive analysis would help to further clarify the role of mode of primary and secondary transmission in secondary case rates.

Also, since age had statistically significant role in determining the rate of secondary cases, an analysis of the role of age using exact age data should be a focus in any future investigations. Median age data is of use for looking at broad patterns, but especially for small outbreaks or outbreaks with widely varying groups of patient ages, the median may not be an accurate measure of the ages within an outbreak. For instance, if an outbreak has two person of age 2 and three persons of age 98, the median age would be 98, despite 40% of patients being only two years old. Thus using the average age instead of median age or even also looking at factors such as the percentage of persons in an outbreak below (or above) a certain age might provide a more accurate picture of how age affects the number of secondary cases. However, patient by patient data on ages or even enough data to determine average age was not available from enough of the published reports to permit such in depth analysis. Another factor which was not included in the Chapter 7 analyses due to lack of information in the papers was the male to female ratio of cases, both primary and secondary cases. It has been noted that women generally make up a greater percentage of cases, with the exception usually being young children (Locking et al. 2003b; Locking et al. 2003a; Michel et al. 1998; Parry & Palmer 2005). Thus the extension of this study to look at the sex ratio of outbreak cases, specifically primary cases versus secondary cases, might help to further elucidate the factors driving secondary case rates, especially if infection rates are linked to both patient sex and age.

These are all possibilities for future research, but the results in the study do suggest actions that can be taken or reinforced without any further studies. Given that the results indicated that young age and person to person contact in nurseries were statistically significantly associated with higher rates of secondary cases, the study reaffirms the need to promote proper hygiene practices amongst pupils and staff in nursery schools, and amongst young persons in general (Al-Jader et al. 1999). Prevention measures, advocated already in many countries (Chin 2000; Hawker et al. 2001; Keene et al. 1994; O'Brien et al. 2001b), such as hand-washing and careful monitoring or exclusion of non toilet trained children from the water in pools and other swimming areas not only could aid in preventing infections in nursery schools, but also in places like petting zoos and water parks.

9.5 Exploration of a framework for estimating the prevalence of *E. coli* O157 infections

The final study of this thesis suggested that a framework based on the distribution of detected event sizes has potential as a method for more accurately estimating the true prevalence of *E. coli* O157 infection. Though the framework was only explored using Scottish data because of limitations with the United States and Canadian data sets, it was possible to construct models that predicted distributions of actual event sizes which were similar to the distribution of detected event sizes in Scotland. Even though these models may not be completely accurate, it was of interest to note that the models suggested that if under-reporting rates in Scotland and other countries are similar, the estimates of under-reporting from existing studies may be too low. The model in the study suggested that fewer than 0.5% of outbreak cases were reported, as compared 2% (low end of the range) in studies in Canada and the United States (Bender et al. 2004; Michel et al. 2000). More specifically, the model indicated that even in reported outbreaks, many cases may be going undetected, and thus efforts should be continued to enhance and improve surveillance to detect cases that are not being reported. Whether or not these findings are representative of all countries is unknown, and a number of issues were discussed which should be further explored to affirm or question the proposed framework as a new method for estimating *E. coli* O157 prevalence.

9.6 Future directions in estimating the true prevalence of *E. coli* O157

The framework presented has potential promise as a different and potentially more accurate and more descriptive method for estimating case prevalence. Yet, the study presented should only be a first step, as there are a number of future pathways to expand and enhanced the current findings. These pathways could not be included in the current study because they either required data that was not available or significantly more complicated statistical analyses. Firstly, the use of a framework based on a distribution of events was hampered by differences in the definition of outbreaks. Specifically, in Scotland, cases that were part of outbreaks within a single household were considered to be sporadic cases (Locking et al. 2003a), and since information on these cases was not available for this thesis, it was not possible to create an entirely accurate plot of the distribution of event sizes. Specifically, models underestimated the number of single case events and overestimated the number of small (<5) case events. The explorations of the framework suggested that appropriately fitting models could possibly be constructed for distributions with fewer single case events and more events of 2 – 5 cases – the case if single household outbreaks were classified as single events. Thus, it would be of great benefit if more information on the single household outbreaks – ideally the number and size - could be obtained. Additionally, if data on sporadic case data for *E. coli* O157 cases alone could be obtained from Canada and Scotland (currently the number of sporadic cases is estimated from the number of total cases and the number of outbreak cases and only the total number of VTEC cases is available for Canada), the framework could be applied to these event distributions in these countries. The ability to explore the framework with several different data sets could provide more evidence as to whether the framework is appropriate, and if so, whether it is `appropriate for countries other than Scotland. Most importantly, the low number of events in Scotland has made statistically validating the fit of the framework to the data not possible. However there are sufficient numbers of events in the United States to allow for the fit to be statistically validated.

Another future direction would be to adjust the framework equations in order to take into account the fact that the probability of case detection may not only be influenced

by the size of an event, but by the order in which is a case is detected. For instance, the probability of the first case in an event being detected may be less than the probability of a subsequent case being detected. This is likely to be a point of serious consideration in large outbreaks where some cases may be detected due to testing or surveys initiated after the discovery of an initial cluster of cases. However, this adjustment would have required a much more complex statistic analysis, which was not feasible as part of this thesis. Additional factors could influence case detection probabilities, for instance if outbreak investigators are more likely to screen contacts of existing cases if an outbreak is in a certain location or spread by a certain method. For instance, an outbreak in a restaurant or food store might result in the staff being tested (McDonnell et al. 1997), whilst the staff would not be tested if the outbreak were in a pool or a petting zoo (Gage et al. 2001; Verma et al. 2007).

As discussed in Chapter 8, it would be of interest to explore further the possibility that the distribution of actual event sizes may not be a single distribution, but rather a mixture of distributions. Magurran and Henderson present a mixture distribution for describing the distribution of species in an environment, which was developed to explain distributions with a greater abundance of rare (or in outbreak terms, small) species (Magurran & Henderson 2003). This mixture distribution, which includes a log series distribution and a log normal distribution could considered rather than a single distribution for actual event sizes. However, caution has to be taken to avoid over-parameterisation, as one of the issues with the current framework is that low count numbers with the Scottish data and large numbers of parameters make it impossible to statistically asses the model fits.

The last future direction for the framework would be to apply the model to infectious diseases other than *E. coli* O157 which occur singly and in outbreaks. Two strong possibilities are *Campylobacter* and *Salmonella,* the numbers of which are thought to be underreported at even greater rates than *E. coli* O157 (Thomas et al. 2006; Voetsch et al. 2004b). Not only would better estimates of cases numbers for these pathogens be helpful in assessing prevention programs, but there are several advantages to using such data for testing the framework. First of all, these pathogens have been under surveillance for many more years in some countries, so data is

available in greater depth and for a longer period of time. Additionally, the reported rates of these two pathogens are much higher than for *E. coli* O157 (Centers for Disease Control and Prevention 2005e), thus the counts per year and/or month are much higher (the yearly reported counts of *Salmonella* in Scotland ranging from approximately 1000 to 3300, and the counts of *Campylobacter* from approximately 4300 to 6400) (Health Protection Scotland 2007a; Health Protection Scotland 2007e). A larger quantity of data not only makes it easier to test the fit of models, but allows the construction of a distribution of detected event sizes that better represents the reported population of a pathogen.

9.7 Overall conclusion

In this thesis, studies have been presented which advance and enhance the current knowledge of *E. coli* O157 trends, outbreak case characteristics and prevalence estimation techniques. Of the more important findings are the fact that it appears that country is not an important factor in outbreak and case trends or in the rates of secondary cases in outbreaks. Additionally, the thesis indicated that many outbreak and case trends can be described adequately by simple linear regression models, and that mode of secondary transmission and median age are significant factors in determining the rate of secondary cases in outbreaks. Also, a framework based on event size distributions has the potential for more accurately estimating the true prevalence of *E. coli* O157 cases. These findings not only statistically confirm existing statements about *E. coli* O157 epidemiology and prevention methods, but also suggest new ways of examining epidemiological factors, in particular looking at outbreaks as whole, rather than outbreak by outbreak. Most importantly, it is hoped that the results from studies in this thesis will also be considered as the impetus for future research to build and expand on the epidemiological research into *E. coli* O157 as well as infectious intestinal diseases in general.

References

Abbas, Z., Balram, C., MacDonald, B. W., Giffin, C. S., Aramini, J., & Panaro, L. (2005), "An Investigation of Two Simultaneous *E. coli* O157:H7 Outbreaks in Health Region 3, New Brunswick, August to September 2003", *Canadian Communicable Disease Report*, 31(22): 229-235.

Ackers, M. L., Mahon, B. E., Leahy, E., Goode, B., Damrow, T., Hayes, P. S., Bibb, W. F., Rice, D. H., Barrett, T. J., Hutwagner, L., Griffin, P. M., & Slutsker, L. (1998), "An outbreak of Escherichia coli O157 : H7 infections associated with leaf lettuce consumption", *Journal of Infectious Diseases*, 177(6): 1588-1593.

Ackman, D., Marks, S., Mack, P., Caldwell, M., Root, T., & Birkhead, G. (1997), "Swimming-associated haemorrhagic colitis due to *Escherichia coli* O157:H7 infection: evidence of prolonged contamination of a fresh water lake", *Epidemiology and Infection*, 119(1): 1-8.

Adak, G., Long, S., & O'Brien, S. (2002), "Trends in indigenous foodborne disease and deaths, England and Wales: 1992 to 2000", *Gut*, 51: 832-841.

Afza, M., Hawker, J., Thurston, H., Gunn, K., & Orendi, J. (2006), "An Outbreak of *Escherichia coli* O157 gastroenteritis in a care home for the elderly", *Epidemiology and Infection*, 134(6): 1276-1281.

Agha, M., DiMonte, B., Greenberg, M., Greenberg, C., Barr, R., & McLaughlin, J. R. (2006), "Incidence trends and projections for childhood cancer in Ontario", *International Journal of Cancer*, 118(11): 2809-2815.

Agnusdei, D., Camporeale, A., Gerardi, D., Rossi, S., Bocchi, L., & Gennari, C. (1993), "Trends in the Incidence of Hip Fracture in Siena, Italy, from 1980 to 1991", *Bone*, 14: S31-S34.

Ahmed, R., Bopp, C., Borczyk, A., & Kasatiya, S. (1987), "Phage-typing scheme for *Escherichia coli* O157:H7", *Journal of Infectious Diseases*, 155(4): 806-809.

Akashi, S., Joh, K., Tsuji, A., Ito, H., Hoshi, H., Hayakawa, T., Ihara, J., Abe, T., Hatori, M., Mori, T., & Nakamura, T. (1994), "A severe outbreak of haemorrhagic colitis and haemolytic uraemic syndrome associated with *Escherichia coli* O157:H7 in Japan", *European Journal of Pediatrics*, 153(9): 650-655.

Al-Jader, L., Salmon, R., Walker, A. M., Williams, H. M., Willshaw, G., & Cheasty, T. (1999), "Outbreak of *Escherichia coli* O157 in a nursery: lessons for prevention", *Archives of Disease in Childhood*, 81(1): 60-63.

Allaby, M. A. K. & Mayon-White, R. (1995), "*Escherichia coli* O157: outbreak in a day nursery", *CDR Review*, 5(1): R4-R6.

Anderson, R. M. & May, R. M. (1978), "Regulation and Stability of Host-Parasite Population Interactions: I. Regulatory Processes", *The Journal of Animal Ecology*, 47(1): 219-247.

Angulo, F., Voetsch, A. C., Vugia, D., Hadler, J., Farley, M., Hedberg, C. W., Cieslak, P. R., Morton, S., Dzogan, S., & Swerdlow, D. (1998), "Determining the Burden of Human Illness from Food Borne Diseases", *Veterinary Clinics of North America: Food Animal Safety*, 14(1): 165-172.

Anonymous (1997), "Surveillance of outbreaks of infectious intestinal disease, first quarter 1997", *SCIEH Weekly Report*, 31(97/20): 101-104.

Anonymous (2004), "Outbreak of VTEC O157:H7 Related to Farm Open to the Public", *Epi-News*, 25: 2-3.

Anonymous (2005), "Large *E. coli* O157 outbreak in Ireland, October-November 2005", *Eurosurveillance Weekly*, 10(12).

Anthony, R. G., Forsman, E. D., Franklin, A. B., Anderson, D. R., Burnham, K. P., White, G. C., Schwarz, C. J., Nichols, J. D., Hines, J. E., Olson, G. S., Ackers, S. H., Andrews, L. S., Biswell, B. L., Carlson, P. C., Diller, L. V., Dugger, K. M., Fehring, K. E., Fleming, T. L., Gerhardt, R. P., Gremel, S. A., Gutierrez, R. J., Happe, P. J., Herter, D. R., Higley, J. M., Horn, R. B., Irwin, L. L., Loschl, P. J., Reid, J. A., & Sovern, S. G. (2006), "Status and trends in demography of northern spotted owls, 1985-2003", *Wildlife Monographs*(163): 1-48.

Armitage, P., Berry, G., & Matthews, J. N. S. (2002), *Statistical Methods in Medical Research*, Blackwell Science Ltd., Oxford.

Armstrong, G., Hollingsworth, J., & Morris, J. (1996), "Emerging Foodborne Pathogens: *Escherichia coli* O157:H7 as a Model of Entry of a New Pathogen into the Food Supply of the Developed World", *Epidemiologic Reviews*, 18(1): 29-51.

Azlaf, R., Dakkak, A., Chentoufi, A., & El Berrahmani, M. (2007), "Modelling the transmission of Echinococcus granulosus in dogs in the northwest and in the southwest of Morocco", *Veterinary Parasitology*, 145(3-4): 297-303.

Bach, S. R., McAllister, T. A., Veira, D. M., Gannon, V. P. J., & Holley, R. A. (2002), "Transmission and Control of *Escherichia coli* O157:H7 - A review", *Canadian Journal of Animal Science*, 82: 475-490.

Banatvala, N., Magnano, A. R., Cartter, M. L., Barrett, T. J., Bibb, W. F., Vasile, L. L., Mshar, P., Lambert-Fair, M.-A., Green, J. H., Bean, N., & Tauxe, R. V. (1996), "Meat Grinders and Molecular Epidemiology: Two Supermarket Outbreaks of *Escherichia coli* O157:H7 Infection", *Journal of Infectious Diseases*, 173: 480-483.

BBC (2006), "Child dies from *E. coli* infection", retrieved on 15 October 2007 from: http://news.bbc.co.uk/1/hi/scotland/glasgow_and_west/5220768.stm.

Beall, G. & Rescia, R. R. (1953), "A Generalization of Neyman Contagious Distributions", *Biometrics*, 9(3): 354-386.

Bell, B., Goldoft, M., Griffin, P. M., Davis, M., Gordon, D., Tarr, P., Bartleson, C., Lewis, J., Barrett, T., Wells, J., Baron, R., & Kobayashi, J. (1994), "A Multistate

Outbreak of *Escherichia coli* O157:H7 - Associated Bloody Diarrhea and Hemolytic Uremic Syndrome From Hamburgers", *Journal of the American Medical Association*, 272(17): 1349-1353.

Belongia, E. A., Macdonald, K. L., Parham, G. L., White, K. E., Korlath, J. A., Lobato, M. N., Strand, S. M., Casale, K. A., & Osterholm, M. T. (1991), "An Outbreak of *Escherichia coli* O157:H7 Colitis Associated with Consumption of Precooked Meat Patties", *Journal of Infectious Diseases*, 164(2): 338-343.

Belongia, E. A., Osterholm, M. T., Soler, J. T., Ammend, D. A., Braun, J. E., & MacDonald, K. L. (1993), "Transmission of *Escherichia coli* O157:H7 Infection in Minnesota Child Day-care Facilities", *Journal of the American Medical Association*, 269(7): 883-888.

Bender, J. B., Hedberg, C. W., Besser, J. M., Boxrud, D. J., Macdonald, K. L., & Osterholm, M. T. (1997), "Surveillance by molecular subtype for *Escherichia coli* O157:H7 infections in Minnesota by molecular subtyping", *New England Journal of Medicine*, 337(6): 388-394.

Bender, J. B., Smith, K. E., McNees, A. A., Rabatsky-Ehr, T. R., Segler, S. D., Hawkins, M., Spina, N. L., Keene, W. E., Kennedy, M. H., Van Gilder, T. J., & Hedberg, C. W. (2004), "Factors Affecting Surveillance Data on *Escherichia coli* O157 Infections Collected from FoodNet Sites, 1996-1999", *Clinical Infectious Diseases*, 38(Suppl 3): S157-S164.

Beutin, L. (2006), "Emerging Enterohaemorrhagic *Escherichia coli*, Causes and Effects of the Rise of a Human Pathogen", *Journal of Veterinary Medicine Series B-Infectious Diseases and Veterinary Public Health*, 53(7): 299-305.

Bhat, M., Denny, J., MacDonald, K., Hofmann, J., Jain, S., & Lynch, M. (2007), "*Escherichia coli* O157:H7 Infection Associated with Drinking Raw Milk - Washington and Oregon, November-December 2005", *MMWR Weekly*, 56(8): 165-167.

Bisset, J. G., Brown, M. I., Bimson, J., Jones, K., McGulgan, M., Griffiths, A., & Hosie, B. (1992), "An Outbreak of Haemorrhagic Colities Due to *E. coli* O157 in Two Residential Homes for the Elderly in the Borders", *Communicable Disease and Environmental Health Scotland*, 26(19): 5-6.

Black, M. A. & Craig, B. A. (2002), "Estimating disease prevalence in the absence of a gold standard", *Statistics in Medicine*, 21: 2653-2669.

Blackmore, C. & Ginzl, D. (2005), "An *E. coli* O157:H7/HUS Outbreak Associated with Three Petting Zoos in Florida 2005: A Summary Report", *Epi Update.*

Bliss, C. I. & Fisher, R. A. (1953), "Fitting the Negative Binomial Distribution to Biological Data", *Biometrics*, 9: 176-199.

Bolduc,D. (2004), "Severe Outbreak of *Escherichia coli* O157:H7 in Health Care Institutions in Charelottetown, Prince Edward Island, Fall, 2002", retrieved on 9

October 2004 from: http://www.phac-aspc.gc.ca/publicat/ccdr-rmtc/04vol30/dr3009ea.html.

Bradshaw, D., Norman, R., Pieterse, D., & Levitt, N. S. (2007), "Estimating the burden of disease attributable to diabetes in South Africa in 2000", *Samj South African Medical Journal*, 97(8): 700-706.

Brenner, R. A., Overpeck, M. D., Trumble, A. C., DerSimonian, R., & Berendes, H. (1999), "Deaths attributable to injuries in infants, United States, 1983-1991", *Pediatrics*, 103(5): 968-974.

Breuer, T., Benkel, D., Shapiro, R., Hall, W., Winnett, M., Linn, M., Neimann, J., Barrett, T., Dietrich, S., Downes, F., Toney, D., Pearson, J., Rolka, H., Slutsker, L., & Griffin, P. M. (2001), "A Multistate Outbreak of *Escherichia coli* O157:H7 Infections Linked to Alfalfa Sprouts Grown from Contaminated Seeds", *Emerging Infectious Diseases*, 7(6): 977-982.

Brewster, D. H., Brown, M. I., Robertson, D., Houghton, G. L., Bimson, J., & Sharp, J. C. M. (1994), "An outbreak of *Escherichia coli* O157 associated with a children's paddling pool", *Epidemiology and Infection*, 112: 441-447.

Brown, H. & Prescott, R. (1999), *Applied Mixed Models in Medicine*, John Wiley & Sons Ltd, Chichester.

Bruce, M. G., Curtis, M. B., Payne, M. M., Gautom, R. K., Thompson, E. C., Bennett, A. L., & Kobayashi, J. I. (2003), "Lake-associated outbreak of *Escherichia coli* O157:H7 in Clark County, Washington, August 1999", *Archives of Pediatrics & Adolescent Medicine*, 157(10): 1016-1021.

Bruneau, A., Rodrigue, H., Ismäel, J., Dion, R., & Allard, R. (2004), "Outbreak of *E. coli* O157 Associated with Bathing at a Public Beach in the Montreal-Centre Region", *Canadian Communicable Disease Report*, 30(15).

Bryant, H. E., Athar, M. A., & Pai, C. H. (1989), "Risk Factors for *Escherichia coli* O157:H7 Infection in an Urban Community", *The Journal of Infectious Diseases*, 160(5): 858-864.

California Food Emergency Response Team (2007) *Investigation of an Escherichia coli O157:H7 Outbreak Associated with Dole Pre-Packaged Spinach*, State of California - Health and Human Services Agency, Department of Health Services, Food and Drug Branch.

Campos-Herrera, R., Escuer, M., Labrador, S., Robertson, L., Barrios, L., & Gutierrez, C. (2007), "Distribution of the entomopathogenic nematodes from La Rioja (Northern Spain)", *Journal of Invertebrate Pathology*, 95(2): 125-139.

Canadian Food Inspection Agency (2007), "*E. coli* O157:H7 Food Safety Facts - Preventing foodborne illness", retrieved on 28 January 2008 from: http://www.inspection.gc.ca/english/fssa/concen/cause/ecolie.shtml.

Caprioli, A., Morabito, S., Brugere, H., & Oswald, E. (2005), "Enterohaemorrhagic *Escherichia coli*: emerging issues on virulence and modes of transmission", *Veterinary Research*, 36(3): 289-311.

Cassie, R. M. (1962), "Frequency Distribution Models in the Ecology of Plankton and Other Organisms", *The Journal of Animal Ecology*, 31(1): 65-92.

CDC (1983), "International Notes Outbreak of Hemorrhagic Colitis -- Ottawa, Canada", *MMWR Weekly*, 32(10): 133-134.

CDC (2007), "Surveillance", retrieved on 16 November 2007 from: http://www.cdc.gov/enterics/about_surveillance.html.

Centers for Disease Control and Prevention (1994), "*Escherichia coli* O157:H7 Outbreak Linked to Home-Cooked Hamburger - - California, July 1993", *MMWR Weekly*, 43: 213-216.

Centers for Disease Control and Prevention (1995a), "*Escherichia coli* O157:H7 Outbreak Linked to Commercially Distributed Dry-Cured Salami -- Washington and California, 1994", *MMWR Weekly*, 44(09): 157-160.

Centers for Disease Control and Prevention (1995b), "Summary of Notifiable Diseases, United States, 1994", *MMWR Weekly*, 43(53).

Centers for Disease Control and Prevention (1996a), "Summary of Notifiable Diseases, United States, 1995", *MMWR Weekly*, 44(53).

Centers for Disease Control and Prevention (1996b), "Surveillance for Waterborne-Disease Outbreaks -- United States, 1993-1994", *MMWR Weekly*, 45(SS-1): 1-33.

Centers for Disease Control and Prevention (1997), "Summary of Notifiable Diseases, United States 1996", *MMWR Weekly*, 45(53): 1-87.

Centers for Disease Control and Prevention (1998), "Summary of Notifiable Diseases, United States, 1998", *MMWR Weekly*, 47(13): 1-93.

Centers for Disease Control and Prevention (1999a), "Public Health Dispatch: Outbreak of *Escherichia coli* O157:H7 and Campylobacter Among Attendees of the Washington County Fair -- New York, 1999", *MMWR Weekly*, 48(36): 803.

Centers for Disease Control and Prevention (1999b), *Surveillance for Outbreaks of Escherichia coli O157:H7 Infection: Summary of 1998 Data*, Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention (2000), *Summary of Outbreaks of Escherichia coli O157 and other Shiga toxin producing E. coli reported to the CDC in 1999*, Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention (2001a), *Reported Outbreaks of Shiga toxin-producing Escherichia coli (including O157:H7) occuring January 1 through December 31, 2000*, Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention (2001b), "Summary of Notifiable Diseases, United States, 1999", *MMWR Weekly*, 48(53): 1-104.

Centers for Disease Control and Prevention (2002a), *Reported Outbreaks of Shiga toxin-producing Escherichia coli (including O157:H7) occuring January 1 through December 31, 2001*, Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention (2002b), "Summary of Notifiable Diseases --- United States, 2000", *MMWR Weekly*, 49(53): 1-102.

Centers for Disease Control and Prevention (2002c), "Surveillance for Waterborne-Disease Outbreaks --- United States, 1999-2000", *MMWR Weekly*, 51(SS-8).

Centers for Disease Control and Prevention (2003a) *Shiga toxin-producing Escherichia coli - United States, 2002*, Accessed at: http://www.cdc.gov/foodborneoutbreaks/ecoli/CSTE2002.pdf.

Centers for Disease Control and Prevention (2003b), "Summary of Notifiable Diseases --- United States, 2001", *MMWR Weekly*, 50(53): 1-108.

Centers for Disease Control and Prevention (2003c), "Waterborne Diseases Outbreak Report", retrieved on 21 March 2007c.

Centers for Disease Control and Prevention. (2004a), Form 52.13 - Investigation of a Foodborne Outbreak. 2004a.

Centers for Disease Control and Prevention (2004b), "Summary of Notifiable Diseases --- United States, 2002", *MMWR Weekly*, 51(53): 1-84.

Centers for Disease Control and Prevention (2004c), "Surveillance for Waterborne-Disease Outbreaks Associated with Drinking Water --- United States, 2001-2002", *MMWR Weekly*, 53(SS08): 23-45.

Centers for Disease Control and Prevention (2005a), "*Campylobacter* Infections", retrieved on 16 April 2007a from: http://www.cdc.gov/ncidod/dbmd/diseaseinfo/campylobacter_g.htm.

Centers for Disease Control and Prevention (2005b), "*Escherichia coli* O157:H7", retrieved on 28 January 2008b from: http://www.cdc.gov/ncidod/dbmd/diseaseinfo/escherichiacoli_t.htm.

Centers for Disease Control and Prevention (2005c), "Outbreaks of *Escherichia coli* O157:H7 Associated with Petting Zoos --- North Carolina, Florida, and Arizona, 2004 and 2005", *MMWR Weekly*, 54(50): 1277-1280.

Centers for Disease Control and Prevention (2005d), "Preliminary FoodNet Data on the Incidence of Infection with Pathogens Transmitted Commonly Through Food -- 10 Sites, United States, 2004", *MMWR Weekly*, 54(14): 352-356.

Centers for Disease Control and Prevention (2005e), "Summary of Notifiable Diseases --- United States, 2003", *MMWR Weekly*, 52(54): 1-85.

Centers for Disease Control and Prevention (2006a), "Enterohemorrhagic *Escherichia coli* (EHEC) - 2000 Case Definition", retrieved on 21 March 2007a from: http://www.cdc.gov/epo/dphsi/casedef/escherichia_coli_current.htm.

Centers for Disease Control and Prevention (2006b), "*Escherichia coli* O157:H7 (*E. coli* O157:H7) - 1996 Case Definition", retrieved on 21 March 2007b from: http://www.cdc.gov/epo/dphsi/casedef/escherichia_coli_1996.htm.

Centers for Disease Control and Prevention (2006c) *FoodNet Surveillance Report for 2004 (Final Report)*, Centers for Disease Control and Prevention, Accessed at: http://www.cdc.gov/foodnet/annual/2004/Report.pdf.

Centers for Disease Control and Prevention (2006d) *Multistate Outbreak of E. coli O157 infections, November - December 2006*, Accessed at: http://www.cdc.gov/ecoli/2006/december/121406.htm.

Centers for Disease Control and Prevention (2006e), "Ongoing Multistate Outbreak of *Escherichia coli* serotype O157:H7 Infections Associated with Consumption of Fresh Spinach --- United States, September 2006", *MMWR Weekly*, 55(38): 1045-1046.

Centers for Disease Control and Prevention (2006f), "Passive vs. Active Surveillance", slide retrieved on 30 October 2007f.

Centers for Disease Control and Prevention (2006g), "Surveillance for Waterborne-Disease Outbreaks Associated with Recreational Water --- United States, 2003-2004 and Surveillance for Waterborne-Disease and Outbreaks Associated with Drinking Water and Water not Intended for Drinking --- United States, 2003-2004", *MMWR Weekly*, 55(SS-12).

Centers for Disease Control and Prevention (2006h) *Update on Multi-State Outbreak of E. coli O157:H7 Infections From Fresh Spinach, October 6, 2006*, Accessed at: http://www.cdc.gov/ecoli/2006/september/updates/100606.htm.

Centers for Disease Control and Prevention (2007a), "About Healthy People 2010", retrieved on 13 February 2008a from: http://www.cdc.gov/nchs/about/otheract/hpdata2010/abouthp.htm.

Centers for Disease Control and Prevention (2007b), "Compendium of Measures to Prevent Disease Associated with Animals in Public Settings, 2007", *MMWR Weekly*, 56(RR05): 1-13.

Centers for Disease Control and Prevention (2007c), "Healthy Swimming", retrieved on 31 January 2008c from: http://0-www.cdc.gov.pugwash.lib.warwick.ac.uk/healthyswimming/index.htm.

Centers for Disease Control and Prevention (2007d), "Multistate Outbreak of *E. coli* O157 Infections Linked to Topp's Brand Ground Beef Patties", retrieved on 15 October 2007d from: http://www.cdc.gov/ecoli/2007/october/100207.html.

Centers for Disease Control and Prevention (2007e), "Outbreak Surveillance Data" from: http://www.cdc.gov/foodborneoutbreaks/outbreak_data.htm.

Centers for Disease Control and Prevention (2007f), "Preliminary FoodNet Data on the Incidence of Infection with Pathogens Transmitted Commonly Through Food -- 10 States, 2006", *MMWR Weekly*, 56(14): 336-339.

Centers for Disease Control and Prevention (2007g), "Summary of Notifiable Diseases --- United States, 2004", *MMWR Weekly*, 53(53): 1-79.

Chalker, R. B. & Blaser, M. (1988), "A Review of Human Salmonellosis: III. Magnitude of Salmonella Infection in the United States", *Reviews of Infectious Disease*, 10(1): 111-124.

Chapman, P. A. (2000), "Sources of *Escherichia coli* O157 and experiences over the past 15 years in Sheffield, UK", *Journal of Applied Microbiology Symposium Supplement*, 88: 51S-60S.

Chapman, P. A., Wright, D., & Higgins, R. (1993), "Untreated milk as a source of verotoxigenic *E. coli* O157", *Veterinary Record*, 133: 171-2.

Chart, H., Rowe, B., vd, K. N., & Monnens, L. A. (1991), "Serological identification of *Escherichia coli* O157 as cause of haemolytic uraemic syndrome in Netherlands", *Lancet*, 337(8738): 437.

Chatfield, C., Ehrenberg, A. S. C., & Goodhardt, G. J. (1966), "Progress on a simplified model of stationary purchasing behaviour (with discussion)", *Journal of the Royal Statistical Society, Series A*, 129: 317-367.

Chaudhuri, K. (1988), "Dynamic optimizatino of combined harvesting of a two-species fishery", *Ecological Modelling*, 41(1-2): 17-25.

Cheasty, T., Allerberger, F., Beutin, L., Caprioli, A., Heuvelink, A., Karch, H., Lofdahl, S., Pierard, D., Scheutz, F., Siitonen, A., & Smith, H. (2000), "A comparison of Verocytotoxin-producing *Escherichia coli* O157 phage types isolated in England and Wales with those from 13 other European countries: January 1997 to June 1999", *SCIEH Weekly Report*, 34(2000/05): 34.

Cheasty, T., Robertson, R., Chart, H., Mannion, P., Syed, Q., Garvey, R., & Rowe, B. (1998), "The use of serodiagnosis in the retrospective investigation of a nursery outbreak associated with *Escherichia coli* O157:H7", *Journal of Clinical Pathology*, 51(7): 498-501.

Chin, J. (2000), *Control of Communicable Diseases Manual*, American Public Health Association, Washington, D.C.

Cimolai, N., Carter, J. E., Morrison, B. J., & Anderson, J. D. (1990), "Risk factors for the progression of *Escherichia coli* O157:H7 enteritis to hemolytic-uremic syndrome", *The Journal of Pediatrics*, 116(4): 589-592.

Clark, A., Morton, S., Wright, P., Corkish, J., Bolton, F., & Russell, J. (1997), "A community outbreak of Vero cytotoxin producing *Escherichia coli* O157 infection linked to a small farm dairy", *CDR Review*, 7(13): R206-R211.

Cody, S. H., Glynn, M. K., Farrar, J. A., Cairns, K. L., Griffin, P. M., Kobayashi, J., Fyfe, M., Hoffman, R., King, A. S., Lewis, J. H., Swaminathan, B., Bryant, R. G., & Vugia, D. J. (1999), "An Outbreak of *Escherichia coli* O157:H7 Infection from Unpasteurized Commercial Apple Juice", *Annals of Internal Medicine*, 130(3): 202-209.

Coia, J. E. (1998a), "Clinical, microbiological and epidemiological aspects of *Escherichia coli* O157 infection", *FEMS Immunology and Medical Microbiology*, 20: 1-9.

Coia, J. E. (1998b), "Nosocomial and laboratory-acquired infection with *Escherichia coli* O157", *Journal of Hospital Infection*, 40: 107-113.

Coia, J. E. (1999), "Controlling *Escherichia coli* O157: the emerging challenge", *Journal of Hospital Infection*, 43 (supplement): S175-S181.

Coia, J. E., Curnow, J., Reilly, W., Synge, B. A., & Sharp, J. C. M. (1995), "Ten years' experience of *Escherichia coli* O157 infection in Scotland", *SCIEH Weekly Report*, 29(95/01): 3-4.

Coia, J. E., Curnow, J., & Tolland, J. (1993), "*Escherichia coli* O157 Infections in Scotland, 1992", *Communicable Disease and Environmental Health Scotland*, 27(93/38): 5-7.

Coia, J. E., Davis, B., & Reilly, W. (1996), "*E. coli* O157 infection in Scotland, 1994", *SCIEH Weekly Report*, 30(96/05): 29-30.

Coia, J. E., Sharp, J. C. M., Campbell, D. M., Curnow, J., & Ramsay, C. N. (1998), "Environmental Risk Factors for Sporadic *Escherichia coli* O157 Infection in Scotland: Results of a Descriptive Epidemiology Study", *Journal of Infection*, 36: 317-321.

Conedera, G., Mattiazzi, E., Russo, F., Chiesa, E., Scorzato, I., Grandesso, S., Bessegato, A., Fioravanti, A., & Caprioli, A. (2007), "A family outbreak of *Escherichia coli* O157 haemorrhagic colitis caused by pork meat salami", *Epidemiology and Infection*, 135(2): 311-314.

Conway, D. I., Stockton, D. L., Warnakulasuriya, K. A. A. S., Ogden, G., & Macpherson, L. M. D. (2006), "Incidence of oral and oropharyngeal cancer in United

Kingdom (1990-1999) - recent trends and regional variation", *Oral Oncology*, 42(6): 586-592.

Cowden, J. M. (1997a), "For debate: a new system for surveillance of outbreaks of infectious intestinal disease in Scotland", *SCIEH Weekly Report*, 29(95/33): 183-184.

Cowden, J. M. (1997b), "Recent Outbreaks of *E. coli* O157 in Scotland", *SCIEH Weekly Report*, 1(97/13): 6-7.

Cowden, J. M., Ahmed, S., Donaghy, M., & Riley, A. (2001), "Epidemiological investigation of the Central Scotland outbreak of *Escherichia coli* O157 infection, November to December 1996", *Epidemiology and Infection*, 126(335-341).

Crampin, M., Willshaw, G., Hancock, R., Djuretic, T., Elstob, C., Rouse, A., Cheasty, T., & Stuart, J. (1999), "Outbreak of *Escherichia coli* O157 Infection Associated with a Music Festival", *European Journal of Clinical Microbiology and Infectious Disease*, 18: 286-288.

Crawley, M. J. (2002), *Statistical Computing: An Introduction to Data Analysis using S-Plus*, John Wiley & Sons Ltd, Chichester, England.

Crawley, M. J. (2005), *Statistics: An Introduction using R*, John Wiley & Sons Ltd., Chichester.

Crawley, M. J. (2007), *The R Book*, John Wiley & Sons, Ltd., Chichester.

Croft, D., Archer, J., Roberts, C. M., Johnson, R. A., Monson, T., Lucas, D., Kurzynski, T., Hoang-Johnson, D., Machmueller, L., Kelly, L., Grossfield, D., Rosario, G., Kaspar, K., Crump, J., & Davis, J. P. (2002) *An outbreak of Escherichia coli O157:H7 infections associated with a pancake breakfast served in a stock pavilion with contaminated livestock bedding - Wisconsin, 2001*, EIS Conference - April 2002.

Crofton, H. D. (1971), "A quantitative appraoch to parasitism", *Parasitology*, 62: 170-193.

Crump, J., Braden, C., Dey, M., Hoekstra, R., Rickelman-Apisa, J., Baldwin, D., De Fijter, S., Nowicki, S., Koch, E., Bannerman, T., Smith, F., Sarisky, J., Hochberg, N., & Mead, P. S. (2003), "Outbreaks of *Escherichia coli O157* infections at multiple county agricultural fairs: a hazard of mixing cattle, concession stands and children", *Epidemiology and Infection*, 131: 1055-1062.

Cumberland, P., Sethi, D., Roderick, P. J., Wheeler, J. G., Cowden, J. M., Roberts, J. A., Rodrigues, L. C., Hudson, M. J., & Tompkins, D. S. (2003), "The Infectious Intestinal Disease Study of England: A prospective evaluation of symptoms and health care use after an acute episode", *Epidemiology and Infection*, 130: 453-460.

Das, P. K., Manoharan, A., Srividya, A., Grenfell, B. T., Bundy, D. A., & Vanamail, P. (1990), "Frequency distribution of *Wuchereria bancrofti* microfilariae in human populations and its relationships with age and sex", *Parasitology*, 101(3): 429-434.

David, S. T., MacDougall, L., Louie, K., McIntyre, L., Paccagnella, A. M., Schleicher, S., & Hamade, A. (2004), "Petting zoo-associated *Escherichia coli* O157:H7 - Secondary transmission, asymptomatic infection, and prolonged shedding in the classroom", *Canadian Communicable Disease Report*, 30(20): 173-180.

Davis, B. & Brogan, R. (1995), "A widespread community outbreak of *E. coli* O157 infection in Scotland", *Public Health*, 109: 381-388.

Day, N. P., Scotland, S. M., Cheasty, T., & Rowe, B. (1983), "*Escherichia coli* O157:H7 Associated with Human Infections in the United-Kingdom", *Lancet*, 1(8328): 825.

De Boer, E. & Heuvelink, A. E. (2000), "Methods for the detection and isolation of Shiga toxin-producing *Escherichia coli*", *Journal of Applied Microbiology*, 88: 133S-143S.

de Vlas, S. J. & Gryseels, B. (1992), "Underestimation of *Schistosoma mansoni* Prevalences", *Parasitology Today*, 8(8): 274-277.

Deneen, V., Wicklund, J., Shallow, S., Stegler, S., Townes, J., Reddy, S., & Hennessey, T. (1998a), "The Impact of Physician Knowledge of Laboratory Practices on Detection of *E. coli* O157:H7", 1st International Conference on Emerging Infectious Diseases, Atlanta, GA.

Deneen, V., Wicklund, J., Shallow, S., Stegler, S., Townes, J., Reddy, S., & Hennessey, T. (1998b), "The Impact of Physician Knowledge of laboratory Practices on Detection of *E. coli* O157:H7", 1st International Conference on Emerging Infectious Diseases, Atlanta, GA.

Douglas, A. S. & Kurien, A. (1997), "Seasonality and Other Epidemiological Features of Haemolytic Uraemic Syndrome and *E. coli* O157 Isolates in Scotland", *Scottish Medical Journal*, 42: 166-171.

Duffy, G. (2003), "Verocytoxigenic *Escherichia coli* in animal faeces, manures and slurries", *Journal of Applied Microbiology*, 94: 94S-103S.

Dupont, W. & Plummer, W. (1998), "Power and sample size calculations for studies involving linear regression", *Controlled Clinical Trials*, 19: 589-601.

Elashoff, J. D. (2005), nQuery Advisor. [6.0]. 2005.

Elliott, J. M. (1971a), *Some Methods for the Statistical Analysis of Samples of Benthic Invertebrates*, Freshwater Biological Association, Ambleside.

Elliott, J. M. (1971b), *Some Methods for the Statistical Analysis of samples of Benthic Invertebrates*, Freshwater Biological Association, Ambleside.

Espie, E., Vaillant, V., Mariani-Kurkdjian, P., Grimont, F., Martin-Schaller, R., De Valk, H., & Vernozy-Rozand, C. (2006), "*Escherichia coli* O157 outbreak associated

with fresh unpasteurized goats' cheese", *Epidemiology and Infection*, 134(1): 143-146.

EURODIAB ACE Study Group (2000), "Variation and trends in incidence of childhood diabetes in Europe", *The Lancet*, 355: 873-876.

European Commission (2006) *ENTER-NET Annual Report 2004 - Surveillance of Enteric Pathogens in Europe and Beyond.*

European Commission (2007) *ENTER-NET Annual Report 2005.*

Evans, D. A. (1953), "Experimental Evidence Concerning Contagious Distributions in Ecology", *Biometrika*, 40(1-2): 186-211.

Eves, A., Bielby, G., Egan, B., & Lumbers, M. (2006), "Food hygiene knowledge and self-reported behaviours of UK school children (4-14 years)", *British Food Journal*, 108(9): 706-720.

Feldman, K. A., Mohle-Boetani, J. C., Ward, J., Furst, K., Abbott, S. L., Ferrero, D. V., Olsen, A., & Werner, S. B. (2002), "A cluster of *Escherichia coli* O157: nonmotile infections associated with recreational exposure to lake water", *Public Health Reports*, 117(4): 380-385.

Ferguson, D. D., Scheftel, J., Cronquist, A., Smith, K., Woo-Ming, A., Anderson, E., Knutsen, J., De, A. K., & Gershman, K. (2005), "Temporally distinct *Escherichia coli* O157 outbreaks associated with alfalfa sprouts linked to a common seed source - Colorado and Minnesota, 2003", *Epidemiology and Infection*, 133(3): 439-447.

Fewster, R. M., Buckland, S. T., Siriwardena, G. M., Baillie, S. R., & Wilson, J. D. (2000), "Analysis of Population Trends For Farmland Birds Using Generalized Additive Models", *Ecology*, 81(7): 1970-1984.

Finney, D. J. & Varley, G. C. (1955), "An Example of the Truncated Poisson Distribution", *Biometrics*, 11(3): 387-394.

Fisher, I. (2002), "From Enter-net: International surveillance network for the enteric infections Salmonella and VTEC O157", *IVC News 16*, 15(12 Supplement 1).

Fisher, R. A. (1941), "The negative binomial distribution", *Annals of Eugenics*, 11: 182-187.

Flint, J. A., Van Duynhoven, Y. T., Angulo, F. J., DeLong, S. M., Braun, P., Kirk, M., Scallan, E., Fitzgerald, M., Adak, G. K., Sockett, P., Ellis, A., Hall, G., Gargouri, N., Walke, H., & Braam, P. (2005), "Estimating the burden of acute gastroenteritis, foodborne disease, and pathogens commonly transmitted by food: An international review", *Clinical Infectious Diseases*, 41(5): 698-704.

Food Standards Agency (2007), "4 C's strategy", retrieved on 17 November 2007 from: http://www.food.gov.uk/safereating/safcom/fdscg/4cstrategy.

FoodNet Working Group (1997), "Foodborne Diseases Active Surveillance Network (FoodNet)", *Emerging Infectious Diseases*, 3(4): 579-581.

Frenzen, P. D., Drake, A., & Angulo, F. J. (2005), "Economic cost of illness due to *Escherichia coli* O157 infections in the United States", *Journal of Food Protection*, 68(12): 2623-2630.

Friedman, M. S., Roels, T., Koehler, J. E., Feldman, L., Bibb, W. F., & Blake, P. (1999), "*Escherichia coli* O157:H7 outbreak associated with an improperly chlorinated swimming pool", *Clinical Infectious Diseases*, 29(2): 298-303.

Friedrich, A. W., Bielaszewska, M., Zhang, W.-L., Pulz, M., Kuczius, T., Ammon, A., & Karch, H. (2002), "*Escherichia coli* Harboring Shiga Toxin 2 Gene Variants: Frequency and Association with Clinical Symptoms", *Journal of Infectious Disease*, 185(1): 74-84.

Furowicz, A. J. & Orskov, F. (1972), "2 New *Escherichia coli* O Antigens, 0159 and 0157, and One New K Antigen, K92, in Strains Isolated from Veterinary Diseases", *Acta Pathologica et Microbiologica Scandinavica Section B-Microbiology and Immunology*, B 80(3): 441.

Gage, R., Crielly, A., Baysinger, M., & Chernak, E. (2001), "Outbreaks of *Escherichia Coli* O157:H7 Infections Among Children Associated with Farm Visits - Pernnsylvania and Washington, 2000", *MMWR Weekly*, 50(15): 293-97.

Galanis, E., Longmore, K., Hasselback, P., Swann, D., Ellis, A., & Panaro, L. (2003), "Investigation of an *E. coli* O157:H7 Outbreak in Brooks, Alberta, June-July 2002: The Role of Occult Cases in the Spread of Infection Within a Daycare Setting", *Canadian Communicable Disease Report*, 29(03).

Gammie, A. J., Mortimer, P. R., Hatch, L., Brierley, A. F., Chada, N., & Walters, J. B. (1996), "Outbreak of Verocytotoxin-producing *Escherichia coli* O157 associated with cooked ham from a single source", *PHLS Microbiology Digest*, 13(3): 142-145.

Garg, A. X., Moist, L., Matsell, D., Thiessen-Philbrook, H. R., Haynes, R. B., Suri, R. S., Salvadori, M., Ray, J., Clark, W. F., & for The Walkerton Health Study Investigators (2005), "Risk of hypertension and reduced kidney function after acute gastroenteritis from bacteria-contaminated drinking water", *Canadian Medical Association Journal*, 173(3): 261-268.

General Register Office for Scotland (2006), "Registrar General Announces Further Results From The 2001 Census", retrieved on 24 September 2007 from: http://www.gro-scotland.gov.uk/press/news2003/cenresprs.html.

Gerrodette, T. (1987), "A Power Analysis for Detecting Trends", *Ecology*, 68(5): 1364-1372.

Gerrodette, T. (1993a), TRENDS. 1st. 1993a. La Jolla, California. 1-1-1993a.

Gerrodette, T. (1993b), "Trends: Software for a Power Analysis of Linear Regression", *Wildlife Society Bulletin*, 21: 515-516.

Goh, S., Newman, C., Knowles, M., Bolton, F. J., Hollyoak, V., Richards, S., Daley, P., Counter, D., Smith, H. R., & Keppie, N. (2002), "*E. coli* O157 phage type 21/28 outbreak in North Cumbria associated with pasteurized milk", *Epidemiology and Infection*, 129(3): 451-457.

Goldwater, P. N. & Bettelheim, K. A. (1995), "*Escherichia coli* serotypes other than O157:H7 as causes of disease in Australia", *Communicable Diseases Intelligence*, 19: 2-4.

Gouveia, S., Proctor, M., Lee, M.-S., Luchansky, J. B., & Kaspar, C. W. (1998), "Genomic Comparisons and Shiga Toxin Production among *Escherichia coli* O157:H7 Isolates from a Day Care Center Outbreak and Sporadic Cases in Southeastern Wisconsin", *Journal of Clinical Microbiology*, 36(3): 727-733.

Grafen, A. & Woolhouse, M. E. J. (1993), "Does the Negative Binomial Distribution Add Up", *Parasitology Today*, 9(12): 475-477.

Grenfell, B. T., Das, P. K., Rajagopalan, P. K., & Bundy, D. A. (1990), "Frequency distribution of lymphatic filariasis microfilariae in human populations; population processes and statistical estimation", *Parasitology*, 101(3): 417-427.

Grif, K., Orth, D., Lederer, I., Berghold, C., Roedl, S., Mache, C. J., Dierich, M. P., & Wurzner, R. (2005), "Importance of environmental transmission in cases of EHEC O157 causing hemolytic uremic syndrome", *European Journal of Clinical Microbiology & Infectious Diseases*, 24(4): 268-271.

Griffin,P.M. (1998) "Epidemiology of Shiga Toxin-Producing *Escherichia coli* Infections in Humans in the United States," in *Escherichia coli O157:H7 and Other Shiga Toxin-Producing E. coli Strains*, J. Kaper & A. O'Brien, eds., ASM Press: Washington, DC.

Griffin, P. M. & Tauxe, R. V. (1991), "The Epidemiology of Infections Caused by *Escherichia coli* O157:H7, Other Enterohemorrhagic *E coli*, and the Associated Hemolytic Uremic Syndrome", *Epidemiologic Reviews*, 13: 60-98.

Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C., & Gelbart, W. M. (1996), *An Introduction to Genetic Analysis*, W.H. Freeman and Company, New York.

Groenewald, P., Vos, T., Norman, R., Laubscher, R., Van Walbeek, C., Saloojee, Y., Sitas, F., & Bradshaw, D. (2007), "Estimating the burden of disease attributable to smoking in South Africa in 2000", *Samj South African Medical Journal*, 97(8): 674-681.

Guerin, M. T., Martin, S. W., & Darlington, G. A. (2005a), "Temporal clusters of *Salmonella* serovars in humans in Alberta 1990-2001", *Canadian Journal of Public Health-Revue Canadienne de Sante Publique*, 96(5): 390-395.

Guerin, M. T., Martin, S. W., Darlington, G. A., & Rajic, A. (2005b), "A temporal study of *Salmonella* serovars in animals in Alberta", *The Canadian Journal of Veterinary Research*, 69: 88-99.

Haas, C. N., Thayyar-Nadabusi, A., Rose, J. B., & Gerba, C. P. (2000), "Development of a dose-response relationship for *Escherichia coli* O157:H7", *International Journal of Food Microbiology*, 1748: 153-159.

Halliday, J. E. B., Chase-Topping, M. E., Pearce, M. C., McKendrick, I. J., Allison, L., Fenlon, D., Low, C., Mellor, D. J., Gunn, G. J., & Woolhouse, M. E. J. (2006), "Herd-level risk factors associated with the presence of Phage type 21/28 E-coli O157 on Scottish cattle farms", *Bmc Microbiology*, 6.

Hamdan, M. A. (1975), "Correlation between the numbers of two types of children when the family size distribution is zero-truncated negative binomial", *Biometrics*, 31(3): 765-769.

Handysides, S. (1999), "Underascertainment of infectious intestinal disease", *Communicable Disease and Public Health*, 2(2): 78-79.

Harrison, S. & Kinra, S. (2004), "Outbreak of *Escherichia coli* O157 associated with a busy bathing beach", *Communicable Disease and Public Health*, 7(1): 47-50.

Hawker, J., Begg, N., Blair, I., Reintjes, R., & Weinberg, J. (2001), *Communicable Disease Control Handbook*, Blackwell Science Ltd., London.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C. G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M., & Shinagawa, H. (2001), "Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12", *DNA Research*, 8(1): 11-22.

Health Canada (1999), "An Outbreak of *Escherichia coli* O157:H7 Infection Associated with Unpasteurized Non-commercial, Custom-pressed Apple Cider - Ontario, 1998", retrieved on 9 October 2004 from: http://www.phac-aspc.gc.ca/publicat/ccdr-rmtc/99vol25/dr2513e.html.

Health Canada (2000a), "Case Definitions for Diseases Under National Surveillance", *Canadian Communicable Disease Report*, 26(S3): 1-73.

Health Canada (2000b), "Case Definitions for Diseases Under National Surveillance", *Canadian Communicable Disease Report*, 26(S3): 1-73.

Health Protection Agency (1999), "VTEC O157 outbreak linked to beach holidays", *CDR Weekly*, 9: 327.

Health Protection Agency (2006a), "Surveillance of waterborne disease outbreaks summary of 2004", *CDR Weekly*, 16(15).

Health Protection Agency (2006b), "Vero cytotoxin-producing *Escherichia coli* O157: 2005", *CDR Weekly*, 16(23): 3-4.

Health Protection Scotland. (2004), General outbreaks of infectious intestinal disease - form. 2004.

Health Protection Scotland (2007a), "*Campylobacter*, Scotland, Annual Totals", retrieved on 18 November 2007a from: http://www.documents.hps.scot.nhs.uk/giz/10-year-tables/campy.pdf.

Health Protection Scotland (2007b), "*E. coli* O157, Scotland, Annual Totals", retrieved on 31 October 2007b.

Health Protection Scotland (2007c), "*E. coli* O157 cases - Grampian", *HPS Weekly Report*, 41(38).

Health Protection Scotland (2007d), "*E. coli* O157 cases - Paisley", *HPS Weekly Report*, 41(33).

Health Protection Scotland (2007e), "*Salmonella*, Scotland, Annual Totals", retrieved on 18 November 2007e from: http://www.documents.hps.scot.nhs.uk/giz/10-year-tables/salmonella.pdf.

Health Protection Scotland (2007f), "Surveillance Systems - Outbreaks of Infectious Intestinal Disease in Scotland", retrieved on 9 October 2007f from: http://www.hps.scot.nhs.uk/giz/ssdetail.aspx?id=131.

Hedberg, C. W., Angulo, F., Townes, J., Hadler, J., Vugia, D., Farley, M., & CDC/USDA/FDA Foodborne Diseases Active Surveillance Network (1997), "Differences in *Escherichia coli* O157:H7 annual incidence among FoodNet active surveillance sites", in 5th International VTEC Producing *Escherichia coli* Meeting, Baltimore, MD.

Heuvelink, A., Van Heerwaarden, C., Zwartkruisnahuis, J. T. M., Van Oosterom, R., Edink, K., Van Duynhoven, Y. T. H. P., & De Boer, E. (2002), "*Escherichia coli* O157 infection associated with a petting zoo", *Epidemiology and Infection*, 129: 295-302.

Hilborn, E. D., Mermin, J. H., Mshar, P., Hadler, J., Voetsch, A. C., Wojtkunski, C., Swartz, M., Mshar, R., Lambert-Fair, M.-A., Farrar, J. A., Glynn, M. K., & Slutsker, L. (1999), "A Multistate Outbreak of *Escherichia coli* O157:H7 Infections Associated With Consumption of Mesclun Lettuce", *Archives of Internal Medicine*, 159: 1758-1764.

Hildebrand, J. M., Maguire, H. C., Holliman, R. E., & Kangesu, E. (1996), "An outbreak of *Escherichia coli* O157 infection linked to paddling pools", *CDR Review*, 6(2): R33-R36.

Honish, L., Predy, G., Hislop, N., Chui, L., Kowalewska-Grochowska, K., Trottier, L., Kreplin, C., & Zazulak, I. (2005), "An Outbreak of *E. coli* O157:H7 Hemorrhagic

Colitis Associated with Unpasteurized Gouda Cheese", *Canadian Journal of Public Health*, 96(3): 182-184.

Honish, L., Zazulak, I., Mahabeer, R., Krywiak, K., Leyland, R., Hislop, N., & Chui, L. (2007), "Outbreak of *Escherichia coli* O157:H7 gastroenteritis associated with consumption of beef donairs, Edmonton, Alberta, May-June 2006", *Canadian Communicable Disease Report*, 33(2): 14-20.

Howie, H., Mukerjee, A., Cowden, J. M., Leith, J., & Reid, T. M. S. (2003), "Investigation of an outbreak of *Escherichia coli* O157 infection caused by environmental exposure at a scout camp", *Epidemiology and Infection*, 131: 1063-1069.

Iebin, B., Ison, A., Ombos, M., Oleszczuk, P., Hmed, R., & Amieson, F. "An *E. coli* O157:H7 Outbreak Associated with Consumption of Haggis", in CACMID 2003 Annual Conference.

Imhoff, B., Morse, D., Shiferaw, B., Hawkins, M., Vugia, D., Lance-Parker, S., Hadler, J., Medus, C., Kennedy, M. H., Moore, M. R., & Van Gilder, T. J. (2004), "Burden of Self-Reported Acute Diarrheal Illness in FoodNet Surveillance Areas, 1998 - 1999", *Clinical Infectious Diseases*, 38(S3): S219-S226.

Incident Control Team NHS Fife (2007), "The *E. coli* O157 outbreak at Careshare Nursery Lauder, Dunfermline in May 2006", retrieved on 18 October 2007 from: http://www.nhsfife.scot.nhs.uk/about_us/corporatedocuments.html.

Innocent, G. T., Mellor, D. J., McEwan, S. A., Reilly, W., Smallwood, J., Locking, M. E., Shaw, D. J., Michel, P., Taylor, D. J., Steele, W. B., Gunn, G. J., Ternent, H. E., Woolhouse, M. E. J., & Reid, S. W. J. (2006), "Spatial and temporal epidemiology of sporadic human cases of *Escherichia coli* O157 in Scotland, 1996 - 1999", *Epidemiology and Infection*, 133: 1033-1041.

Izsak, J. & Hunter, P. R. (1992), "Serotype abundance distributions in reports of *Salmonella* incidents in domestic livestock as indicators of the population biology of *Salmonella* infections", *Functional Ecology*, 6: 154-159.

Jensen, C., Ethelberg, S., Gervelmeyer, A., Nielsen, E. M., Olsen, K. E. P., & Molbak, K. (2006), "First general outbreak of Vercytotoxin-producing *Escherichia coli* O157 in Denmark", *Eurosurveillance Monthly*, 11(2).

Joe, S. & Kaplan, M. S. (2002), "Firearm-Related Suicide Among Young African-American Males", *Psychiatric Services*, 53(3): 332-334.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992), *Univariate Discrete Distributions*, John Wiley & Sons, Inc., New York.

Johnson, W. M., Lior, H., & Bezanson, G. S. (1983), "Cyto-Toxic *Escherichia coli* O157:H7 Associated with Hemorrhagic Colitis in Canada", *Lancet*, 1(8314): 76.

Jones, I. G. & Roworth, M. (1996), "An outbreak of *Escherichia coli* 0157 and campylobacteriosisassociated with contamination of a drinking water supply", *Public Health*, 110(5): 277-282.

Jones, T., Imhoff, B., Samuel, M., Mshar, P., McCombs, K., Hawkins, M., Deneen, V., Cambridge, M., & Olsen, S. J. (2004), "Limitations to Successful Investigation and Reporting of Foodborne Outbreaks: An Analysis of Foodborne Disease Outbreaks in FoodNet Catchment Areas, 1998 - 1999", *Clinical Infectious Diseases*, 38(Suppl 3): S297-S302.

Joubert, J., Norman, R., Lambert, E. V., Groenewald, P., Schneider, M., Bull, F., & Bradshaw, D. (2007), "Estimating the burden of disease attributable to physical inactivity in South Africa in 2000", *Samj South African Medical Journal*, 97(8): 725-731.

Kamper-Jorgensen, M., Wohlfahrt, J., Simonsen, J., Thrane, N., & Benn, C. S. (2006), "Temporal trend in paediatric infections in Denmark", *Archives of Disease in Childhood*, 91(5): 401-404.

Kaper, J. B., Nataro, J. P., & Mobley, H. L. (2004), "Pathogenic *Escherichia coli*", *Nature Reviews*, Microbiology.(2): 123-140.

Karch, H., Bielaszewska, M., Bitzan, M., & Schmidt, H. (1999), "Epidemiology and Diagnosis of Shiga Toxin-Producing *Escherichia coli* Infections", *DIagnostic Microbiology and Infectious Disease*, 34(3): 229-243.

Karch, H., Rüssmann, H., Schmidt, H., Schwarzkopf, A., & Heesemann, J. (1995), "Long-Term Shedding and Clonal Turnover of Enterohemorrhagic *Escherichia coli* O157 in Diarrheal Diseases", *Journal of Clinical Microbiology*, 33(6): 1602-1605.

Karch, H., Tarr, P. I., & Bielaszewska, M. (2005), "Enterohaemorrhagic *Escherichia coli* in human medicine", *International Journal of Medical Microbiology*, 295(6-7): 405-418.

Karmali, M. A. (2004), "Infection by Shiga toxin-producing *Escherichia coli*: an overview", *Molecular Biotechnology*, 26(2): 117-122.

Karmali, M., Petric, M., Lim, C., Fleming, P. C., & Steele, B. T. (1983), "*Escherichia coli* cytotoxin, haemolytic-uraemic syndrome, and haemorrhagic colitis.", *Lancet*, 2(8362): 1299-1300.

Keene, W. E., Mcanulty, J. M., Hoesly, F. C., Williams, L. P., Hedberg, K., Oxman, G. L., Barrett, T. J., Pfaller, M. A., & Fleming, D. W. (1994), "A Swimming-Associated Outbreak of Hemorrhagic Colitis Caused by *Escherichia coli* O157:H7 and Shigella-Sonnei", *New England Journal of Medicine*, 331(9): 579-584.

Keene, W. E., Hedberg, K., Herriott, D. E., Hancock, D. D., McKay, R. W., Barrett, T., & Fleming, D. W. (1997), "A Prolonged Outbreak of *Escherichia coli* O157:H7 Infections Caused by Commercially Distributed Raw Milk", *The Journal of Infectious Diseases*, 176: 815-818.

Kendal, W. S. (2007), "The number distribution for involved lymph nodes in cancer", *Mathematical Biosciences*, 205(1): 32-43.

Khakhria, R., Duck, D., & Lior, H. (1990), "Extended phage-typing scheme for *Escherichia coli* O157:H7", *Epidemiology and Infection*, 105: 511-520.

Khakhria, R., Woodward, D., Johnson, W. M., & Poppe, C. (1997), "*Salmonella* isolated from humans, animals and other sources in Canada, 1983-92", *Epidemiology and Infection*, 119(1): 15-23.

Kirkwood, B. R. & Sterne, J. A. C. (2003), *Medical Statistics*, Blackwell Science Ltd, Oxford, UK.

Koch, A. L. (1975), "The kinetics of mycelial growth", *Journal of General Microbiology*, 89(2): 209-216.

Kohli, H. S., Chaudhuri, A. K. R., Todd, W. T. A., Mitchell, A. A. B., & Liddell, K. G. (1994), "A Severe Outbreak of *Escherichia coli* O157 in 2 Psychogeriatric Wards", *Journal of Public Health Medicine*, 16(1): 11-15.

Kuczius, T., Bielaszewska, M., Friedrich, A. W., & Zhang, W.-L. (2004), "A rapid method for the discrimination of genes encoding classical Shiga toxin (Stx) 1 and its variants, Stx1c and Stx1d, in *Escherichia coli*", *Molecular Nutrition and Food Research*, 48: 515-521.

Kudva, I. T., Hatfield, P. G., & Hovde, C. J. (1996), "*Escherichia coli* O157:H7 in microbial flora of sheep", *Journal of Clinical Microbiology*, 34(2): 431-433.

Kuhn, L., Davidson, L. L., & Durkin, M. S. (1994), "Use of Poisson Regression and Time Series Analysis for Detecting of Changes over Time in Rates of Child Injury following a Prevention Program", *American Journal of Epidemiology*, 140(10): 943-955.

Levine, W., Smart, J. F., Archer, D. L., Bean, N., & Tauxe, R. V. (1991), "Foodborne Disease Outbreaksin Nursing Homes, 1975 Through 1987", *Journal of the American Medical Association*, 266(15): 2105-2109.

Licence, K., Oates, K. R., Synge, B. A., & Reid, T. M. S. (2001), "An outbreak of *E. coli* O157 infection with evidence of spread from animals to man through contamination of a private water supply", *Epidemiology and Infection*, 126: 135-138.

Lior, H. (1983), "Hemorrhagic Colitis in a Home for the Aged, Ontario", *Canada Diseases Weekly Report*, 9(8): 29-32.

Locking, M. E., Allison, L. J., Rae, L., & Hanson, M. (2003a), "VTEC in Scotland 2002: enhanced surveillance and Reference Laboratory data", *SCIEH Weekly Report*, 37(2003/49): 304-307.

Locking, M. E., Allison, L. J., Rae, L., & Hanson, M. (2004), "VTEC in Scotland 2003: enhanced surveillance and reference laboratory data", *SCIEH Weekly Report*, 38(2004/49): 294-297.

Locking, M. E., Allison, L. J., Rae, L., Pollock, K., & Hanson, M. (2006a), "VTEC in Scotland 2004: Enhanced surveillance and reference laboratory data", *HPS Weekly Report*, 39(51-52): 290-295.

Locking, M. E., O'Brien, S., Reilly, W., Wright, E., Campbell, D. M., Coia, J. E., Browning, L. M., & Ramsay, C. N. (2001), "Risk Factors for sporadic cases of *Escherichia coli* O157 infection: the importance of contact with animal excreta", *Epidemiology and Infection*, 127: 215-220.

Locking, M. E., Reilly, W., & Rae, L. (2003b), Improving our understanding: enhanced surveillance of *E. coli* O157 and health outcomes in Scotland 1999-2002. VTEC 2003 . 2003b.

Locking, M. E., Smith-Palmer, A., Cowden, J. M., Allison, L. J., Rae, L., Hanson, M., & Reilly, W. "Outbreaks of *E. coli* O157 infection in Scotland, 1996-2006: Does a decade make a difference?", in VTEC 2006 Abstracts.

Locking, M. E., Smith-Palmer, A., Cowden, J. M., & Reilly, W. (2003c), "From Food to Farm: The Changing Profile of *Escherichia coli* O157 Outbreaks in Scotland, 1996 to 2002", in Abstracts of the 5th International Symposium on VTEC, p. 206.

Lynch, M., Painter, J., Woodruff, R., & Braden, C. (2006), "Surveillance for Foodborne-Disease Outbreaks --- United States, 1998--2002", *MMWR Weekly*, 55(SS10): 1-34.

MacDonald, A. R., Gould, I. M., & Curnow, J. (1996), "Epidemiology of infection due to *Escherichia coli O157*: 3-year prospective study", *Epidemiology and Infection*, 116: 279-284.

MacDonald, C., Drew, J., Carlson, R., Dzogan, S., Tataryn, S., MacDonald, A. R., Ali, A., Amhed, R., Easy, R., Clark, C., & Rodgers, F. (2000), "Outbreak of *Escherichia coli* O157:H7 leading to the recall of retail ground beef - Winnipeg, Manitba, May 1999", *Canadian Communicable Disease Report*, 26(13).

MacDonald, D. M., Fyfe, M., Paccagnella, A. M., Trinidad, A., Louie, K., & Patrick, D. (2004), "*Escherichia coli* O157:H7 outbreak linked to salami, British Columbia, Canada, 1999", *Epidemiology and Infection*, 132: 283-289.

MacDougall, L., Majowicz, S. E., Dore, K. A., Flint, J. A., Thomas, K., Kovacs, S., & Sockett, P. N. (2007), "Under-reporting of infectious gastrointestinal illness in British Columbia, Canada: who is counted in provincial communicable disease statistics", *Epidemiology and Infection*, Forthcoming article.

Magurran, A. E. & Henderson, P. A. (2003), "Explaining the excess of rare species in natural species abundance distributions", *Nature*, 422: 714-716.

Majowicz, S. E., Edge, V. L., Fazil, A., McNab, W. B., Dore, K. A., Sockett, P. N., Flint, J. A., Middleton, D., McEwen, S. A., & Wilson, J. B. (2005), "Estimating the Under-reporting Rate for Infectious Gastrointestinal Illness in Ontario", *Canadian Journal of Public Health*, 96(3): 178-181.

Marcus,R., Gerber,D.E., Smith,K.E., Keene,W.E., Holtry,R., Vugia,D., Chaves,S., Hoefer,D., and Moore,M.R. (2004a), "The Influence of Outbreak Cases on Trends in *E. coli* O157 Infection, FoodNet Sites, 1996 - 2002", retrieved on 1 November 2007a.

Marcus, R., Rabatsky-Ehr, T., Mohle-Boetani, J. C., Farley, M., Medus, C., Shiferaw, B., Carter, M., Zansky, S., Kennedy, M., Van Gilder, T., & Hadler, J. L. (2004b), "Dramatic decrease in the incidence of Salmonella serotype enteritidis infections in 5 FoodNet sites: 1996-1999", *Clinical Infectious Diseases*, 38: S135-S141.

Marsh, J., MacLeod, A. F., Hanson, M., Emmanuel, F. X. S., Frost, J., & Thomas, A. (1992), "A restaurant-associated outbreak of *E. coli* O157 infection", *Public Health*, 14(1): 78-83.

Marshall, J. K., Thabane, M., Garg, A. X., Clark, W. F., Salvadori, M., & Collins, S. M. (2006), "Incidence and Epidemiology of Irritable Bowel Syndrome After a Large Waterborne Outbreak of Bacterial Dysentery", *Gastroenterology*, 131(2): 445-450.

Martin, D. C. & Katti, S. K. (1965), "Fitting of Certain Contagious Distributions to Some Available Data by Maximum Likelihood Method", *Biometrics*, 21(1): 34-48.

Maruzumi, M., Morita, M., Matsuoka, Y., Uekawa, A., Nakamura, T., & Fuji, K. (2005), "Mass Food Poisoning by Beef Offal Contaminated by *Escherichia coli* O157", *Japanese Journal of Infectious Diseases*, 58: 397.

Matthews, L., Low, J. C., Gally, D. L., Pearce, M. C., Mellor, D. J., Heesterbeek, J. A. P., Chase-Topping, M., Naylor, S. W., Shaw, D. J., Reid, S. W. J., Gunn, G. J., & Woolhouse, M. E. J. (2006a), "Heterogeneous shedding of *Escherichia coli O157* in cattle and its implications for control", *Proceedings of the National Academy of Sciences of the United States of America*, 103(3): 547-552.

Matthews, L., McKendrick, I. J., Ternent, H., Gunn, G. J., Synge, B., & Woolhouse, M. E. (2006b), "Super-shedding cattle and the transmission dynamics of *Escherichia coli* O157", *Epidemiology & Infection*, 134(1): 131-142.

Maule, A. (1999), "Environmental aspects of *E. coli* O157", *International Food Hygiene*, 9: 21-23.

Maule, M. M., Zuccolo, L., Magnani, C., Pastore, G., Dalmasso, P., Pearce, N., Merletti, F., & Gregori, D. (2006), "Bayesian methods for early detection of changes in childhood cancer incidence: Trends for acute lymphoblastic leukaemia are consistent with an infectious aetiology", *European Journal of Cancer*, 42(1): 78-83.

McCall, B., Strain, D., Hills, S., Heymer, M., Bates, J., Murphy, D., Kelly, R., & Price, D. (1996), "An Outbreak of *Escherichia coli* O157 Infection on the Gold Coast", *Communicable Diseases Intelligence*, 20(10): 236-239.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Chapman & Hall.

McDonald, L. C., Banerjee, S. N., Jarvis, W. R., & National Nosocomial Infections Surveillance System (1999), "Seasonal Variation of Acinetobacter Infections: 1987 - 1996", *Clinical Infectious Diseases*, 29: 1133-1137.

McDonnell, R., Rampling, A., Crook, S., Cockcroft, P., Willshaw, G., Cheasty, T., & Stuart, J. (1997), "An outbreak of Vero cytotoxin producing *Escherichia coli* O157 infection associated with takeaway sandwiches", *CDR Review*, 7(13): R201-R205.

McIntyre, L., Fung, J., Paccagnella, A. M., Isaac-Renton, J., Rockwell, F., Emerson, B., & Preston, T. (2002), "*Escherichia coli* O157 Outbreak Associated with the Ingestion of Unpasteurized Goat's Milk in British Columbia, 2001", *Canadian Communicable Disease Report*, 28(01).

Mead, P. S., Finelli, L., Lambert-Fair, M.-A., Champ, D., Townes, J., Hutwagner, L., Barrett, T., Spitalny, K., & Mintz, E. (1997), "Risk Factors for Sporadic Infection with *Escherichia coli* O157:H7", *Archives of Internal Medicine*, 157: 204-208.

Mead, P. S. & Griffin, P. M. (1998), "*Escherichia coli* O157:H7", *The Lancet*, 352: 1207-1212.

Mead, P. S., Slutsker, L., Dietz, V., McCaig, L., Bresee, J., Shapiro, C., Griffin, P. M., & Tauxe, R. V. (1999), "Food-Related Illness and Death in the United States", *Emerging Infectious Diseases*, 5(5): 607-625.

Melton-Celsa,A. & O'Brien,A. (1998) "Structure, Biology and Relative Toxicity of Shiga Toxin Family Members for Cells and Animals," in *Escherichia coli O157:H7 and Other Shiga Toxin-Producing E. coli Strains*, J. Kaper & A. O'Brien, eds., ASM Press: Washington, DC, pp. 121-128.

Menzies, F. D., Neill, S. D., Goodall, E. A., & Mcilroy, S. G. (1994), "Avian *Salmonella* Infections in Northern-Ireland, 1979-1991", *Preventive Veterinary Medicine*, 19(2): 119-128.

Michel, P., Wilson, J. B., Martin, S. W., Clarke, R. C., McEwan, S. A., & Gyles, C. L. (2000), "Estimation of the under-reporting rate for the surveillance of *Escherichia coli* O157:H7 cases in Ontario, Canada", *Epidemiology and Infection*, 125: 35-45.

Michel, P., Wilson, J. B., Martin, S. W., Clarke, R. C., McEwen, S. A., & Gyles, C. L. (1998), "A Descriptive Study of Verocytotoxigenic *Escherichia coli* (VTEC) Cases Reported in Ontario, 1990-1994", *Canadian Journal of Public Health*, 89(4): 253-257.

Michel, P., Wilson, J. B., Martin, S. W., McEwen, S. A., & Gyles, C. L. (1999), "Temporal and geographical distributions of reported cases of *Escherichia coli* O157:H7 infection in Ontario", *Epidemiology and Infection*, 122: 193-200.

Michino, H., Araki, K., Minami, S., Takaya, S., Sakai, N., Miyazaki, M., Ono, A., & Yanagawa, H. (1999), "Massive outbreak of *Escherichia coli* O157 : H7 infection in schoolchildren in Sakai City, Japan, associated with consumption of white radish sprouts", *American Journal of Epidemiology*, 150(8): 787-796.

Michino,H., Araki,K., Minami,S., Nakayama,T., Ejima,Y., Hiroe,K., Tanaka,H., Fujita,N., Usami,S., Yonekawa,M., Sadamoto,K., Takaya,S., & Sakai,N. (1998) "Recent Outbreaks of Infections Caused by *Escherichia coli* O157:H7 in Japan," in *Escherichia coli O157:H7 and Other Shiga Toxin-Producing E. coli Strains*, J. Kaper & A. O'Brien, eds., ASM Press: Washington DC, pp. 73-82.

Miller, G., Dunn, G. M., Smith-Palmer, A., Ogden, I. D., & Strachan, N. J. C. (2004a), "Human campylobacteriosis in Scotland: seasonality, regional trends and bursts of infection", *Epidemiology and Infection*, 132(4): 585-593.

Miller, G., Dunn, G. M., Smith-Palmer, A., Ogden, I. D., & Strachan, N. J. C. (2004b), "Human campylobacteriosis in Scotland: seasonality, regional trends and bursts of infection", *Epidemiology and Infection*, 132: 585-593.

Monnens,L.A., Savage,C.O., & Taylor,C. (1998) "Pathophysiology of Hemolyic-Uremic Syndrome," in *Escherichia coli O157:H7 and Other Shiga Toxin-Producing E. coli Strains*, J. B. Kaper & A. D. O'Brien, eds., American Society for Microbiology: Washington, DC, pp. 287-292.

Morgan, G. M., Newman, C., Palmer, S. R., Allen, J. B., Shepherd, W., Rampling, A., Warren, R. E., Gross, R. J., Scotland, S. M., & Smith, H. R. (1988), "First recognized community outbreak of haemorrhagic colitis due to verotoxin-producing *Escherichia coli* O157.H7 in the UK", *Epidemiology and Infection*, 101: 83-91.

Moura, M. F., Picanco, M. C., Guedes, R. N. C., Barros, E. C., Chediak, M., & Morais, E. G. F. (2007), "Conventional sampling plan for the green leafhopper *Empoasca kraemeri* in common beans", *Journal of Applied Entomology*, 131(3): 215-220.

Moxley,R.A. & Francis,D.H. (1998) "Overview of Animal Models," in *Escherichia coli O157:H7 and Other Shiga Toxin-Producing E. coli Strains*, J. B. Kaper & A. D. O'Brien, eds., ASM Press: Washington, D.C., pp. 249-260.

Murray, P. R., Rosenthal, K. S., Kobayashi, G. S., & Pfaller, M. A. (2002), *Medical Microbiology*, Mosby, London.

Myers, C., Farrar, J. A., & Waddell, J. M. (2002) *Final Report - E. coli O157:H7 Illnesses in Washington - July, 2002*, State of California - Health and Human Serives Agency, Department of Health Services, Food and Drug Branch.

Nataro, J. P. & Kaper, J. B. (1998), "Diarrheagenic *Escherichia coli*", *Clinical Microbiology Reviews*, 11(1): 142-201.

National Institute of Infectious Diseases (1996), "Table 1: Outbreaks of VTEC O157:H7 infection, 1996", retrieved on 13 December 2005 from: http://idsc.nih.go.jp/iasr/18/209/graph/t2091.gif.

National Laboratory for Enteric Pathogens (2002), "Laboratory Surveillance Data for Enteric Pathogens in Canada", retrieved on 2 August 2007 from: http://www.nml-lnm.gc.ca/english/PDF/1999AnnualSummary.pdf.

National Laboratory for Enteric Pathogens (2006), "Laboratory Surveillance Data For Enteric Pathogens in Canada", retrieved on 3 August 2007 from: http://www.nml-lnm.gc.ca/english/PDF/2004AnnualSummary.pdf.

Naylor, S. W., Gally, D. L., & Low, J. C. (2005), "Enterohaemorrhagic *E. coli* in veterinary medicine", *International Journal of Medical Microbiology*, 295(6-7): 419-441.

New York State Department of Health (2000), "Health Commissioner Releases *E. coli* Outbreak Report" from: http://www.health.state.ny.us/press/releases/2000/ecoli.htm.

Neyman, J. (1939), "On a new class of "contagious" distributions, applicable in entomology and bacteriology", *Annals of Mathematical Statistics*, 10: 35-57.

NHS Lothians. (2005), "Scottish *E. coli* O157 Reference Laboratory User Manual", Edinburgh, NHS Lothians.

Noël, J. M. & Boedeker, E. C. (1997), "Enterohemorrhagic *Escherichia coli*: a family of emerging pathogens", *Digestive Diseases*, 15(1-2): 67-91.

Norman, R., Bradshaw, D., Steyn, K., & Gaziano, T. (2007), "Estimating the burden of disease attributable to high cholesterol in South Africa in 2000", *Samj South African Medical Journal*, 97(8): 708-715.

O'Brien, A. O., Lively, T. A., Chen, M. E., Rothman, S. W., & Formal, S. B. (1983), "*Escherichia coli* O157:H7 strains associated with haemorrhagic colitis in the United States produce a Shigella dysenteriae 1 (SHIGA) like cytotoxin", *Lancet*, 1(8326): 702.

O'Brien, S. & Adak, G. (2002), "*Escherichia coli* O157:H7 - Piecing together the jigsaw puzzle", *New England Journal of Medicine*, 347(8): 608-609.

O'Brien, S., Adak, G., & Gilham, C. (2001a), "Contact with Farming Environment as a Major Risk Factor for Shiga Toxin (Vero Cytotoxin)-Producing *Escherichia coli* O157 Infection in Humans", *Emerging Infectious Diseases*, 7(6).

O'Brien, S., Adak, G., & Reilly, W. (2001b), "The Task Force on *E. coli* O157 Final Report: the view from here", *Communicable Disease and Public Health*, 4(3): 154-156.

O'Brien, S., Murdoch, P. S., Riley, A., King, I., Barr, M., Murdoch, S., Greig, A., Main, R., Reilly, W., & Thomson-Carter, F. (2001c), "A foodborne outbreak of Vero cytotoxin-producing *Escherichia coli* O157:H-phage type 8 in hospital", *Journal of Hospital Infection*, 49: 167-172.

O'Connor,D. (2002), "Report of the Walkerton Commission of Inquiry", retrieved on 14 December 2005 from: http://www.attorneygeneral.jus.gov.on.ca/english/about/pubs/walkerton/part1/.

O'Donnell, J. M., Thornton, L., McNamara, E. B., Prendergast, T., Igoe, D., & Cosgrove, C. (2002), "Outbreak of Vero cytotoxin-producing *Escherichia coli* O157 in a child day care facility", *Communicable Disease and Public Health*, 5(1): 54-58.

O'Neill, M. F. & Faddy, M. J. (2003), "Use of binary and truncated negative binomial modelling in the analysis of recreational catch data", *Fisheries Research*, 2: 471-477.

Office of the Chief Medical Officer (2005), "South Wales *E. coli* Outbreak - September 2005: A Report Commissioned by the Chief Medical Officer", retrieved on 6 September 2007 from: www.cmo.wales.gov.uk/content/ publications/reports/e-coli-e.pdf.

Olsen, S. J., Bishop, R., Brenner, F. W., Roels, T. H., Bean, N., Tauxe, R. V., & Slutsker, L. (2001), "The Changing Epidemiology of *Salmonella*: Trends in Serotypes Isolated from Humans in the United States, 1987-1997", *The Journal of Infectious Diseases*, 183: 753-761.

Orskov, F. & Orskov, I. (1992), "*Escherichia coli* serotyping and disease in man and animals", *Canadian Journal of Microbiology*, 38(7): 699-704.

Ostroff, S. M., Griffin, P. M., Tauxe, R. V., Shipman, L. D., Greene, K. D., Wells, J. G., Lewis, J. H., Blake, P. A., & Kobayashi, J. M. (1990), "A Statewide Outbreak of *Escherichia coli* O157:H7 Infections in Washington-State", *American Journal of Epidemiology*, 132(2): 239-247.

Ostroff, S. M., Tarr, P. I., Neill, M., Lewis, J. H., Hargrett-Bean, N., & Kobayashi, J. (1989), "Toxin Genotypes and Plasmid Profiles as Determinants of Systemic Sequelae in *Escherichia coli* O157:H7 Infections", *The Journal of Infectious Diseases*, 160(6): 994-998.

Outbreak Control Team - NHS Wales (2007) *Outbreak of Verotoxin positive Escherichia coli O157 Infection in South Wales - October 2005*, Health of Wales Information Service.

Paiba, G. A., Gibbens, J. C., Pascoe, S. J. S., Wilesmith, J. W., Kidd, S. A., Byrne, C., Ryan, J. B. M., Smith, R. R., McLaren, I. M., Futter, R. J., Kay, A. C. S., Jones,

Y. E., Chappell, S. A., Willshaw, G. A., & Cheasty, T. (2002), "Faecal carriage of verocytotoxin-producing *Escherichia coli* O157 in cattle and sheep at slaughter in Great Britain", *Veterinary Record*, 150(19): 593-+.

Palumbo,M.S., Sigl,J., Farrar,J.A., and Waddell,J.M. (2004), "Investigation of *E. coli* O157:H7 Outbreak at San Mateo County Retirement Facility" from: http://www.dhs.ca.gov/ps/fdb/local/PDF/2006 Spinach Report Final redacted.PDF.

Parry, S. M. & Palmer, S. R. (2005), "The public health significance of VTEC O157", *Journal of Applied Microbiology Symposium Supplement*, 88: 1S-9S.

Parry, S. M. & Salmon, R. L. (1998), "Sporadic STEC O157 infection: Secondary household transmission in Wales", *Emerging Infectious Diseases*, 4(4): 657-661.

Parry, S. M., Salmon, R. L., Willshaw, G. A., & Cheasty, T. (1998), "Risk Factors for and prevention of sporadic infections with vero cytotoxin (shiga toxin) producing *Escherichia coli* O157", *The Lancet*, 351: 1019-1022.

Paunio, M., Pebody, R., Keskimaki, M., Kokki, M., Ruutu, P., Oinonen, S., Vuotari, V., Siitonen, A., Lahti, E., & Leinikki, P. (1999), "Swimming-associated outbreak of *Escherichia coli* O157:H7", *Epidemiology and Infection*, 122(1): 1-5.

Pavia, A. T., Nichols, C. R., Green, D. P., Tauxe, R. V., Mottice, S., Greene, K. D., Wells, J. G., Siegler, R. L., Brewer, E. D., Hannon, D., & Blake, P. A. (1990), "Hemolytic-Uremic Syndrome During An Outbreak of *Escherichia coli* O157:H7 Infections in Institutions for Mentally-Retarded Persons - Clinical and Epidemiologic Observations", *Journal of Pediatrics*, 116(4): 544-551.

Payne, C. J. I., Petrovic, M., Roberts, R. J., Paul, A., Linnane, E., Walker, M., Kirby, D., Burgess, A., Smith, R. M. M., Cheasty, T., Willshaw, G., & Salmon, R. L. (2003), "Vero cytotoxin-producing *Escherichia coli* O157 gastroenteritis in farm visitors, North Wales", *Emerging Infectious Diseases*, 9(5): 526-530.

Pearl, D. L., Louie, M., Chui, L., Dore, K., Grinisrud, K. M., Leedell, D., Martin, S. W., Michel, P., Svenson, L. W., & McEwen, S. A. (2006), "The use of outbreak information in the interpretation of clustering of reported cases of *Escherichia coli O157* in space and time in Alberta, Canada, 2000-2002", *Epidemiology and Infection*, 134(4): 699-711.

Pebody, R. G., Furtado, C., Rojas, A., McCarthy, N., Nylen, G., Ruutu, P., Leino, T., Chalmers, R., deJong, B., Donnelly, M., Fisher, I., Gilham, C., Graverson, L., Cheasty, T., Willshaw, G., Navarro, M., Salmon, R., Leinikki, P., Wall, P., & Bartlett, C. (1999), "An international outbreak of Vero cytotoxin-producing *Escherichia coli* O157 infection amongst tourists; a challenge for the European infectious disease surveillance network", *Epidemiology and Infection*, 123: 217-223.

Pennington, T. H. (1998), "Factors involved in recent outbreaks of *Escherichia coli* O157:H7 in Scotland and recommendations for its control", *Journal of Food Safety*, 18: 383-391.

Petrie, A. & Sabin, C. (2000), *Medical Statistics at a Glance*, Blackwell Science, Padstow, UK.

Pickering, L. K., Bartlett, A. V., & Woodward, W. E. (1986), "Acute Infectious Diarrhea Among Children in Day Care: Epidemiology and Control", *Reviews of Infectious Diseases*, 8(4): 539-547.

Poisson, S. D. (1837) *Recherches sur la Probabilité des Jugements en Matiére Criminelle et en Matiére Civile, Précédées des Regles Générales du Calcul des Probabilitiés*, Bachelier, Imprimeur-Libraire pour les Mathematiques, la Physique, etc..

Prattley, D. J., Cannon, R. M., Wilesmith, J. W., Morris, R. S., & Stevenson, M. A. (2007), "A model (BSurvE) for estimating the prevalence of bovine spongiform encephalopathy in a national herd", *Preventive Veterinary Medicine*, 80(4): 330-343.

Proctor, M. E. (2000a) *Investigation of an Outbreak of E. coli O157:H7 Infection at the Layton Avenue Sizzler Restaurant Associated with Cross-Contamination of Watermelon with Raw Meat, Milwaukee, WI; July - August, 2000*, Communicable Disease Epidemiology Section, Bureau of Communicable Diseases, Wisconsin Division of Public Health, Department of Health and Family Services (Wisconsin).

Proctor, M. E. (2000b) *Investigation of an Outbreak of Gastrointestinal Illness at the Mayfair Road Sizzler Restaurant, Wauwatosa, WI: July - August 2000*, Department of Health and Family Servicesm, Wisconsin Division of Public Health.

ProMED-mail (2007), "*E. coli* O157, restaurant - Canada (ON)(03)", *ProMED-mail*, 20 June: 20070620.1989.

Province, M. A., Tishler, P., Rao, D. C., & Eaves, L. J. (1989), "Repeated-measures model for the investigation of temporal trends using longitudinal family studies: Application to systolic blood pressure", *Genetic Epidemiology*, 6(2): 333-347.

Public Health Agency of Canada (2005), "Notifiable Diseases Monthly Report", retrieved on 30 October 2007 from: http://www.phac-aspc.gc.ca/bid-bmi/dsd-dsm/ndmr-rmmdo/index.html.

Public Health Agency of Canada (2006), "Notifiable Diseases On-Line" from: http://dsol-smed.phac-aspc.gc.ca/dsol-smed/ndis/c_time_e.html.

Public Health Agency of Canada (2007), "Surveillance" from: http://www.phac-aspc.gc.ca/surveillance_e.html.

Public Health Laboratory Service (1995), "An outbreak of *Escherichia coli* O157 in Southern England", *CDR Weekly*, 5(22): 103.

Public Health Laboratory Service (1997), "Two outbreaks of Vero cytotoxin producing *Escherichia coli* O157 infection associated with farms", *CDR Weekly*, 7(30): 263-266.

Public Health Laboratory Service (1998a), "Cases of *Escherichia coli* O157 infection associated with unpasteurised cream", *CDR Weekly*, 8(43): 377.

Public Health Laboratory Service (1998b), "Outbreak of Vero cytotoxin producing *Escherichia coli* O157 infection in Dorset", *CDR Weekly*, 8(21): 183-186.

Public Health Laboratory Service (1999a), "*Escherichia coli* O157 associated with eating unpasteurised cheese", *CDR Weekly*, 9(13): 113-116.

Public Health Laboratory Service (1999b), "*Escherichia coli* O157 associated with eating unpasteurised cheese - update", *CDR Weekly*, 9(15): 131-134.

Public Health Laboratory Service (1999c), "Outbreak of VTEC O157 infection at a prison in the Midlands", *CDR Weekly*, 9(32): 281-284.

Public Health Laboratory Service (1999d), "Outbreak of VTEC O157 infection in East Sussex", *CDR Weekly*, 9(25): 219-222.

Public Health Laboratory Service (1999e), "VTEC O157 outbreak associated with a farm visitor centre in North Wales", *CDR Weekly*, 9(26): 227-230.

Public Health Laboratory Service (2000a), "Outbreak of Vero cytotoxin-producing *Escherichia coli* O157 infection in a children's nursery in Suffolk", *CDR Weekly*, 11(26).

Public Health Laboratory Service (2000b), "Outbreaks of VTEC O157 infection linked to comsumption of unpasteurised milk", *CDR Weekly*, 10(23): 203-206.

Public Health Laboratory Service (2000c), "Two outbreaks of VTEC O157 infection in northern England", *CDR Weekly*, 10(26): 229.

Public Health Laboratory Service (2002), "Surveillance of waterborne disease and water quality: July to December 2001", *CDR Weekly*, 12(11).

Public Health Laboratory Service (2006), "Two contiguous but unconnected outbreaks of Vero cytotoxin-Producing *E. coli* O157 Infection in South East London", *CDR Weekly*, 15(39).

R Development Core Team. (2007), R: A Language and Environment for Statistical Computing. 2007. Vienna, Austria, R Foundation for Statistical Computing.

Rangel, J. M., Sparling, P. H., Crowe, C., Griffin, P. M., & Swerdlow, D. (2005), "Epidemiology of *Escherichia coli* O157:H7 Outbreaks, United States, 1982-2002", *Emerging Infectious Diseases*, 11(4).

Reida, P., Wolff, M., Pohls, H. W., Kuhlmann, W., Lehmacher, A., Aleksi, S., Karch, H., & Bockemuhl, J. (1994), "An outbreak due to enterohaemorrhagic *Escherichia coli* O157:H7 in a children day care centre characterized by person-to-person transmission and environmental contamination", *Zentralblatt fur Bakteriologie*, 281(4): 534-543.

Reilly, W. (1997), "*E. coli* O157 in Scotland-An Overview", *SCIEH Weekly Report*, 1(97/13): 4-5.

Reiss, G., Kunz, P., Koin, D., & Keeffe, E. B. (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.

Renter, D. & Sargeant, J. (2002), "Enterohemorrhagic *Escherichia coli* O157: epidemiology and ecology in bovine production environments", *Animal Heath Research Reviews*: 83-94.

Riley, L., Remis, R., & Helgerson, S. (1983), "Hemorrhagic colitis associated with a rare *Escherichia coli* serotype", *New England Journal of Medicine*, 308: 681-685.

Roberts, J. A. (2000), "Economic aspects of food-borne outbreaks and their control", *British Medical Bulletin*, 56(1): 133-141.

Roberts, J. A., Upton, P. A., & Azene, G. (2000), "*Escherichia coli* O157:H7; an economic assessment of an outbreak", *Journal of Public Health Medicine*, 22(1): 99-107.

Roberts, J. A. & Upton, P. A. (2000), *E. coli O157: An Economic Assessment of an Outbreak*, Lothian Health Board.

Rodriguez, C., Naves, J., Fernandez-Gil, A., Obeso, J. R., & Delibes, M. (2007), "Long-term trends in food habits of a relict brown bear population in northern Spain: the influence of climate and local factors", *Environmental Conservation*, 34(1): 36-44.

Rota, M. C., Cawthorne, A., Bella, A., Caporali, M. G., Filia, A., & D'Ancona, F. (2007), "Capture-recapture estimation of underreporting of legionellosis cases to the National Legionellosis Register: Italy 2002", *Epidemiology and Infection*, 135(6): 1030-1036.

Sakuma, M., Urashima, M., & Okabe, N. (2006), "Verocytotoxin-producing *Escherichia coli*, Japan, 1999-2004", *Emerging Infectious Diseases*, 12(2).

Samadpour, M., Stewart, J., Steingart, K., Addy, C., Louderback, J., McGinn, M., Ellington, J., & Newman, T. (2002a), "Laboratory Investigation of an *E. coli* O157:H7 OUtbreak Associated with Swimming in Battle Ground Lake, Vancouver, Washington", *Journal of Environmental Health*, 64(10): 16-20.

Samadpour, M., Stewart, J., Steingart, K., Addy, C., Louderback, J., McGinn, M., Ellington, J., & Newman, T. (2002b), "Laboratory Investigation of an *E. coli* O157:H7 Outbreak Associated with Swimming in Battle Ground Lake, Vancouver, Washington", *Journal of Environmental Health*, 64(10): 16-20.

Sampford, M. R. (1955), "The Truncated Negative Binomial Distribution", *Biometrika*, 42(1-2): 58-69.

San Mateo County Health Services Agency (2004), "Outbreak Investigation of *Escherichia coli* O157:H7", *Epidemiological Bulletin - San Mateo County Health Services Agency*, 3(1): 6-7.

Sandberg, M., Nygard, K., Meldal, H., Valle, P. S., Kruse, H., & Skjerve, E. (2006), "Incidence trend and risk factors for *campylobacter* infections in humans in Norway", *Bmc Public Health*, 6.

Sartz, L., deJong, B., Hjertqvist, M., Plym-Forshell, L., Alsterlund, R., Löfdahl, S., Osterman, B., Ståhl, A., Eriksson, E., Hansson, H.-B., & Karpman, D. (2007), "An outbreak of *Escherichia coli* O157:H7 infection in southern Sweden associated with consumption of fermented sausage; aspects of sausage production that increase the risk of contamination", *Epidemiology and Infection*, Forthcoming article.

SAS Institute Inc. (2003), SAS. [9.1]. 2003. Cary, NC, SAS Institute, Inc.

Schneider, M., Norman, R., Parry, C., Bradshaw, D., & Pluddemann, A. (2007a), "Estimating the burden of disease attributable to alcohol use in South Africa in 2000", *Samj South African Medical Journal*, 97(8): 664-672.

Schneider, M., Norman, R., Steyn, N., & Bradshaw, D. (2007b), "Estimating the burden of disease attributable to low fruit and vegetable intake in South Africa in 2000", *Samj South African Medical Journal*, 97(8): 717-723.

SCIEH (1997a), "Surveillance of outbreaks of infectious intestinal disease, 1996", *SCIEH Weekly Report*, 31(97/21): 1.

SCIEH (1997b), "Surveillance of outbreaks of infectious intestinal disease, first quarter 1997", *SCIEH Weekly Report*, 31(97/20): 101-104.

SCIEH (1998a), "Outbreak of *Escherichia coli* O157 Infection in Greater Glasgow", *SCIEH Weekly Report*, 35(98/35): 191.

SCIEH (1998b), "Surveillance of outbreaks of infectious intestinal disease, 1997, and first quarter of 1998", *SCIEH Weekly Report*, 32(98/17): 98-99.

SCIEH (1999), "*E. coli* outbreak in Grampian", *SCIEH Weekly Report*, 33(99/12): 156.

SCIEH (2000a), *SCIEH Weekly Report*, 34: 135.

SCIEH (2000b), "Surveillance Report of outbreaks of infectious intestinal disease for 1999", *SCIEH Weekly Report*, 34(00/14): 82.

SCIEH (2000c), "Surveillance Report: Gastro-intestinal Infections", *SCIEH Weekly Report*, 34(00/02): 10.

SCIEH (2001a), "*E. coli* O157 outbreak associated with a garden centre", *SCIEH Weekly Report*, 35(40): 245.

SCIEH (2001b), "Surveillance Report: Gastro-intestinal Infections", *SCIEH Weekly Report*, 35(01/13): 87.

SCIEH (2002a), "Sorbitol-fermenting, verotoxigenic *E. coli* O157", *SCIEH Weekly Report*, 36(02/43): 281.

SCIEH (2002c), "Surveillance Report: Gastro-intestinal Infections", *SCIEH Weekly Report*, 36(02/17): 118-119.

SCIEH (2002d), "Surveillance Report: Gastro-intestinal Infections", *SCIEH Weekly Report*, 36(02/13): 91.

SCIEH (2002b), "Surveillance Report: Gastro-intestinal Infections", *SCIEH Weekly Report*, 36(02/01): 2-3.

SCIEH (2003a), "Decline in reports of *E. coli* O157", *SCIEH Weekly Report*, 37(03/41): 249.

SCIEH (2003d), "Surveillance Report: Gastro-intestinal Infections", *SCIEH Weekly Report*, 37(03/17): 114-115.

SCIEH (2003b), "Surveillance Report: Gastro-intestinal Infections", *SCIEH Weekly Report*, 37(03/31): 2-3.

SCIEH (2003c), "Surveillance Report: Gastro-intestinal Infections", *SCIEH Weekly Report*, 37(03/22): 142.

Scottish Executive (2007a), "Scottish Executive - Topics - Statistics in Scotland", retrieved on 9 July 2007a from: http://www.scotland.gov.uk/Topics/Statistics/.

Scottish Executive (2007b), "Shedding light on *E. coli* O157", retrieved on 17 November 2007b from: http://www.infoscotland.com/handsclean/CCC_FirstPage.jsp.

Shah, S., Hoffman, R., Shillam, P., & Wilson, J. B. (1996), "Prolonged Faecal Shedding of *Escherichia coli* O157:H7 During an Outbreak at a Day Care Center", *Clinical Infectious Diseases*, 23: 835-836.

Sharp, J. C. M., Coia, J. E., Curnow, J., & Reilly, W. (1994a), "*Escherichia coli* O157 infections in Scotland", *Journal of Medical Microbiology*, 40: 3-9.

Sharp, J. C. M., Ritchie, L. D., Curnow, J., & Reid, T. M. S. (1994b), "High incidence of haemorrhagic colitis due to *Escherichia coli* O157 in one Scottish town: clinical and epidemiological features", *Journal of Infection*, 29: 343-350.

Shaw, D. J., Grenfell, B. T., & Dobson, A. P. (1998), "Patterns of macroparasite aggregation in wildlife host populations", *Parasitology*, 117: 597-610.

Shaw, D. J., Jenkins, C., Pearce, M. C., Cheasty, T., Gunn, G. J., Dougan, G., Smith, H. R., Woolhouse, A. E. J., & Frankel, G. (2004), "Shedding patterns of

verocytotoxin-producing *Escherichia coli* strains in a cohort of calves and their dams on a Scottish beef farm", *Applied and Environmental Microbiology*, 70(12): 7456-7465.

Sheng, H. Q., Knecht, H. J., Kudva, I. T., & Hovde, C. J. (2006), "Application of bacteriophages to control intestinal *Escherichia coli* O157:H7 levels in ruminants", *Applied and Environmental Microbiology*, 72(8): 5359-5366.

Shukla, R., Slack, R., George, A., Cheasty, T., Rowe, B., & Scutter, J. (1995), "*Escherichia coli* O157 infection associated with a farm visitor centre", *CDR Review*, 5(6): R86-R90.

Silk, B. J. & Berkelman, R. L. (2005), "A Review of Strategies for Enhancing the Completeness of Notifiable Disease Reporting", *Journal of Public Health Management and Practice*, 11(3): 191-200.

Silvestro, L., Caputo, M., Blancato, S., DeCastelli, L., Fioravanti, A., Tozzoli, R., Morabito, S., & Caprioli, A. (2004), "Asymptomatic carriage of verocytotoxin-producing *Escherichia coli* O157 in farm workers in Northern Italy", *Epidemiology and Infection*, 132: 915-919.

Slutsker, L., Ries, A. A., Maloney, K., Wells, J. G., Greene, K. D., & Griffin, P. M. (1998), "A Nationwide Case-Control Study of *Escherichia coli* O157:H7 Infection in the United States", *The Journal of Infectious Diseases*, 177: 962-966.

Smith, H., Cheasty, T., Willshaw, G., Caprioli, A., Tozzi, A., Coia, J. E., Fisher, I., O'Brien, S., Fruth, A., Tschäpe, H., & Reilly, W. (2002), "Changing patterns of VTEC Infection in Britain and continental Europe", *IVC News 16*, 15(12): Supplement 1.

Smith-Palmer, A. & Cowden, J. M. (2004), "Annual report of general outbreaks of infectious intestinal disease in Scotland, 2003", *SCIEH Weekly Report*, 38(04/32): 190-194.

Smith-Palmer, A., Cowden, J. M., & Locking, M. E. (2005), "Annual report of general outbreaks of infectious intestinal disease in Scotland, 2004", *HPS Weekly Report*, 39(05/50): 282-285.

Smith-Palmer, A., Stewart, W. C., Mather, H., Greig, J., Cowden, J. M., & Reilly, W. (2003), "Epidemiology of *Salmonella enterica* serovars Enteritidis and Typhimurium in animals and people in Scotland between 1990 and 2001", *Veterinary Record*, 153: 517-520.

Snedecor, G. W. & Cochran, W. G. (1971), *Statistical Methods*, Iowa State University Press, Ames, Iowa.

Sockett, P. N., Ng, L.-K., Doré, K., Ellis, A., Demczuk, W., Ciampa, N., Flint, J. A., & Tinga, C. (2006) *Epidemiology of Shiga Toxin (Verocytotoxin) Producing E. coli infections in Canada from 1994 to 2003*, Public Health Agency of Canada.

Soderstrom, A., Lindberg, A., & Andersson, Y. (2005), "EHEC O157 outbreak in Sweden from locally produced lettuce, August-September 2005", *Eurosurveillance Weekly*, 10(9).

Spika, J. S., Parsons, J. E., Nordenberg, D., Wells, J. G., Gunn, R. A., & Blake, P. A. (1986), "Hemolytic Uremic Syndrome and Diarrhea Associated with *Escherichia coli* O157:H7 in A Day-Care-Center", *Journal of Pediatrics*, 109(2): 287-291.

Statistics Canada (2006), "Population estimates and projections", retrieved on 7 July 2007 from: http://www40.statcan.ca/l01/cst01/demo02a.htm.

StatsDirect Ltd. (2007), StatsDirect statistical software. 2.6.2. 2007. England.

Steinmuller, N., Demma, L., Bender, J. B., Eidson, M., & Angulo, F. J. (2006), "Outbreaks of Enteric Disease Associated with Animal Contact: Not Just a Foodborne Problem Anymore", *Clinical Infectious Diseases*, 43: 1596-1602.

Strachan, N. J. C., Doyle, M. P., Kasuga, F., Rotariu, O., & Ogden, I. D. (2005), "Dose response modelling of *Escherichia coli* O157 incorporating data from foodborne and environmental outbreaks", *International Journal of Food Microbiology*, 103: 35-47.

Strachan, N. J. C., Dunn, G. M., Locking, M. E., Reid, T. M. S., & Ogden, I. D. (2006), "*Escherichia coli* O157: Burger bug or environmental pathogen?", *International Journal of Food Microbiology*, 112(2): 129-137.

Strockbine,N., Wells,J., Bopp,C., & Barrett,T. (1998) "Overview of Detection and Subtyping Methods," in *Escherichia coli O157:H7 and Other Shiga Toxin-Producing E. coli Strains*, J. Kaper & A. O'Brien, eds., ASM Press: Washington, DC, pp. 331-356.

Sugiyama, A., Iwade, Y., Akachi, S., Nakano, Y., Matsuno, Y., Yano, T., Yamauchi, A., Nakayama, O., Sakai, H., Yamamoto, K., Nagasaka, Y., Nakano, T., Ihara, T., & Kamiya, H. (2005), "An Outbreak of Shigatoxin-Producing *Escherichia coli* O157:H7 in a Nursery School in Mie Prefecture", *Japanese Journal of Infectious Diseases*, 58: 398-400.

Sussman, M. (1997), *Escherichia coli: Mechanisms of Virulence*, Cambridge University Press, Cambridge.

Sutcliffe, P., Picard, L., Fortin, B., Malaviarachchi, D., Hohenadel, J., & O'Donnell, B. (2004), "*Escherichia coli* O157:H7 outbreak at a summer hockey camp, Sudbury, 2004", *Canadian Communicable Disease Report*, 30(22).

Swerdlow, D. & Griffin, P. M. (1997), "Duration of faecal shedding of *Escherichia coli* O157:H7 among children in day-care centres", *The Lancet*, 349: 745-746.

Swerdlow, D., Woodruff, B., Brady, R., Griffin, P. M., Tippen, S., Donnell, H., Geldreich, E., Payne, B., Meyer, A., Wells, J., Greene, K. D., Bright, M., Bean, N., & Blake, P. A. (1992), "A waterborne outbreak in Missouri of *Escherichia coli*

O157:H7 associated with bloody diarrhea and death", *Annals of Internal Medicine*, 117(10): 812-819.

Tarr, P. I. (1995), "*Escherichia coli* O157:H7 - Clinical, Diagnostic, and Epidemiologic Aspects of Human Infection", *Clinical Infectious Diseases*, 20(1): 1-10.

Tarr, P. I., Gordon, C. A., & Chandler, W. L. (2005), "Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome", *The Lancet*, 365(9464): 1073-1086.

Tarr, P. I., Niell, M. A., Clausen, C. R., Watkins, S. L., Christie, D. L., & Hickman, R. O. (1990), "*Escherichia coli* O157:H7 and the hemolytic uremic syndrome: importance of early cultures in establishing the etiology", *Journal of Infectious Disease*, 162: 553-556.

Task Force on E.coli O157 (2001) *Task Force on E. coli O157 - Final Report*, Accessed at: http://www.food.gov.uk/multimedia/pdfs/ecolitaskfinreport.pdf.

Tauxe, R. V. (1998), "New Approaches to Surveillance and Control of Emerging Foodborne Infectious Diseases", *Emerging Infectious Diseases*, 4(3): 455-456.

Tennis, P., Takumi, K., & Shinagawa, K. (2004), "Dose Response for Infection by *Escherichia coli* O157:H7 from Outbreak Data", *Risk Analysis*, 24(2): 401-407.

Terajima, J., Izumiya, H., Iyoda, S., Tamura, K., & Watanabe, H. (1999), "Detection of a multi-prefectural *E. coli* O157:H7 outbreak caused by contaminated Ikura-Sushi ingestion", *Japanese Journal of Infectious Diseases*, 52(2): 52-53.

The R Project (2007), "Test of Equal or Given Proportions", retrieved on 18 June 2007 from: http://finzi.psych.upenn.edu/R/library/stats/html/prop.test.html.

Thomas, M. K., Majowicz, S. E., Sockett, P. N., Fazil, A., Pollett, G., Dore, K. A., Flint, J. A., & Edge, V. L. (2006), "Estimated numbers of community cases of illness due to *Salmonella*, *Campylobacter* and verotoxigenic *Escherichia coli*: Pathogen-specific community rates", *Canadian Journal of Infectious Diseases and Medical Microbiology*, 17(4): 229-234.

Thomson-Carter, F., Allison, L. J., & Pennington, T. H. (1996), "*Escherichia coli* O157 in Scotland, 1995", *SCIEH Weekly Report*, 30(96/13): 70.

Tinga, C., Valcour, J. E., & Dore, K. A. (2005) *Preliminary Report of the National Database of Enteric Outbreak Summaries, 1996 - 2003*, Foodborne, Waterborne, and Zoonotic Infections Division, Centre for Infectious Disease Prevention & Control; Public Health Agency of Canada.

Tinga, C., Valcour, J., & Doré, K. (2006) *Provincial/Territorial Enteric Outbreaks in Canada, 1996-2003*, Public Health Agency of Canada.

Tleyjeh, I. M., Steckelberg, J. M., Murad, H. S., Anavekar, N. S., Ghomrawi, H. M. K., Mirzoyev, Z., Moustafa, S. E., Hoskin, T. L., Mandrekar, J. N., Wilson, W. R., & Baddour, L. M. (2005), "Temporal trends in infective endocarditis - A population-based study in Olmsted County, Minnesota", *JAMA*, 293(24): 3022-3028.

Todd, E. (2000), "*Escherichia coli* O157:H7 Infections Associated With Ground Beef and Their Control in Canada", *Canadian Communicable Disease Report*, 26(13): 111-116.

Todd, E., Chatman, P., & Rodrigues, V. (2000) *Annual summaries of foodborne and waterborne diseases in Canada, 1994 and 1995*, Health Products and Food Branch, Health Canada, Polyscience Publications Inc..

Todd, W. T. A. & Dundas, S. (2001), "The management of VTEC O157 infection", *International Journal of Food Microbiology*, 66: 103-110.

Trevena, W. B., Willshaw, G., Cheasty, T., Domingue, G., & Wray, C. (1999), "Transmission of Vero cytotoxin producing *Escherichia coli* O157 infection from farm animals to humans in Cornwall and West Devon", *Communicable Disease and Public Health*, 2: 263-268.

Tuttle, J., Gomez, T., Doyle, M. P., Wells, J. G., Zhao, T., Tauxe, R. V., & Griffin, P. M. (1999), "Lessons from a large outbreak of *Escherichia coli* O157:H7 infections: insights into the infectious dose and method of widespread contamination of hamburger patties", *Epidemiology and Infection*, 122: 185-192.

U.S.Census Bureau (2006), "American Fact Finder - Population Estimates", retrieved on 8 October 2007 from: http://factfinder.census.gov/servlet/DTTable?_bm=y&-geo_id=01000US&-ds_name=PEP_2006_EST&-state=dt&-mt_name=PEP_2006_EST_G2006_T001.

United States Department of Agriculture (2007a), "2002 Census of Agriculture", retrieved on 8 October 2007a from: http://www.nass.usda.gov/Census/Pull_Data_Census.

United States Department of Agriculture (2007b), "Recall Release: New Jersey firm expands recall of ground beef products due to possible *E. coli* O157:H7 contamination", retrieved on 1 October 2007b from: http://www.fsis.usda.gov/PDF/040_2007_Expanded_Recall.pdf.

Upton, P. A. & Coia, J. E. (1994), "Outbreak of *Escherichia coli* O157 infection associated with pasteurised milk supply", *The Lancet*, 344: 1015.

Valcour, J. E., Michel, P., McEwen, S. A., & Wilson, J. B. (2002), "Associations between Indicators of Livestock Farming Intensity and Incidence of Human Shiga Toxin-Producing *Escherichia coli* Infection", *Emerging Infectious Diseases*, 8(3): 252-256.

Vallance, B. A., Chan, C., Robertson, M. L., & Finlay, B. B. (2002), "Enteropathogenic and enterohemorrhagic *Escherichia coli* infections: emerging

themes in pathogenesis and prevention", *Canadian Journal of Gastroenterology*, 16(11): 771-778.

van de Giessen, A. W., Bouwknegt, M., Dam-Deisz, W. D. C., van Pelt, W., Wannet, W. J. B., & Visser, G. (2006), "Surveillance of *Salmonella* spp. and *Campylobacter* spp. in poultry production flocks in the Netherlands", *Epidemiology and Infection*, 134: 1266-1275.

Van der Heyden, J. H. A., Catchpole, M. A., Paget, W. J., Stroobant, A., & European Study Group (2000), "Trends in gonorrhoea in nine western European countries, 1991 - 6", *Sexually Transmitted Infections*, 76: 110-116.

van Rest, E. D. & Parkin, E. A. (1933), "Poisson series and biological data", *Nature*, 132: 445.

Varma, J. K., Greene, K. D., Reller, M. E., DeLong, S. M., Trottier, J., Nowicki, S. F., DiOrio, M., Koch, E. M., Bannerman, T. L., York, S. T., Lambert-Fair, M. A., Wells, J. G., & Mead, P. S. (2003), "An outbreak of *Escherichia coli* O157 infection following exposure to a contaminated building", *JAMA*, 290(20): 2709-2712.

Venzon, D. J. & Moolgavkar, S. H. (1984), "Cohort Analysis of Malignant Melanoma in Five Countries", *American Journal of Epidemiology*, 119(1): 62-70.

Verma, A., Bolton, F. J., Fiefield, D., Lamb, P., Woloschin, E., Smith, N., & McCann, R. (2007), "An Outbreak of *E. coli* O157 associated with a swimming pool: an unusual vehicle of transmission", *Epidemiology and Infection*.

Voetsch, A. C., Angulo, F., Rabatsky-Ehr, T. R., Shallow, S., Cassidy, M., Thomas, S., Swanson, E., Zansky, S., Hawkins, M., Jones, T., Shillam, P. J., Van Gilder, T. J., Wells, J. G., & Griffin, P. M. (2004a), "Laboratory Practices for Stool-Specimen Culture for Bacterial Pathogens, Including *Escherichia coli* O157:H7, in the FoodNet Sites, 1995 - 2000", *Clinical Infectious Diseases*, 38(S3): S190-S197.

Voetsch, A. C., Kennedy, M., Keene, W. E., Smith, K. E., Rabatsky-Ehr, T., Zansky, S., Thomas, S. M., Mohle-Boetani, J. C., Sparling, P., McGavern, M. B., & Mead, P. S. (2006), "Risk factors for sporadic Shiga toxin-producing *Escherichia coli* O157 infections in FoodNet sites, 1999-2000", *Epidemiology and Infection*.

Voetsch, A. C., Van Gilder, T. J., Angulo, F., Farley, M., Shallow, S., Marcus, R., Cieslak, P. R., Deneen, V., & Tauxe, R. V. (2004b), "FoodNet Estimate of the Burden of Illness Caused by Nontyphoidal *Salmonella* Infections in the United States", *Clinical Infectious Diseases*, 38(Suppl 3): S127-S134.

Wachsmuth, I., Sparling, P., Barrett, T., & Potter, M. (1997), "Enterohemorrhagic *Escherichia coli* in the United States", *FEMS Immunology and Medical Microbiology*, 18: 233-239.

Wahl, M. & Andersson, Y. (2004), "EHEC cases from an international football tournament (Gothia Cup) in Sweden, July 2004", *Eurosurveillance Weekly*, 8(36).

Wall, P., McDonnell, R., Adak, G., Cheasty, T., Smith, H., & Rowe, B. (1996), "General outbreaks of Vero cytotoxin producing *Escherichia coli* O157 in England and Wales from 1992 to 1994", *CDR Review*, 6(2): R26-R33.

Warshawsky, B., Gutmanis, I., Henry, B., Dow, J., Reffle, J., Pollett, G., Ahmed, R., Aldom, J., Alves, D., Chagla, A., Ciebin, B., Kolbe, F., Jamieson, F., & Rodgers, F. (2002), "Outbreak of *Escherichia coli* O157:H7 related to animal contact at a petting zoo", *Canadian Journal of Infectious Diseases*, 13(3): 175-181.

Waters, J. R., Sharp, J. C. M., & Dev, V. J. (1994), "Infection Caused by *Escherichia coli* O157:H7 in Alberta, Canada and in Scotland: A Five-Year Review, 1987-1991", *Clinical Infectious Diseases*, 19: 834-843.

Welinder-Olsson, C. (2005), "Enterohemorrhagic *Escherichia coli* (EHEC)", *Scandinavian Journal of Infectious Diseases*, 37: 403-416.

Welinder-Olsson, C., Stenqvist, K., Badenfors, M., Brandberg, A., Floren, K., Holm, M., Holmberg, L., Kjellin, E., Marild, S., Studahl, A., & Kaijser, B. (2004), "EHEC outbreak among staff at a children's hospital - use of PCR for verocytotoxin detection and PFGE for epidemiological investigation", *Epidemiology and Infection*, 132(1): 43-49.

Weller, S. C. & Stanberry, L. R. (2007), "Estimating the population prevalence of HPV", *JAMA*, 297(8): 876-878.

Wells, J., Davis, B., & Wachsmuth, I. (1983), "Laboratory investigation of hemorrhagic colitis outbreaks associated with a rare *Escherichia coli* serotype", *Journal of Clinical Microbiology*, 18: 512-520.

Werber, D., Behnke, S. C., Fruth, A., Merle, R., Menzler, S., Glaser, S., Kreienbrock, L., Prager, R., Tschape, H., Roggentin, P., Bockemuhl, J., & Ammon, A. (2007), "Shiga toxin-producing *Escherichia coli* infection in Germany - Different risk factors for different age groups", *American Journal of Epidemiology*, 165(4): 425-434.

Wheeler, J. G., Sethi, D., Cowden, J. M., Wall, P., Rodrigues, L. C., Tompkins, D. S., Hudson, M. J., & Roderick, P. J. (1999), "Study of infectious intestinal disease in England: rates in the commnunity, presenting to general practice, and reported to national surveillance", *British Medical Journal*, 318: 1046-1050.

Whittam,T.S. (1998) "Evolution of *Escherichia coli* O157:H7 and Other Shiga Toxin-Producing *E. coli* Strains," in *Escherichia coli O157:H7 and Other Shiga Toxin-Producing E. coli Strains*, J. B. Kaper & A. D. O'Brien, eds., ASM Press: Washington, D.C., pp. 195-209.

WHO (1996), "Food Safety: Enterohaemorrhagic *Escherichia coli* infection", *WHO Weekly Epidemiological Record*, 71(35): 267-268.

Williams, R. C., Isaacs, S., Decou, M. L., Richardson, E. A., Buffett, M. C., Slinger, R. W., Brodsky, M. H., Ciebin, B. W., Ellis, A., & Hockin, J. (2000), "Illness

outbreak associated with *Escherichia coli* O157:H7 in Genoa salami.", *Canadian Medical Association Journal*, 162(10): 1409-1413.

Willshaw, G., Cheasty, T., & Pritchard, G. "Outbreaks of VTEC O157 infection linked to animal contact in England amd Wales between 1996 and 2005", in VTEC 2006 Abstracts.

Willshaw, G., Cheasty, T., Smith, H., O'Brien, S., & Adak, G. (2001), "Verocytotoxin-producing *Escherichia coli* (VTEC) O157 and other VTEC from human infections in England and Wales: 1995-1998", *Journal of Medical Microbiology*, 50: 135-142.

Willshaw, G., Smith, H., Cheasty, T., Wall, P., & Rowe, B. (1997), "Vero Cytotoxin-Producing *Escherichia coli* O157 Outbreaks in England and Wales, 1995: Phenotypic Methods and Genotypic Subtyping", *Emerging Infectious Diseases*, 3(4): 561-565.

Wilson, J. B., Johnson, R., Clarke, R. C., Rahn, K., Renwick, S., Alves, D., Karmali, M., Michel, P., Orrbine, E., & Spika, J. (1997), "Canadian Perspectives on Verocytotoxin-Producing *Escherichia coli* Infection", *Journal of Food Protection*, 60(11): 1451-1453.

Wilson, J. B., Spika, J., Clarke, R. C., McEwen, S. A., Johnson, R., Rahn, K., Renwick, S., Karmali, M., Lior, H., Alves, D., Gyles, C. L., & Sandhu, K. (1996), "Verocytotoxigenic *Escherichia coli* infection in dairy farm families", *Journal of Infectious Disease*, 174: 1021-1027.

Winnipeg Regional Health Authority (2007) *Outbreak of Verotoxigenic E. coli in the Winnipeg Health Region, Summer 2006.*

Wong, C. S., Jelacic, S., Habeeb, R. L., Watkins, S. L., & Tarr, P. I. (2000), "The risk of the hemolytic-uremic syndrome after antibiotic treatment of *Escherichia coli* O157:H7 infections", *New England Journal of Medicine*, 342(26): 1930-1936.

Woodward, D., Clark, C., Caldeira, R. A., Ahmed, R., & Rodgers, F. (2002), "Verotoxigenic *Escherichia coli* (VTEC): A major public health threat in Canada", *Canadian Journal of Infectious Diseases*, 13(5): 321-330.

Woolhouse, M. E. J. (1998), "Patterns in Parasite Epidemiology: The Peak Shift", *Parasitology Today*, 14(10): 428-434.

Woolhouse, M. E. J., Taylor, L. H., & Haydon, D. T. (2001), "Population Biology of Multihost Pathogens", *Science*, 292(5519): 1109-1112.

Yamamoto, J., Ishikawa, A., Miyamoto, M., Nomura, T., Uchimura, M., & Koiwai, K. (2001), "Outbreak of enterohemorrhagic *Escherichia coli* O157 mass infection caused by 'whole roasted cow'", *Japanese Journal of Infectious Diseases*, 54(2): 88-89.

Zeger, S. L., Irizarry, R., & Peng, R. D. (2006), "On time series analysis of public health and biomedical data", *Annual Review of Public Health*, 27: 57-79.

Ziese, T., Anderson, Y., de Jong, B., Lofdahl, S., & Ramberg, M. (1996), "Outbreak of *Escherichia coli* O157 in Sweden", *Eurosurveillance Monthly*, 1(1): 2-3.

Zimmerhackl, L. B. (2000), "*E. coli*, Antibiotics, and the Hemolytic-Uremic Syndrome", *The New England Journal of Medicine*, 342(26): 1990-1991.