**EVALUATION OF OUTBREAK DETECTION ALGORITHMS:
AN EMPIRICAL AND A SIMULATION STUDY**

by

SRI VIDYA THUPPAL, B.Sc.

A thesis submitted to

The Faculty of Graduate Studies and Research

in partial fulfilment of

the requirements for the degree of

Master of Science

School of Mathematics and Statistics

Ottawa-Carleton Institute of Mathematics and Statistics

Carleton University

Ottawa, Ontario, Canada

© Copyright

March 2004, Sri Vidya Thuppal

Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-93990-1

Our file Notre référence

ISBN: 0-612-93990-1

The author has granted a non-exclusive license allowing the Library and Archives Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

PAGINATION ERROR.

TEXT COMPLETE.

ERREUR DE PAGINATION.

LE TEXTE EST COMPLET.

Abstract

The purpose of this research is to compare a selected group of algorithms that detect outbreaks of communicable diseases. Five algorithms used nationally and internationally to detect outbreaks of communicable diseases, are compared in this study using empirical and simulated data sets. In the empirical study, the algorithms were run on seven data sets from common, rare, seasonal or secular types of communicable diseases. The evaluation measures of sensitivity, specificity, positive predictive value, and times to detection of outbreaks were estimated (along with their standard deviations) and studied. The estimated measures were also circulated to a selected group of experts including epidemiologists and statisticians. The analysis of the measures (and the expert opinion) was used to identify differences or similarities in the algorithms. In the simulation study, 50 data sets that depict patterns of common, rare and seasonal diseases were generated using the characteristics of the data sets that were used in the empirical study such as empirical probability distribution and number of outbreaks. Similar to the empirical study, the algorithms were run on the 150 data sets (50 data sets for each of the three types of diseases) and the evaluation measures were estimated and analyzed. Along with the estimated measures, additional statistical analysis such as cumulative distribution function plots, interaction plots, and multivariate analysis of variance were carried out. The analyses showed that interactions existed between the algorithms and the type of disease and hence that the ability of the algorithms to detect outbreaks was highly influenced by the type of disease.

Acknowledgements

I am deeply indebted to my supervisors Dr. Shirley E. Mills and Dr. Patrick Farrell, for their invaluable guidance, exceptional suggestions, continued encouragement and support in all aspects of my research. I have greatly enjoyed working under their supervision.

I would like to thank Dr. Cyril W.L Garner, Director of the School of Mathematics and Statistics, and Dr. Sam Melkonian, Associate Director of the School of Mathematics and Statistics, and the faculty members of the School of Mathematics and Statistics for their support during my study.

I am grateful to Ms. Lisa Belzak, Dr. Jamie Hockin, and Dr. Michael Goddard for providing me with directions, and for their constant support in pursuing my research. I would also like to express my gratitude to them for providing financial support while working at Health Canada. I sincerely thank Ms. Lisa Belzak for her excellent guidance, and incessant encouragement during my research work at Health Canada.

I would also like to express thanks to Mr. Nick Andrews, Mr. Arnold Bosman, Ms. Lori Hutwagner, Mr. Mark Tinga, Dr. Wilfrid Van Pelt, and Mr. Rick White, for providing access to their algorithms for the purposes of my research. Their invaluable advice and suggestions during the course of my research are greatly appreciated. I would also like to express my gratitude to Dr. Howard Burkom for his excellent advice and encouragement.

My sincere thanks to the Alberta Provincial Laboratory, the British Columbia Centre for Disease Control, and the National Institute of Public Health and the Environment for giving me access to the data files required for my research.

Also I would like to thank the epidemiologists who provided their support during the course of my research.

I would like to express my gratitude to Dr. Matias Salibian-Barrera for his time, patience, and valuable guidance. I would also like to thank Mr. Andrew Williams for this excellent suggestions and ideas.

I would like to express my special thanks to my parents Mr. & Mrs. Chakravarthi Srinivasan, my father-in-law and mother-in-law Mr. & Mrs. Desika Chari Thuppal, my sister and brother-in-law Mr. & Mrs. Balaji Narayanan, my brother Ramapriyan Chakravarthi, and most of all my little nephew Kesav Narayanan, for their immeasurable love and support.

Finally, I would like to thank my beloved husband, Rajagopalan Thuppal for all his patience, immense support, constant encouragement, invaluable advice, and for taking care of me throughout my study. He has been very encouraging and has always been boosting my confidence.

Table of Contents

Abstract.....	iii
Acknowledgements	iv
Table of Contents	vi
List of Tables	viii
List of Figures.....	ix
Chapter 1. Introduction.....	1
Chapter 2. Literature Review	4
Chapter 3. Outbreak Detection Algorithms and Evaluation Criteria	10
3.1. <i>Outbreak Detection Algorithms</i>	10
3.1.1. EARS-Historical	11
3.1.2. EARS-Non Historical	12
3.1.3. Generalized Additive Model.....	13
3.1.4. Log Linear Regression Model (HPA).....	14
3.1.5. Log Linear Regression Model (RIVM)	15
3.2. <i>Evaluation Criteria</i>	16
Chapter 4. Empirical Study	23
4.1.1. Data Sources and Diseases.....	23
4.1.2. Gold Standard: Consensus method	24
4.2. <i>Empirical Evaluation</i>	25
4.2.1. Common Disease Type	26
4.2.2. Rare Disease Type	27
4.2.3. Seasonal Disease Type.....	30
4.2.4. Secular Trend Disease Type	32
4.2.5. Overall Performance	34
Chapter 5. Simulation Study.....	37
5.1. <i>Data simulation</i>	37
5.2. <i>Gold Standard</i>	40
5.3. <i>Simulation Evaluation</i>	45
5.3.1. Cumulative Distribution Function Plot.....	45
5.3.2. Box Plots of Times to Detection.....	48

5.3.3.	Two-way Multivariate Analysis of Variance (MANOVA)	49
5.3.4.	Interaction Plots between Algorithms and Diseases	51
5.3.5.	Univariate Analysis of Variance (ANOVA)	54
Chapter 6.	Conclusion	62
Bibliography		65
Appendix A: Tables		68
Appendix B: Figures		85
Appendix C: R Codes for Simulation		123
Appendix D: Exact Binomial Confidence Interval for Empirical Study		134

List of Tables

Table 1: Possible test outcomes	17
Table 2: Data sets used in Empirical Study	24
Table 3: Univariate analysis for testing the interaction effects (Simulation Study)	55
Table 4: Algorithms suitable in detecting outbreaks by type of disease (Simulation Study)	61
Table 5: Estimates of measures for Hepatitis A data set (Empirical Study)	69
Table 6: Estimates of measures for Hanta Virus data set (Empirical Study)	70
Table 7: Estimates of measures for Salmonella Dublin data set (Empirical Study)	71
Table 8: Estimates of measures for Influenza A & B data set (Empirical Study)	72
Table 9: Estimates of measures for Salmonella Enteritidis data set (Empirical Study)	73
Table 10: Estimates of measures for Gonorrhea data set (Empirical Study)	74
Table 11: Estimates of measures for Meningococcal data set (Empirical Study)	75
Table 12: Summary measures for Common disease (Simulation study)	77
Table 13: Times to detection for Common disease (Simulation study)	77
Table 14: Summary measures for Rare disease (Simulation study)	78
Table 15: Times to detection summary for Rare Disease (Simulation Study)	78
Table 16: Summary measures for Seasonal Disease (Simulation Study)	80
Table 17: Times to detection summary for Seasonal Disease (Simulation Study)	80
Table 18: Tests of hypothesis for interaction effects for Sensitivity Weeks (Simulation Study)	81
Table 19: Tests of hypothesis for interaction effects for Sensitivity Run (Simulation Study)	82
Table 20: Tests of hypothesis for interaction effects for Specificity (Simulation Study)	83
Table 21: Tests of hypothesis for interaction effects for PPV (Simulation Study)	84

List of Figures

Figure 1: CDF plot of Sensitivity Weeks for Common Disease (Simulation Study)	46
Figure 2: Flags of algorithms on Hepatitis A data set (Empirical Study)	86
Figure 3: Flags of algorithms on Hanta Virus data set (Empirical Study)	87
Figure 4: Flags of algorithms on Salmonella Dublin data set (Empirical Study)	88
Figure 5: Flags of algorithms on Influenza A & B data set (Empirical Evaluation)	89
Figure 6: Flags of algorithms on Salmonella Enteritidis data set (Empirical Study)	90
Figure 7: Flags of algorithms on Gonorrhea data set (Empirical Study)	91
Figure 8: Flags of algorithms on Meningococcal data set (Empirical Study)	92
Figure 9 Example of simulation (Common disease)	93
Figure 10 Example of Simulation (Rare disease)	94
Figure 11 Example of Simulation (Seasonal disease)	95
Figure 12: Box plot of Times to detection for Common disease (Simulation Study)	96
Figure 13: Box Plot of Times to detection for Rare Disease (Simulation Study)	97
Figure 14: Box Plot of Times to detection for Seasonal Disease (Simulation Study)	98
Figure 15: CDF Plot of Sensitivity run for Common Disease (Simulation Study)	99
Figure 16: CDF Plot of Specificity for Common Disease (Simulation Study)	100
Figure 17: CDF Plot of PPV for Common Disease (Simulation Study)	101
Figure 18: CDF Plot of Sensitivity Weeks for Rare Disease (Simulation Study)	102
Figure 19: CDF plot for Sensitivity run for Rare disease (Simulation Study)	103
Figure 20: CDF plot of Specificity for Rare disease (Simulation Study)	104
Figure 21: CDF plot of PPV for Rare disease (Simulation Study)	105
Figure 22: CDF plot of Sensitivity weeks for Seasonal disease (Simulation Study)	106
Figure 23: CDF plot of Sensitivity run for Seasonal disease (Simulation Study)	107
Figure 24: CDF plot of Specificity for Seasonal disease (Simulation Study)	108
Figure 25: CDF plot of PPV for Seasonal disease (Simulation Study)	109
Figure 26: Average Sensitivity Weeks for Algorithms by Disease (Simulation Study)	110
Figure 27: Average Sensitivity Weeks by Disease for each Algorithm (Simulation Study)	110
Figure 28: Average Sensitivity Run for Algorithms by Disease (Simulation Study)	111
Figure 29: Average Sensitivity Run by Disease for each Algorithm (Simulation Study)	111
Figure 30: Average Specificity for Algorithms by Disease (Simulation Study)	112
Figure 31: Average Specificity by Disease for each Algorithm (Simulation Study)	112
Figure 32: Average PPV for Algorithms by Disease (Simulation Study)	113
Figure 33: Average PPV by Disease for each Algorithm (Simulation Study)	113
Figure 34: Normal Probability Plot for Sensitivity Weeks (Simulation Study)	114
Figure 35: Normal Probability Plot for Sensitivity Run (Simulation Study)	115
Figure 36: Normal Probability Plot for Specificity (Simulation Study)	116
Figure 37: Normal Probability Plot for PPV (Simulation Study)	117
Figure 38: Box Plot of Residuals for Sensitivity Weeks (Simulation Study)	118
Figure 39: Box Plot of Residuals for Sensitivity Run (Simulation Study)	119
Figure 40: Box Plot of Residuals for Specificity (Simulation Study)	120
Figure 41: Box Plot of residuals for PPV (Simulation Study)	121

CHAPTER 1. INTRODUCTION

Disease surveillance is a crucial part of the health care system of a country. Outbreaks in diseases such as severe acute respiratory syndrome (SARS) and salmonella have a great impact on the health of a community. It is important to identify unusual outbreaks in such diseases so as to instigate health care actions. Generally nurses, practitioners, and doctors at hospitals and local health care departments identify outbreaks in communicable diseases.

Disease outbreak detection algorithms play an important role in detecting outbreaks of communicable diseases. In general, an outbreak detection algorithm detects outbreaks in diseases by comparing the current counts of cases in a disease, with an expected number of cases, which is determined using previous years of historical information. If the current count of cases exceeds the expected number, the cases are flagged as potential outbreaks.

Many countries across the world use disease outbreak detection algorithms at various levels of jurisdiction and for various types of diseases. This research is directed towards a comparison of a selected group of outbreak detection algorithms, used nationally and internationally. The study is divided into two major sections: an empirical study that compares the algorithms using empirical data sets and pre-determined criteria for evaluation; and a simulation study that compares the algorithms using simulated data sets, the same pre-determined evaluation criteria as used in the empirical study, and

extensive statistical analysis such as multivariate analysis of variance and interaction plot analysis.

The empirical study is a part of a major project undertaken at the Center for Surveillance Coordination, Health Canada, with the collaboration of partners at the National Institute of Public Health and the Environment (RIVM), Netherlands; the Health Protection Agency (HPA), United Kingdom; the Centers for Disease Control and Prevention (CDC), United States; and the British Columbia Centre for Disease Control (BCCDC), Canada. The purpose of the project was to compare existing outbreak detection algorithms and propose one or a combination of algorithms that may be suitable for outbreak detection in Canada. The evaluation measures of sensitivity, specificity, positive predictive value, and times to detection of outbreaks have been used to compare the five algorithms (see section 4.1). The methodology and the results of the comparisons for the empirical and simulation studies are presented in this thesis. The thesis is organized into six major chapters: introduction, literature review, outbreak detection algorithms and evaluation criteria, empirical study, simulation study, and conclusion. In Chapter 2 details on the literature that was reviewed and the methodologies used to compare surveillance systems in the past is discussed. Chapter 3 introduces the five outbreak detection algorithms and the evaluation criteria that were used in this study to compare these algorithms. The empirical and simulation studies used to compare the five algorithms are discussed in Chapter 4 and 5 respectively. These chapters provide information on the data sources, methodologies and gold standard (see sections 4.1.2 and 5.2) used in estimating the

evaluation measures to compare the algorithms, as well as the results from these studies. Tables and figures accompanying the analysis are presented in the Appendix A and B.

CHAPTER 2. LITERATURE REVIEW

This chapter presents the literature review of previous studies that evaluated different surveillance systems, which detected outbreaks in diseases. The performance measures of sensitivity, specificity and positive predictive value (PPV) have been used to evaluate surveillance systems in the past. The timeliness for detection of outbreaks of the surveillance systems has also been studied to determine how early (or late) the systems detect outbreaks.

Takahashi et al (2001) evaluated the School Health Surveillance System (SHSS) for detecting outbreaks in paediatric influenza in comparison with the Influenza Sentinel Surveillance System (ISSS), using the surveillance guidelines (Centers for Disease Control and Prevention, 1998). Data from school absenteeism (ages 4-15 years old) was compared with data from influenza-like illness (ILI) for sentinel sites during the second to the 11th weeks of 1998 and 1999. Among other measures determining the quality of the data itself, sensitivity, specificity, and positive predictive value were used to evaluate the effectiveness of the SHSS in measuring the impact of influenza. The sensitivity, specificity, and PPV of the SHSS compared to the ISSS were calculated as 80%, 100% and 100% respectively. The SHSS system had higher levels of sensitivity, specificity and PPV. The authors however suggested further analysis to determine if the system would perform similarly during less severe influenza seasons.

Fujii et al (2002) compared a physical sentinel surveillance system that detected outbreaks in influenza, the Influenza Sentinel Surveillance System (ISSS), with the

Microbial Isolation Surveillance System (MISS). The ISSS system was evaluated during five consecutive influenza seasons (1995 to 2000) in Kyoto City. Estimates of sensitivity, specificity, and positive predictive value of the ISSS were calculated using the MISS by influenza season and by month as a reference. They found that the sensitivity of the ISSS ranged from 79.1% to 96.8%, specificity ranged from 55% to 76.8%, and PPV ranged from 28.9% to 52.1% by season, and that the sensitivity ranged from 85.7% to 100%, specificity ranged from 38.4% to 92.5%, and PPV ranged from 31.4% to 50.4% by month. They concluded that the ISSS was in general a good surveillance system for monitoring influenza activities.

Hutwagner et al (1997) developed an algorithm for detecting salmonella outbreaks using laboratory-based surveillance data. The algorithm is based on seasonal adjustments to the cumulative sums (CUSUM) method, a quality control method commonly used in manufacturing. Hutwagner et al used different averages, the mean of 5 weeks, median of 5 weeks, and mean of 15 weeks over the past 5 years as the expected value in the algorithm, to identify which average provided the best sensitivity, specificity, and false-positive rate. They chose to use the mean of five weeks to summarize the historical data because it produced the highest specificity (the difference in sensitivity was considered negligible). The sensitivity of the algorithm, using the mean of five weeks, ranged from 0% to 100%, specificity ranged from 64% to 100% and false positive rate ranged from 0% to 1%. Hutwagner et al concluded that the algorithm with mean of five weeks as the expected value produced the best sensitivity, specificity, and PPV for detecting salmonella outbreaks.

Barbujani and Calzolari (1984) performed a study to compare two statistical techniques for the surveillance of birth defects, the Sets Technique (Chen 1978) and the CUSUM (Weatherall 1976), through a Monte Carlo simulation. The two techniques were compared using sensitivity, specificity, accuracy, alarm delays and a direct comparison method. Barbujani and Calzolari concluded based on their methods of evaluation that CUSUM was a more effective method than the Sets technique for detecting increases in congenital malformation rates

Hashimoto et al (2000) evaluated the paediatric disease surveillance system by calculating the sensitivity, specificity, and positive predictive values for epidemic warnings for a selected set of critical values. The paediatric disease surveillance system was used to detect epidemics for 16 infectious diseases. The sensitivity of the system was more than 60% in ten diseases, and more than 80% in four diseases, the specificity was more than 95%, and the PPV was between 15.6% and 31.4%. They concluded that the early stage of epidemics of some infectious disease might be detected using this method. Hashimoto et al also measured the timeliness of the paediatric disease surveillance system by calculating the proportion of true positive warnings among the epidemic warnings one to four weeks before an epidemic. The proportion of warnings given by the surveillance system three to four weeks before an epidemic was between 16% and 68.6%. They proposed that warning three to four weeks before the start of an epidemic may be useful for instigating control activities before the start of an epidemic.

Belio-Blasco (2000) evaluated two active hospital surveillance methods, the retrospective review of medical records, and the discharge forms, for the detection of nosocomial infection in surgical patients using the prospective system as the reference standard. The prospective surveillance consisted of daily review of patients' records during hospitalization. The sensitivity and specificity of the two surveillance methods were compared to determine which method was best for detecting nosocomial infections in surgical patients. The overall sensitivity of review of the medical records was 88% and that of the discharge form was 56%, with a specificity of 99% in both cases. The retrospective method of review of the medical record was considered by the author to be the most efficient active surveillance method for detecting nosocomial infections in surgical patients.

Quenél et al (1994) evaluated a surveillance system, the health service based indicators that was used to identify influenza epidemics. Sensitivity, specificity and positive predictive value of health service based indicators (indicators of medical activity, absenteeism from work, drug consumption, hospital activity, etc.,) collected by an epidemiological network in the Paris region, was studied using identification of virus as the "gold standard". Based on the sensitivity, specificity and PPV it was concluded that the health service based indicators were useful for the surveillance of influenza epidemics.

The health service based indicators and virological indicators were also compared based on the time at which these indicators detected the outbreak. It was concluded that the

emergency visit indicator was the earliest indicator (one to four weeks before the virological indicator), followed by sick-leave reports made to the National Health Service and prescriptions issued by the general practitioners (one to three weeks).

Tsui et al (2002) assessed the value of ICD-9-coded chief complaints, routinely collected for detection of epidemics, by determining the sensitivity, specificity, positive predictive value and timeliness of influenza detection. A respiratory set (RS) of ICD-9-codes and influenza set (IS) were evaluated using the Serfling method (Serfling 1963) applied to Pneumonia and Influenza (P&I) data as the gold standard. To evaluate the data sets the Serfling method was applied to them and the sensitivity, specificity, PPV and timeliness of detection were studied. The timeliness of influenza detection was calculated by subtracting detection dates computed for P&I data by the Serfling method from those computed for RS and IS. The times of detection achieved with RS and IS data and the Serfling method were on average 0.67 and 0.33 weeks later than the time of detection achieved with P&I data. Based on their study, the authors concluded that, for detecting influenza epidemics using ICD-9 coded chief complaints, the Serfling method and data from a single emergency department are probably not useful for public health surveillance purposes because the timeliness on average was less when compared to that achieved using P&I death data.

Wagner et al (2001) explained the use of timeliness of detection of public health surveillance systems along with sensitivity, specificity and PPV. The authors reviewed previous work relevant to measuring timeliness and defining timeliness requirements (see

Quenél et al. 1994, Hashimoto et al. 2000 and Tsui et al. 2002). They suggested that timeliness of detection could be measured by subtracting the time of detection of outbreaks by the surveillance system from the time when the actual event occurred.

It can therefore be concluded based on the evidence from the literature review that sensitivity, specificity, positive predictive value, and time to detection are valid methods of comparisons of surveillance systems, or outbreak detection algorithms.

CHAPTER 3. OUTBREAK DETECTION ALGORITHMS AND EVALUATION CRITERIA

This chapter introduces the five outbreak detection algorithms and the criteria used in this study to compare these algorithms. A brief description of the algorithms along with the source of references is presented initially, and is followed by the description of the criteria used to compare these algorithms.

3.1. Outbreak Detection Algorithms

As a result of a literature review performed in the Spring 2002 (Center for Surveillance Coordination, Health Canada. 2002), eight outbreak detection algorithms were identified as being suitable for surveillance of nationally notifiable diseases. Due to the unavailability of some of these outbreak detection algorithms, only five of the algorithms were considered for evaluation in this study. These are:

- EARS-Historical (CDC-US)
- EARS-Non Historical (CDC-US)
- General Additive Models (BCCDC-Canada)
- Log Linear Regression Models (HPA-UK)
- Log Linear Regression Models (RIVM-Netherlands)

A brief description of these systems, along with references, is presented in the following sections.

3.1.1. EARS-Historical

The EARS-Historical (EARS-H) algorithm is a combination of the historical limits method (Stroup et al. 1989) and the seasonally adjusted CUSUM method (Hutwagner et al. 1997). The historical limits and the seasonally adjusted CUSUM algorithms were developed, for disease surveillance, at the Centers for Disease Control and Prevention (CDC) of the United States. The historical limits method compares the number of cases of a disease in a current four-week period (y_4), with the mean of the 15 totals of four-week intervals of historical information, (the preceding four-week period, the same four-week period, and the following four-week period of the previous five years). The algorithm flags when the current four-week count exceeds the upper limit of the $100(1 - \alpha)\%$ confidence interval of the expected value for the current four-week period.

$$\frac{y_4}{\mu_{15}} \geq 1 + \frac{2\sigma_{y_{15}}}{\mu_{15}}$$

where μ_{15} and $\sigma_{y_{15}}$ are the mean and standard deviation of the 15 totals of four-week intervals.

The seasonally adjusted CUSUM compares the cumulative sum of the current weekly count of a disease with the summary information of counts, for the same week for the previous five years. The method is based on the formula

$$S_t = \max\left(0, S_{t-1} + \frac{(y_t - (\mu_5 + k\sigma_{y_{15}}))}{\sigma_{y_{15}}}\right), S_0 = 0 \text{ and } k > 0$$

where y_t is the number of cases in the current week, μ_5 and $\sigma_{y_{15}}$ respectively are the mean and standard deviation of five years of historical data, k is a central reference value,

S_t is the CUSUM for the current week, and S_{t-1} ($t = 1, 2, \dots, 5$) represents the CUSUM for same weeks of the previous years. A flag is given if the current CUSUM is greater than a decision value h .

The EARS-H algorithm combines the flags from the historical limits and seasonally adjusted CUSUM algorithms. The algorithm flags if either the historical limits algorithm or the seasonally adjusted CUSUM algorithm produces a flag.

3.1.2. EARS-Non Historical

Hutwagner et al (2003) developed the EARS-Non Historical (EARS-NH) algorithm for the CDC for disease surveillance in the United States. The algorithm is used when historical data on a disease is not available. The algorithm is based on the CUSUM procedure

$$S_t = \max\left(0, S_{t-1} + \frac{(y_t - (\mu + k\sigma_{y_t}))}{\sigma_{y_t}}\right), S_0 = 0 \text{ and } k > 0$$

where y_t is the number of cases on the current day of interest, μ and σ_{y_t} are respectively the mean and standard deviation of the limited historical data, k is a central reference value, S_t is the CUSUM for the current day, and S_{t-1} ($t = 1, 2, \dots, 7$), is the CUSUM for the limited historical data.

The algorithm produces three levels of flags, C1-mild, C2-medium and C3-ultra. The C1-mild method uses the previous seven days of baseline information starting from the current day and flags if the current CUSUM exceeds the mean (μ) plus three (k)

standard deviations (σ_{y_t}). The C2-medium method uses seven days of baseline information starting from the third day back, and flags if the current CUSUM exceeds the mean (μ) plus three (k) standard deviations (σ_{y_t}) while the C3-ultra method uses seven days of baseline information starting from the third day back, and flags if the current CUSUM exceed mean (μ) plus two (k) standard deviations (σ_{y_t}). The EARS-NH algorithm combines the flags from the C1-mild, C2-medium, and C3-ultra methods and produces a flag if any of the three methods produces a flag.

3.1.3. Generalized Additive Model

White (personal correspondence) developed the GAM algorithm for disease surveillance for the British Columbia Centre for Disease Control (BCCDC), Canada. The algorithm is a generalized additive model based on

$$\log(\mu(t_i)) = \alpha + f(t) + c(t)$$

where $\hat{\mu}(t_i) = \exp(\hat{\alpha} + \hat{f}(t) + \hat{c}(t))$ is the expected value of the weekly counts $y(t)$, $t = 1, 2, \dots, T$, $f(t)$ and $c(t)$ respectively are the temporal and seasonal components that are estimated by LOESS, a curve-fitting smoothing technique based on local regression (Cleveland et al. 1988). The non-seasonal rate is estimated as

$$\hat{\lambda}(t) = \exp(\hat{\alpha} + \hat{f}(t)).$$

An odds-ratio is estimated based on whether the counts are greater or lesser than the non-seasonal rate. If the counts in a set of consecutive weeks (known as a streak) including the current week are consecutively greater than the expected value, the odds-ratio is estimated as

$$z_1 = \max_{streak} \left(\log \left(\frac{\hat{L}(t)}{\hat{U}(t)} \right) \right)$$

and if the counts are consecutively lesser than the expected value, the odds-ratio is estimated as

$$z_1 = \min_{streak} \left(\log \left(\frac{\hat{L}(t)}{\hat{U}(t)} \right) \right)$$

where $\hat{L}(t) = P(\sum Y_i \leq \sum y(i))$, $\hat{U}(t) = P(\sum Y_i \geq \sum y(i))$.

The change in non-seasonal rate is estimated as

$$z_2 = \frac{\hat{\mu}_i(t) - \hat{\mu}_i(t-1)}{\varpi \hat{\sigma} \sqrt{\hat{\mu}(t-1)}}$$

where ϖ is an adjustment for the correlation between $\mu(t)$ and $\mu(t-1)$. An alert score is

calculated as $\frac{z_1/2 + z_2/5}{K}$ (K is set to 1 by default) and subsequently an alert level, which

is the largest integer that is less than the alert score, is calculated. The algorithm generates a flag if the alert level is greater than or equal to one.

3.1.4. Log Linear Regression Model (HPA)

Farrington et al (1996) developed a log-linear regression algorithm (referred to as the HPA algorithm in this study) for disease surveillance for the Communicable Disease Surveillance Centre (CDSC) in the United Kingdom. The algorithm is a log-linear regression model adjusted for trends, seasonality, over-dispersion, and past outbreaks, and is based on the model

$$\log(\mu_i) = \alpha + \beta t_i$$

where μ_i is the mean of the base-line (historical data) counts y_i , and t_i represents the baseline weeks.

The algorithm uses 35 weeks of historical information, the preceding three weeks, same week, and the following three weeks of the previous five years. A score is calculated for the current weekly count y_{01} as

$$X = \frac{y_{01} - \hat{\mu}_{01}}{U - \hat{\mu}_{01}}$$

where U is the upper limit of a $100(1 - 2\alpha)\%$ confidence interval for y_{01} , and $\hat{\mu}_{01}$ is the expected count, which is estimated as

$$\hat{\mu}_{01} = \exp(\hat{\alpha} + \hat{\beta}t_{01})$$

The algorithm flags if the score $X > 1$.

3.1.5. Log Linear Regression Model (RIVM)

Widdowson et al (2003) developed a log-linear regression algorithm (referred to as the RIVM algorithm in this study) for infectious disease surveillance for the Infectious Disease Surveillance System (ISIS) in the Netherlands. The algorithm is a log-linear regression model adjusted for trends, seasonality, over-dispersion, and past outbreaks, and is based on a model similar to that of the HPA algorithm,

$$\log(\mu_i) = \alpha + \beta t_i$$

where μ_i is the mean of the base-line (historical data) counts y_i , and t_i represents the baseline weeks. The algorithm uses 45 weeks of baseline data, the preceding 4-week

period, same week and the following 4-week period of the previous five years. The algorithm uses two moving window periods, a 7-day total including the current day and a 4-week total including the current day calculated daily. An expected value for the current week is calculated as

$$\hat{\mu}_{01} = \exp(\hat{\alpha} + \hat{\beta}t_{01})$$

and compared with the current week count. The algorithm flags if the current weekly count $y_{01} > \hat{\mu}_{01} + 2.56\hat{\sigma}_{01}$ (threshold). If the 4-week total is less than five or if the 7-day total is less than three, flags are not generated even if the totals are greater than the threshold.

3.2. Evaluation Criteria

An outbreak detection algorithm used to detect an outbreak in a communicable disease will have the following possible test results:

- The algorithm detects an outbreak when an outbreak is present (True Positive).
- The algorithm does not detect an outbreak when an outbreak is present (False Negative).
- The algorithm concludes that an outbreak is present when an outbreak is not present (False Positive).
- The algorithm correctly concludes that an outbreak is not present (True Negative).

A 2×2 table of the possible test results of an algorithm is shown in Table 1.

		Outbreak Situation	
		Outbreak present	Outbreak not present
Test Result	Detected	n_{TP}	n_{FP}
	Not detected	n_{FN}	n_{TN}

Table 1: Possible test outcomes

where n_{TP} - is the number of true positives
 n_{FP} - is the number of false positives
 n_{FN} - is the number of false negatives
 n_{TN} - is the number of true negatives

The criteria used for comparing the algorithms in this study were how often they:

- Flagged an outbreak that is present in the surveillance data,
- Falsely flagged for the presence of an outbreak when one did not exist and,
- Did not flag an outbreak that is present in the surveillance data.

Additionally, the algorithms were evaluated on their performance based on their timeliness of detection. The following measures were used to compare the algorithms based on the above-mentioned criteria:

- Sensitivity
- Specificity
- Positive Predictive Value

- Timeliness

Two definitions of **sensitivity** were used in this study.

- 1) For the first method of estimating sensitivity, algorithms were applied to weekly observations and the sensitivity of each algorithm was defined as the ratio of the number of true outbreak weeks in a data set flagged by the algorithm to the total number of true outbreak weeks present in the data as defined by the gold standard (see sections 4.1.2 and 5.2) . This estimate of sensitivity, referred to as sensitivity weeks in this study, is useful to measure the duration of detection of an outbreak and is estimated as

$$\hat{Sens}_w = \frac{n_{TP}}{n_{TP} + n_{FN}} .$$

- 2) For the second method of estimating sensitivity, algorithms were still applied to weekly observations. If the algorithm flagged at least once (any week) during an outbreak period the flags were considered as true positives and were used in the estimation of sensitivity. This estimate of sensitivity, referred to as sensitivity run in this study, is useful to measure the detection of an outbreak and is estimated as

$$\hat{Sens}_r = \frac{n_{TP}}{n_{TP} + n_{FN}}$$

The number of true positives n_{TP} can be described as the number of successes in a trial of n^* (the total number of true outbreaks) events, and therefore follows a binomial distribution,

$$n_{TP} \sim B(n^*, p^*)$$

where $n^* = n_{TP} + n_{FN}$ and p^* is estimated as

$$\hat{p}^* = \frac{n_{TP}}{n_{TP} + n_{FN}}.$$

The variance of n_{TP} can thus be estimated as

$$\hat{Var}(n_{TP}) = n^* \hat{p}^* (1 - \hat{p}^*).$$

Using the above estimate of the variance of n_{TP} , the variance of the estimator of sensitivity can be estimated as

$$\hat{Var}(\hat{Sens}) = \hat{Var}\left(\frac{n_{TP}}{n^*}\right) = \frac{1}{n^{*2}} (n^* \hat{p}^* (1 - \hat{p}^*)) = \frac{\hat{p}^* (1 - \hat{p}^*)}{n^*}.$$

The standard deviation of sensitivity weeks and run can thus be obtained as

$$\hat{\sigma}_w = \sqrt{\hat{Var}(\hat{Sens}_w)} \text{ and } \hat{\sigma}_r = \sqrt{\hat{Var}(\hat{Sens}_r)}.$$

The **specificity** of an algorithm is defined as the ratio of the number of true non-outbreak weeks present in a data set that are not flagged by an algorithm to the total number of true non-outbreak weeks present in the data as defined by the gold standard. The specificity is estimated as

$$\hat{Spec} = \frac{n_{TN}}{n_{TN} + n_{FP}}.$$

The number of true negatives n_{TN} can be described as the number of successes in a trial of n' , (*the* total number of true non-outbreaks) events. Therefore,

$$n_{TN} \sim B(n', p')$$

where $n' = n_{TN} + n_{FP}$ and p' is estimated as

$$\hat{p}' = \frac{n_{TN}}{n_{TN} + n_{FP}}.$$

The variance of n_{TN} can thus be estimated as

$$\hat{Var}(n_{TN}) = n' \hat{p}' (1 - \hat{p}').$$

Using the variance estimate of n_{TN} , the variance of the estimator of specificity can be estimated as

$$\hat{Var}(\hat{Spec}) = \hat{\sigma}_s^2 = \hat{Var}\left(\frac{n_{TN}}{n'}\right) = \frac{1}{n'^2} (n' \hat{p}' (1 - \hat{p}')) = \frac{\hat{p}' (1 - \hat{p}')}{n'}.$$

The **Positive Predictive Value (PPV)** of an algorithm is defined as the ratio of the number of true outbreak weeks flagged by the algorithm to the total number of weeks that are flagged by the algorithm. The PPV can be estimated as

$$\hat{PPV} = \frac{n_{TP}}{n_{TP} + n_{FP}}.$$

The number of false positives n_{FP} can be described as the number of successes in a trial of n' (the total number of true non-outbreaks) events and therefore

$$n_{FP} \sim B(n', p'')$$

where, $n' = n_{TN} + n_{FP}$ and p'' is estimated as

$$\hat{p}'' = \frac{n_{FP}}{n_{TN} + n_{FP}}.$$

PPV is a function of n_{TP} and n_{FP} , therefore the variance of the PPV can be estimated using the delta method. The calculations for the variance are presented below.

If we let $X' = (n_{TP}, n_{FP})$, then the variance-covariance matrix of X can be estimated as

$$\hat{Var}(X) = \begin{bmatrix} \hat{Var}(n_{TP}) & \hat{Cov}(n_{TP}, n_{FP}) \\ \hat{Cov}(n_{TP}, n_{FP}) & \hat{Var}(n_{FP}) \end{bmatrix}.$$

Since n_{TP} and n_{FP} are two independent binomial random variables, $\hat{Cov}(n_{TP}, n_{FP}) = 0$.

Therefore,

$$\hat{Var}(X) = \begin{bmatrix} n^* \hat{p}^* (1 - \hat{p}^*) & 0 \\ 0 & n' \hat{p}'' (1 - \hat{p}'') \end{bmatrix}.$$

By the delta method,

$$\begin{aligned} \hat{Var}(PPV) &= \hat{\sigma}_p^2 = \begin{bmatrix} \frac{\partial PPV}{\partial n_{TP}} & \frac{\partial PPV}{\partial n_{FP}} \end{bmatrix} \hat{Var}(X) \begin{bmatrix} \frac{\partial PPV}{\partial n_{TP}} \\ \frac{\partial PPV}{\partial n_{FP}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial PPV}{\partial n_{TP}} & \frac{\partial PPV}{\partial n_{FP}} \end{bmatrix} \begin{bmatrix} n^* \hat{p}^* (1 - \hat{p}^*) & 0 \\ 0 & n' \hat{p}'' (1 - \hat{p}'') \end{bmatrix} \begin{bmatrix} \frac{\partial PPV}{\partial n_{TP}} \\ \frac{\partial PPV}{\partial n_{FP}} \end{bmatrix} \\ &= n^* \hat{p}^* (1 - \hat{p}^*) \left[\frac{\partial PPV}{\partial n_{TP}} \right]^2 + n' \hat{p}'' (1 - \hat{p}'') \left[\frac{\partial PPV}{\partial n_{FP}} \right]^2 \\ &= n^* \hat{p}^* (1 - \hat{p}^*) \left(\frac{n_{FP}}{(n_{TP} + n_{FP})^2} \right)^2 + n' \hat{p}'' (1 - \hat{p}'') \left(\frac{n_{TP}}{(n_{TP} + n_{FP})^2} \right)^2 \\ &= n^{*2} \hat{Var}(Sens) \left(\frac{n_{FP}}{(n_{TP} + n_{FP})^2} \right)^2 + n'^2 \hat{Var}(Spec) \left(\frac{n_{TP}}{(n_{TP} + n_{FP})^2} \right)^2. \end{aligned}$$

The measure **time to detection** is defined as the difference between the time an algorithm detects an outbreak and the time the outbreak started, as defined by the gold standard. A value of 0 for time to detection would indicate that the algorithm detected the outbreak at the beginning of the outbreak as described by the gold standard. A negative value for time to detection indicates that the algorithm detected the outbreak before the onset of the outbreak and a positive value for time to detection indicates that the algorithm detected the outbreak after the onset of the outbreak.

The algorithms were compared in the empirical and simulation study using the estimates of sensitivity, specificity, PPV and times to detection. The variance of these estimates was also estimated and taken into account when comparing the algorithms.

CHAPTER 4. EMPIRICAL STUDY

In the empirical study algorithms were applied to empirical data sets and compared using the evaluation criteria presented in Chapter 3. The algorithms were applied to the data sets on weekly observations to estimate sensitivity, specificity, PPV and times to detection (see section 3.2). While interpreting the estimates of the measures, the variance of the estimates was taken into consideration. The data sources, gold standard outbreaks, and the results from the empirical study are presented in this chapter.

4.1.1. Data Sources and Diseases

To determine the performance of the algorithms on a variety of disease types, the algorithms were applied, using the parameters specified by the developers of the algorithms (not all algorithms allow for change in parameters), to different communicable diseases that depict typical patterns such as seasonal variation, secular trends, and rare and common occurrences. The data sets used in this study were obtained from the Netherlands, Alberta, and British Columbia.

To determine the performance of the algorithms on large versus small data sets, algorithms were applied to regional health unit data, provincial laboratory data and national data. To keep the study at a manageable level, seven diseases were chosen from the above data sources. All data sets had at least five years of historical information. The number of years on which the algorithms were run was determined based upon the number of years of historical information available and the number of years for which outbreaks were identified.

The selected diseases, the disease type, the source of the data set and the number of years of data on which the algorithms were applied to (excluding 5 years of historical data) are displayed in Table 2 below.

No.	Disease	Type	Source	Number of Years
1	Hepatitis A	Common	Canada Provincial (British Columbia)	1997-2002 (6)
2	Hanta Virus	Rare	Canada Provincial (Alberta)	2000-2002 (3)
3	Salmonella Dublin	Rare	National (Netherlands)	2000(1)
4	Influenza (A + B)	Seasonal	Canada Provincial (Alberta)	1998-2002 (5)
5	Salmonella Enteritidis	Seasonal	National (Netherlands)	1996-2002 (7)
6	Gonorrhea	Secular	Canada Provincial (Alberta)	2000-2002 (3)
7	Meningococcal	Secular	Regional Health Unit (Alberta)	1999-2002 (4)

Table 2: Data sets used in Empirical Study

4.1.2. Gold Standard: Consensus method

When the data were obtained from the above-mentioned sources, the information of the true outbreaks that were present in the data were unknown. To calculate the evaluation measures some information regarding the outbreaks in the data sets, the week when the outbreaks had actually started and ended were required. Therefore, the data sets were examined by a group of experienced epidemiologists and, based on the number of cases, periods of interest were identified in each data set. These periods were used as the reference outbreaks or the gold standard outbreaks, to estimate the sensitivity, specificity, PPV and times to detection of outbreaks of the algorithms. Since these periods were

identified based on the disease counts and the epidemiological curve, the start and end of outbreaks as identified may or may not coincide with the actual start and end of the outbreak present in the data set that is unknown. Also some periods identified as possible outbreaks may have not been outbreaks in reality. However, as all the algorithms were applied to the data sets under the same circumstances, this factor may be ignored in this study. The results drawn from the empirical study are based on the assumption that the identified outbreaks are close to the true outbreak situations.

4.2. Empirical Evaluation

Each of the five algorithms was applied to the seven data sets (see Section 4.1.1). For each disease under consideration the sensitivity, specificity, positive predictive value and times to detection of each algorithm were estimated with respect to the gold standard as described in section 4.1.2. For comparison purposes, these measures appear in summary in Tables 5-11 in the Appendix A.

The performance of the algorithms was examined based on these measures and the results of the comparisons are described in the following sections. The tables were also sent to a committee selected from the epidemiologists and statisticians, who identified the outbreaks in the data sets, for expert review. While summarizing the performance of the algorithms on the seven data sets, the comments of the experts were taken into consideration.

4.2.1. Common Disease Type

Hepatitis A

The Hepatitis A data set had 11 years of data from 1992 to 2002. Out of the six years of data (1997 to 2002) on which the algorithms were applied nine outbreaks were identified by the epidemiologists. The number of true positives, false positives, true negatives, false negatives, and the time when algorithms detected the outbreaks were calculated for each algorithm using the outbreaks identified by the epidemiologists as the gold standard. The evaluation measures were estimated using these values and are presented in Table 5 of the Appendix A. The flags produced by the algorithms are presented in Figure 2 of the Appendix B. The small peaks in the figures represent the flags produced by the algorithms. These peaks show when the algorithm detected the outbreak, and how long it took before it stopped flagging. The gold standard outbreaks also presented in the figure with a red line, shows the start, duration and end of each outbreak as identified by the epidemiologists.

The examination of the evaluation measures show that the EARS-NH algorithm had the highest sensitivity run among the five algorithms ($\hat{Sens}_r = 77.7\%$, $\hat{\sigma}_r = 0.139$). The RIVM and EARS-H algorithms had the next highest sensitivity run of 55.5% with a standard deviation of 0.166. The sensitivity weeks of all algorithms were low and ranged from 1% to 32%.

All algorithms had similar specificities ranging from 94% to 100% with the standard deviation ranging from 0 to 0.016. Although the GAM and HPA algorithms had the

highest PPV (100% each, $\hat{\sigma}_p = 0$), these algorithms had lower sensitivity than the other algorithms, which ranged from 68% to 88%. The times to detection of outbreaks for the EARS-NH algorithm was in general lower than the other algorithms, except for one outbreak where the algorithm detected it nine weeks after the outbreak started. The EARS-H algorithm detected this outbreak one week after the outbreak started, while the remaining algorithms missed this outbreak. The times to detection of GAM, EARS-H and RIVM algorithms were similar (median times to detection ranging from 1 to 1.5). It appears that EARS-NH, EARS-H, and RIVM algorithms may be used for detecting outbreaks in Hepatitis A data set.

4.2.2. Rare Disease Type

Hanta Virus

The Hanta Virus data set had nine years of data from 1994 to 2002. The algorithms were applied to three years of data from 2002 to 2002. There were 11 outbreaks identified in these three years of data by the epidemiologists. As Hanta Virus was considered as a rare disease, it was decided that each case should be flagged as an outbreak. The evaluation measures were estimated by calculating the number of true positives, false positive, true negatives, false negatives, and the time at which the algorithm detected the outbreaks and are presented in Table 6 of the Appendix A. The flags produced by the algorithms are presented in Figure 3 of Appendix B.

Table 6 shows that the EARS-NH algorithm had the highest sensitivity of 81.8% with a standard deviation of 0.116 (since each case was considered as an outbreak the estimated

sensitivity run and week are equal). The EARS-H algorithm had the next highest sensitivity of 72.7% with a standard deviation of 0.134. The sensitivity of the remaining algorithms was less than 50% while the HPA algorithm missed all the outbreaks.

All algorithms had similar specificities ranging from 97% to 100% with the standard deviation ranging from 0 to 0.014. The GAM, EARS-NH, and EARS-H algorithms had a PPV of 100% each; the RIVM algorithm had a much lower PPV of 33.3% with a standard deviation of 0.049. All algorithms detected the outbreaks on the same week that the outbreak started (times to detection = 0). It appears that EARS-NH and EARS-H algorithms may be used to detect Hanta Virus outbreaks.

Salmonella Dublin

The Salmonella Dublin data set had 13 years of data from 1990 to 2002. When the epidemiologists examined the data they found that the potential outbreak periods might exist only in the year 2000. Therefore the algorithms were applied to the data set only for the year 2000 using the data from years 1995 to 1999 as the historical data. Two outbreaks were identified in the year 2000 of the Salmonella Dublin data set by the epidemiologists. The estimated measures of evaluation for the Salmonella Dublin data set are presented in Table 7 of the Appendix A, and the flags produced by the algorithms is presented in Figure 4 of Appendix B.

With an exception of RIVM algorithm, which detected only one outbreak, all algorithms detected both the outbreaks in the Salmonella Dublin data set. The sensitivity weeks of

EARS-H algorithm was higher than the other algorithms ($\hat{Sens}_w = 88.8$, $\hat{\sigma}_w = 0.104$). The sensitivity weeks of GAM and RIVM algorithms were each 66.6% with a standard deviation of 0.157. The sensitivity weeks of EARS-NH and HPA algorithms were lower than the other algorithms: 55.5% and 44.4% respectively with a standard deviation of 0.166 each.

The specificity of EARS-NH, GAM, and HPA algorithms were similar ranging from 93% to 100% with standard deviation ranging from 0 to 0.038. The specificity of EARS-H algorithm was slightly lower, 86% with a standard deviation of 0.053, while that of RIVM algorithm was much lower than the remaining algorithms (67.4%, $\hat{\sigma}_s = 0.071$). The PPV of EARS-H algorithm was the highest at 88.8% with a standard deviation of 0.104. GAM and RIVM algorithms had the next highest PPV (both at 66.6%, $\hat{\sigma}_p = 0.157$), while EARS-NH and HPA algorithms had lower PPV of 55.5% ($\hat{\sigma}_p = 0.166$) and 44.4% ($\hat{\sigma}_p = 0.166$) respectively.

The times to detection of EARS-NH and EARS-H algorithms were similar with each of them detecting the outbreaks on the same week the outbreak started. The times to detection of GAM and HPA algorithms were slightly higher with a median of one each, while that of RIVM was much lower. The RIVM algorithm detected one outbreak nine weeks before the start of the outbreak. Overall it appears that with an exception of RIVM algorithm all algorithms may be used to detection Salmonella Dublin outbreaks.

4.2.3. Seasonal Disease Type

Influenza A & B

The Influenza A & B data set had 10 years of data from 1993 to 2002. The algorithms were applied to five years of data from 1998 to 2002. Nine outbreaks were identified in the five of Influenza A & B data. The estimated measures of evaluation for the Influenza A & B data set is presented in Table 8 in the Appendix A, and the flags produced by the algorithms are presented in Figure 5 of Appendix B.

The examination of Table 8 shows that the sensitivity runs of the five algorithms range from 55% to 66% with the standard deviations ranging from 0.157 to 0.166. The sensitivity run of EARS-NH, GAM, and EARS-H algorithms was 66.6% and that of HPA and RIVM algorithms was 55.5%. The sensitivity weeks of the algorithms ranged from 38% to 63%. The sensitivity weeks of EARS-H and GAM algorithms were similar and were higher than the other algorithms (62.8% and 56.9% respectively and $\hat{\sigma}_w$ was 0.052 and 0.053 respectively). The sensitivity weeks of RIVM algorithm was slightly lower at 52.3% with a standard deviation of 0.054, while the sensitivity weeks of EARS-NH and HPA were lowest (around 39%).

The specificity of the algorithms was between 85% and 99%. The HPA algorithm had the highest specificity (99.4%, $\hat{\sigma}_s = 0.058$), while the RIVM algorithm had the lowest specificity (85.5%, $\hat{\sigma}_s = 0.0027$). The PPV of all algorithms followed similar patterns as that of the specificity, with the HPA algorithm having the highest PPV (97.1%, $\hat{\sigma}_p = 0.002$) and RIVM with the lowest PPV (64.3%, $\hat{\sigma}_p = 0.004$).

The times to detection of the algorithms were fairly similar for Influenza A & B data set, with the median times to detection ranging from –1 to 2. Overall it appears that any of the five algorithms may be used to detect Influenza A & B outbreaks.

Salmonella Enteritidis

The Salmonella Enteritidis data set had 13 years of data from 1990 to 2002. The algorithms were therefore applied to seven years of data from 1996 to 2002. Nine outbreaks were identified in the seven years of data. The estimated evaluation measures for Salmonella Enteritidis data set are presented in Table 9 in the Appendix A, and the flags produced by the algorithms are presented in Figure 6 of Appendix B.

Table 9 shows that the EARS-H algorithm had the highest sensitivity run (66.6, $\hat{\sigma}_r = 0.157$) out of the five algorithms. The RIVM algorithm had the next highest sensitivity run with 55.5% and a standard deviation of 0.166. The sensitivity run of the remaining algorithms was much lower and ranged from 22% to 33% with the standard deviations ranging from 0.139 to 0.157. The sensitivity weeks of EARS-H and RIVM algorithms were similar (41.2, $\hat{\sigma}_r = 0.119$) and were higher than the remaining algorithms.

The specificities of the algorithms ranged from 86% to 99% with the HPA algorithm having the highest specificity at 98.5% and a standard deviation of 0.064, and the EARS-NH algorithm with the lowest specificity at 86.5% and a standard deviation of 0.018. The PPV of all algorithms were lower than 40%. The GAM, EARS-H, and HPA algorithms

had higher PPV than RIVM and EARS-NH algorithms. The PPV of these algorithms ranged from 35% to 40% with standard deviation ranging from 0.02 to 0.03.

The times to detection were similar for all algorithms with the median time to detection of zero in each case. Overall it appeared that EARS-H and RIVM algorithms may be used to detect Salmonella Enteritidis outbreaks.

4.2.4. Secular Trend Disease Type

Gonorrhea

The Gonorrhea data set had seven years of data from 1995 to 2002. The algorithms were applied to three years of data from 2000 to 2003 using five years of data from 1995 to 1999 as historical data. Seven outbreaks were identified in the three years of data. The estimated measures of evaluation for Gonorrhea data set are presented in Table 10 in Appendix A, and the flags produced by the algorithms are presented in Figure 7 of Appendix B.

Table 10 shows that the EARS-H algorithm had the highest sensitivity run than the other algorithms (85.7%, $\hat{\sigma}_r = 0.132$). The RIVM algorithm had the next highest sensitivity run of 71.4% with a standard deviation of 0.171. The EARS-NH, GAM and HPA algorithms had the same value for the estimated sensitivity run (57.1, $\hat{\sigma}_r = 0.187$). The sensitivity weeks of RIVM algorithm was much higher than the other algorithms (93.9%, $\hat{\sigma}_w = 0.042$). The sensitivity weeks of the other algorithms ranged from 18% to 64%, GAM with the highest and EARS-NH with the lowest.

The specificity of EARS-NH, GAM, HPA, and EARS-H algorithms were similar ranging from 93% to 98% with standard deviation ranging from 0.011 to 0.022. Comparatively, the RIVM algorithm had a specificity of 71.5% with a standard deviation of 0.041. The PPV of the algorithms followed similar patterns as that of specificity, with the RIVM algorithm with the lowest PPV out of all the algorithms (47%, $\hat{\sigma}_p = 0.004$). The times to detection of all algorithms were very similar with the median of zero in all cases. Overall it appears that any of the five algorithms may be used in detecting outbreaks of Gonorrhea disease.

Meningococcal

The Meningococcal data set had 10 years of data from 1993 to 2002. The algorithms were applied to four years of data from 1999 to 2002. Seven outbreaks were identified in the four years of data by the epidemiologist. The algorithms were applied to the seven years of data and the evaluation measures were estimated using the identified outbreaks as the gold standard. The estimated measures of evaluation for Meningococcal data set are presented in Table 11 of Appendix A, and the flags produced by the algorithms are presented in Figure 8 of Appendix B.

Table 11 shows that the EARS-H algorithm had the highest sensitivity run out of all the five algorithms (100%). The EARS-NH algorithm had the next highest sensitivity run of 85.7% and a standard deviation of 0.132. The RIVM and GAM algorithms had same sensitivity run of 57.1% with a standard deviation of 0.187 in both cases, while the HPA

algorithm had the lowest sensitivity run of 42.9% with a standard deviation of 0.187. The EARS-H and RIVM algorithms had similar sensitivity weeks, 62.5% and 66.6% respectively with standard deviations 0.07 and 0.008 respectively. The sensitivity of these algorithms was higher than the other algorithms.

The specificity of the algorithms ranged from 81% to 99%. The HPA algorithm had the highest specificity (99.3%, $\hat{\sigma}_s = 0.007$), while the RIVM had the lowest specificity (80.9%, $\hat{\sigma}_s = 0.032$). The PPV of the algorithms also followed similar patterns as that of the specificity, the HPA with the highest PPV (90%, $\hat{\sigma}_p = 0.008$) and the RIVM with the lowest PPV (52.5%, $\hat{\sigma}_p = 0.065$).

The times to detection of HPA algorithm were generally higher than that of the other algorithms. The median time to detection of this algorithm was three while that of the other algorithms was zero. In general EARS-H, EARS-NH, GAM, and RIVM algorithms may be used in detecting Meningococcal outbreaks.

4.2.5. Overall Performance

To further support the evidence from the empirical evaluation exact binomial confidence intervals were estimated for the evaluation measures. The confidence intervals were estimated using the exact binomial-beta relationship presented below proposed by Feller 1968.

If $B(k; n, p) = \sum_{v=0}^k b(v; n, p)$ is the probability of at most k successes in n trials, then

$$1 - B(k; n, p) = n \binom{n-1}{k} \int_0^p t^k (1-t)^{n-k-1} dt.$$

The exact binomial confidence intervals of the evaluation measures with the family level of significance at $\alpha = 0.05$ were estimated and are presented in Table 22 in Appendix D. Within each disease, the exact confidence intervals were examined to see if there was any overlap in the confidence intervals for each evaluation measures for the algorithms. If there was an overlap it was concluded that there was no evidence that the evaluation measures of the algorithms were different. The confidence intervals showed that in general there was no significant difference between the evaluation measures of the algorithms for each data set. Therefore, it can be concluded that any of the five algorithms may be used for detecting outbreaks in the seven data sets.

Since the algorithms were applied to only one data set from the vast majority of common disease, it may not be appropriate to generalize the results to all common diseases. However, the results may be extended to the diseases that follow similar patterns as that of the Hepatitis A. By referring to common disease in this study, we intend to cover those diseases with patterns similar to that of Hepatitis A.

Overall any of the five algorithms may be used for outbreaks in common, rare, seasonal diseases and diseases with a secular trend. These algorithms had an overall high sensitivity and a reasonable specificity and PPV. The algorithms were also timely in detecting most of the outbreaks. The GAM and HPA algorithms had overall good specificity and PPV. In most cases the GAM algorithm was better than the HPA and RIVM algorithms. However, it should be noted that the GAM, HPA and RIVM

algorithms could be tweaked to fit the type of disease. The RIVM algorithm did not perform well in detecting outbreaks in rare diseases. However, it should be noted that in practice the threshold in the RIVM algorithms is manually set to zero so as to detect each case of rare disease. Similarly, in the case of HPA, the parameter α can be changed depending on the disease type in order to achieve the desired level of sensitivity and specificity. Therefore, the results may have been different if the parameters in the algorithms were set at a different level. It is totally possible to obtain different results if different data sets were used. Also it should be noted that the algorithms were run on one or at most two data sets for each type of disease, and also that each disease is different by itself. Therefore, it may not be appropriate to generalize the results from this study to all common, rare, seasonal, and secular type of disease. To make such conclusions the algorithms need to be applied to more data sets from each type of disease and to more number of years of data for each disease. Therefore, these aspects have to be considered while interpreting the results.

CHAPTER 5. SIMULATION STUDY

In the simulation study, the data sets were simulated using the empirical probability distribution of the data sets used in the empirical study. To keep the study at a manageable level and yet to cover various types of diseases, data sets depicting common, rare, and seasonal disease patterns, using one data set from each type of disease from the empirical study, were generated for the simulation study. Fifty data sets for the three types of disease were generated for 10 years (almost 3660 observations allowing for leap years) of data. To evaluate the algorithms using simulated data sets, the same evaluation criteria as were used for the empirical study were employed. Additionally, statistical analysis such as multivariate analysis of variance (MANOVA) and graphical analysis such as empirical cumulative distribution function plot (CDF Plot) and interaction plots were produced to enable comparisons of the algorithms. The algorithms were applied to weekly observations in a way similar to that of the empirical study.

5.1. Data simulation

The baseline counts for the simulated data sets were generated using the empirical probability distribution of the baseline counts from the real data sets that were used in the empirical study. The generation of baseline counts was disease-dependent and therefore different scenarios had to be considered in the generation of these data sets. The procedures used in the generation are described in subsequent pages.

To ensure that the generated data sets captured the patterns in the empirical data sets, many simulations of the generated data sets were plotted along with the appropriate empirical data set used in its generation and examined. One such example of data that were generated for common, rare, and seasonal disease is presented in figures 9, 10 and 11 respectively. Upon examination of the graphs, it appeared that the generated data sets reasonably depicted the patterns in the empirical data sets. The conclusions from the simulation study are based on the assumption that the generated data sets sufficiently portray the patterns in the empirical data sets.

Common disease type

The empirical probability distribution of the Hepatitis A data set was used to generate data sets for a common disease in the simulation study. The Hepatitis A data set had a decreasing trend and a change in variation displaying a cyclical pattern. Therefore, to replicate the decreasing trend and the change in variation, the Hepatitis A data series was divided into three sections, such that each section had a different variation and trend. The empirical probability distribution of disease counts for each section was determined using simple statements in the R statistical package (see Appendix C). The baseline series for each section of the data sets was then generated using the empirical probability distribution of the corresponding section of the Hepatitis A baseline counts. For example, the first section of the Hepatitis A baseline data set took on values from 0 to 5 and 7 with probability of occurrence 0.55, 0.24, 0.13, 0.06, 0.009, 0.007, and 0.001 respectively. Therefore, the first section of the generated baseline series would take on values 0 to 5 and 7 with the above probabilities. The baseline counts in the remaining two sections were generated in a similar fashion. These three sections were combined to

form the whole baseline series for the simulated common disease. Using the above procedure, 50 data sets were generated and the algorithms were run on these data sets.

Rare disease type

The empirical probability distribution of the Salmonella Dublin data set (Figure 10) was used as a reference to generate data sets for a rare disease. The disease counts in the Salmonella Dublin data set occurred in a random fashion, and therefore no adjustment as in the case of common diseases was required to replicate the rare disease. The probability of occurrence of baseline counts, for the Salmonella Dublin data set were determined using statements in R statistical package (see Appendix C). The baseline series was then generated using the empirical probability distribution of the baseline Salmonella Dublin counts. For example, the baseline counts of the Salmonella Dublin data series took on values 0 to 3 with probability of occurrence of 0.97 0.03, 0.003, and 0.003 respectively. Therefore the baseline of the generated data would take on the values 0 to 3 with the above probabilities. Fifty data sets for the rare disease were generated using the above methodology, and the algorithms were run on these data sets for comparison.

Seasonal disease type

The empirical probability distribution of the baseline Influenza A & B data set was used to generate the baseline counts for a seasonal disease. The Influenza A & B data set had a seasonal pattern, with a slow increase, a peak, and a slow decrease in the seasonal curve. These curves in the data set were considered as outbreaks. Therefore the empirical probability distribution of the non-curve counts (baseline counts) was used in

the generation of the baseline series for the simulated data sets. The probability of occurrence of the non-curve counts was determined using statements in R and these probabilities were used to generate the baseline counts. For example, the baseline counts in the Influenza A & B data set took on values from 0 to 2 with probabilities of occurrence 0.96, 0.034, and 0.002 respectively. Therefore, the baseline counts for the simulated data set would take on values 0 to 2 with the above probabilities. These counts formed the whole baseline series for the simulated data set. Using the above methodology, 50 simulated data sets were generated for the seasonal disease, and the algorithms were run on these data sets to compare the algorithms.

5.2. Gold Standard

The outbreak counts for the simulated data sets were generated using the empirical probability distribution of the outbreak counts from the real data sets used in the empirical study. These generated outbreak counts were used as gold standard outbreaks, to estimate the sensitivity, specificity, PPV and times to detection of the algorithms. While generating the outbreak counts; a) the number of outbreaks, n , b) the duration of each outbreak (number of weeks the outbreak lasted), $l_i, i = 1, 2, \dots, n$ and c) the placement of the outbreaks in the real data sets were considered. For example, if the real data set had two years worth of data and had five outbreaks, then a simulated data set with 10 years of data was assumed to have 25 outbreaks ($10 \times 5/2$). In true situations outbreaks in a disease do not occur with the same frequencies each time. Therefore the number of outbreaks to be generated in each simulated data set was randomly chosen between $n - 2$ and $n + 2$ (number of outbreaks in the real data set ± 2). The addition and subtraction of

two from n was chosen arbitrarily and is carried out throughout this study. However, it may be possible to choose different numbers to change the variation. In true situations as the length of outbreaks is not fixed; therefore, to randomize the duration of each generated outbreak, the outbreak durations were chosen randomly between $\min(l_i) - 2$ and $\max(l_i) + 2$ (l_i is the duration of each outbreak (in weeks) in the real data set). The outbreak counts, the number of outbreaks, the length of each outbreak, and the points where the outbreaks were inserted were disease-dependent. Therefore different scenarios had to be considered in the generation of data sets for common, rare and seasonal diseases. The methodologies used in the generation are described below.

Common Disease Type

The empirical probability distribution of the Hepatitis A data set was used to generate data sets for the common disease. As mentioned earlier in section 4.2.1, there were six years of data and nine outbreaks in the Hepatitis A data set. Therefore, the number of outbreaks in a 10-year data was determined as 15 ($= 9 \times 10 / 6$). The number of outbreaks (n_{outbk}) to be inserted in the simulated data set was chosen then randomly between 13 and 17 (i.e. 15 ± 2). The minimum and maximum of the duration of the outbreaks in the Hepatitis A data was three and 24 respectively. As the outbreak durations were calculated from weekly series, the outbreak durations were therefore multiplied by seven so as to replicate the durations of outbreaks in a daily series. The outbreak durations for the daily series of the simulated data set were chosen randomly between 19 ($= (3 \times 7) - 2$) and 170 ($= (24 \times 7) + 2$).

As described in section 5.1, the Hepatitis A data set had a decreasing trend and a change in variation displaying a cyclical pattern. Therefore, the data series was divided into three sections and the probability of occurrence of counts, of each section was determined. The n_{outbk} points, where the n_{outbk} outbreaks were to be inserted were chosen randomly (random sample from 1 to 3660). If the random point fell in the first section, then the empirical probability distribution of the first section of the Hepatitis A outbreak counts was used to generate the outbreak counts. If the random point fell in the second section, then the empirical probability distribution of the second section of the Hepatitis A outbreak counts was used to generate the outbreak counts and so on. For example, the first section of the Hepatitis A data set took on values from 0 to 8 with probability of occurrence of 0.43, 0.18, 0.14, 0.12, 0.08, 0.03, 0.02, and 0.004 respectively. Therefore, if the random point to insert outbreak fell in the first section then the outbreak counts at the random point took on the values 0 to 8 with the above probabilities. The outbreaks in the other two sections were generated in a similar fashion. These outbreaks formed the whole outbreak series for the simulated data set.

Rare Disease Type

The empirical probability distribution of the Salmonella Dublin data set was used to generate data sets for a rare disease. The Salmonella Dublin data set had one year of data and two outbreaks. Therefore, the number of outbreaks in a 10-year worth data was calculated as 20 ($10 \times 2/1$). The number of outbreaks (n_{outbk}) to be inserted in the simulated data set was chosen randomly between 18 and 22 (20 ± 2). The durations of the two outbreaks in the Salmonella Dublin data were three and six respectively. As the

outbreak durations were calculated from weekly series, therefore the outbreak durations were multiplied by seven so as to replicate the durations of outbreaks in a daily series. The outbreak durations for the daily series of simulated data set were chosen randomly between $19 (= (3 \times 7) - 2)$ and $40 (= (6 \times 7) + 2)$.

The probability of occurrence of the outbreak counts in the Salmonella Dublin data set was determined. The n_{outbk} points where the n_{outbk} outbreaks were to be inserted was chosen randomly (random sample from 1 to 3660) and the generated outbreaks using the empirical probability distribution of the Salmonella Dublin data were inserted at these points. For example, the Salmonella data set took on values from 0 to 2 with probability of occurrence of 0.76, 0.18, and 0.06 respectively. Therefore, the outbreak counts at the random point would take on the values from 0 to 2 with the above probabilities. These outbreaks formed the whole outbreak series for the simulated data set.

Seasonal Disease Type

The empirical probability distribution of the Influenza A & B data set was used to generate data sets for a seasonal disease. The Influenza A & B data set had five years of data and nine outbreaks. Therefore, the number of outbreaks in a 10-year worth data was determined as 18 (i.e. $10 \times 9 / 5$). The number of outbreaks (n_{outbk}) to be inserted in the simulated data set was then chosen randomly between 16 and 20 (18 ± 2). The minimum and maximum of the duration of the outbreaks in the Hepatitis A data were one and 21 respectively. As the outbreak durations were calculated from weekly series, therefore the outbreak durations were multiplied by seven so as to replicate the durations of outbreaks

in a daily series. The outbreak durations for the daily series of the simulated data set were chosen randomly between 5 ($= (1 \times 7) - 2$) and 149 ($= (21 \times 7) + 2$).

As mentioned in section 5.1, the Influenza A & B data set had a seasonal pattern. To portray the pattern in a seasonal disease with a slow increase, a peak and a slow decrease, the seasonal curve in the Influenza A & B data set was divided into three sections and the probability of occurrence of counts, for each section was determined. Seasonal disease was considered as a special case when choosing points where outbreaks were to be inserted. By definition a seasonal disease occurs at approximately the same time of each year. Using the pattern of occurrence of outbreaks in the real data set, n_{outbk} points, where the n_{outbk} outbreaks were to be inserted were generated roughly at the same point by randomly selecting day 1 (week 1) to day 35 (week 5) of each year. These numbers were chosen arbitrarily by examining the patterns in the Influenza A & B data set. At each random point a seasonal curve was generated, using the empirical probability distributions of the three sections. The increasing and decreasing curves were generated using the empirical probability distribution of the Influenza A & B data counts that were less than their 75th quartile. The peak of the curve was generated using the empirical probability distribution of the Influenza A & B data counts that were greater than their 25th quartile. These quartiles were chosen on a trial and error basis while trying to imitate the Influenza A & B data set. For example, the 25th quartile and the 75th quartile for the Influenza A & B data are one and seven respectively. The probabilities with which the observations 1 to 6 (observations less than 75th quartile) occur are 0.29, 0.23, 0.14, 0.08, 0.09, 0.07, and 0.1 respectively. Therefore, the increasing and decreasing seasonal curve

would take on value 1 to 6 with the above probabilities. The peaks of the seasonal curves were generated as in a similar fashion as described above. These outbreaks formed the whole outbreak series for the simulated data set.

5.3. Simulation Evaluation

The five algorithms were applied to 50 simulated data sets for each of common, rare and seasonal diseases. Estimates of sensitivity (weeks and run), specificity, positive predictive value, and time to detection of each algorithm were calculated with respect to the gold standard as described in section 5.2 for each data set. To facilitate comparison of the algorithms with respect to these measures, the minimum, 25th percentile, mean, median, 75th percentile, maximum and standard deviation of the 50 estimates of sensitivity (weeks and run), specificity, PPV, and time to detection and their standard deviations were estimated and tabulated for each disease. These values are presented in, Tables 12, 14 and 16 of Appendix A.

5.3.1. Cumulative Distribution Function Plot

A cumulative distribution function plot (CDF plot) of sensitivity weeks, sensitivity run, specificity and PPV for each algorithm was plotted to provide a graphical representation of the differences or similarities in the algorithms. The graphs were not plotted for the time to detection since the mean, median and the quartiles of time to detection were calculated only when algorithms detected the outbreaks. The missed outbreaks were not taken into account in the estimation of time to detection. Therefore it was considered to

be more appropriate to represent the time to detection in a box plot (see Figures 12, 13 and 14 in the Appendix B). The results of these plots are described in the section 5.3.2.

Figure 1 shows the cumulative distribution function of sensitivity weeks of the algorithms for the common disease. Ideally it is expected that the sensitivity weeks of the algorithms should be high and that there should be a sharp rise in the CDF curve. The CDF plot in Figure 1 suggests that the RIVM algorithm may have better sensitivity weeks when compared to the other methods. We note that the EARS-NH, HPA and EARS-H are not very different with respect to sensitivity weeks. GAM is somewhere between RIVM and the other methods in terms of its performance in detecting outbreaks. The CDF plots of the remaining measures for each type of disease are displayed in Figures 15 through 25.

Figure 1: CDF plot of Sensitivity Weeks for Common Disease (Simulation Study)

The CDF plots for the four evaluation measures for the diseases show the following results.

Common disease

Figures 1, 15, 16 and 17 show that, the RIVM algorithm had better sensitivity for common diseases than the other algorithms and the rest of the algorithms had almost the same sensitivity. The specificity of the algorithm was greater than 80 and the PPV was greater than 45%. The EARS-H and EARS-NH algorithms had the next best sensitivity. The EARS-H, GAM, and HPA algorithms had better specificity and PPV than the other algorithms.

Rare disease

Figures 18 through 21 present the empirical CDF plot of the measures for rare diseases. The figures show that EARS-H algorithm may have better sensitivity for rare diseases. The specificity of the algorithm was greater than 85% and the PPV was greater than 30%. The RIVM algorithm had the next best sensitivity, although this algorithm had the lowest specificity and PPV. The GAM and HPA algorithms had better specificity and PPV than the other algorithms.

Seasonal disease

Figures 22 through 25 present the empirical CDF plot of the measures for seasonal diseases. The figures show that the GAM and EARS-NH (almost equal sensitivity)

algorithms had better sensitivity than the other algorithms. These algorithms also had good specificity and PPV. The HPA algorithm had the best specificity out of all the algorithms. The EARS-H and RIVM algorithms performed similarly with respect to all the measures.

5.3.2. Box Plots of Times to Detection

The box plots of times to detection of the algorithms for common disease is displayed in Figure 12 of Appendix B. The box plot shows that the median times to detection of the algorithms are fairly close. However, the RIVM algorithm has many outliers, with times to detection taking values from -34 to 15 . In the empirical study we had suggested that the algorithm that has a negative time to detection generally is good in detecting the outbreaks earlier. But in the case of the simulation study the start of the outbreaks were not arbitrary as in the case of the empirical study, but was set at a particular point. So it is not meaningful if the algorithms detect the outbreaks say -15 weeks earlier. These detected outbreaks may be considered as false alarms. Therefore, it can be concluded that all algorithms have good times to detection for common diseases.

Figure 13 displays the box plots of times to detection for rare disease. In the case of rare disease, the EARS-H algorithm has a better time to detection when compared to the other algorithms. This algorithm has the least variability ($\hat{\sigma}_t = 1.16$) and had a median time to detection of one week. As in the case of the common disease the RIVM algorithm has

many outliers. In general all algorithms may be said to have good times to detection for rare diseases.

The box plot of times to detection of algorithms for the seasonal disease is displayed in Figure 14. The box plots show that the median times to detection of the algorithm differ greatly. The EARS-NH algorithm had a median of zero and had the least variability ($SD = 1.59$). The 25th and 75th percentile of times to detection of this algorithm coincide with the median. Therefore, it may be concluded that the EARS-NH algorithm has better times to detection than the other algorithms.

5.3.3. Two-way Multivariate Analysis of Variance (MANOVA)

To obtain a better sense of the effect of the algorithms on the evaluation measures on the types of diseases a two-way MANOVA was performed. The design for MANOVA assumes that the random samples drawn from the different populations are independent. Since the algorithms were applied to all 50 data sets that were generated for each type of disease, if the analysis is performed with algorithms as treatments, and diseases as blocks, the responses which are the evaluation measures obtained by applying the algorithms to the same data set may not be independent. Therefore, for each of the three diseases, the algorithms were randomly assigned to 10 out of 50 generated data sets. The data sets for a disease were randomly assigned numbers from one to fifty, and 10 numbers were chosen at random from these 50 numbers. An algorithm was chosen randomly out of the five and was applied to the 10 data sets that corresponded to the 10 chosen numbers. Ten

more numbers were chosen randomly from the remaining 40 numbers and another algorithm was chosen randomly from the remaining four. This algorithm was in turn applied to the 10 data sets that corresponded to the 10 chosen numbers. Using the same methodology, the remaining three algorithms were assigned to 10 randomly chosen data sets. The same procedure was followed for the other two diseases.

Multivariate Analysis of Variance (MANOVA) assumes that the responses follow a multivariate normal distribution. Therefore, to investigate the normality assumptions on the responses, the evaluation measures, normal probability plots and box plots of the residuals of the measures were graphed. These graphs are displayed in Figures 34 through 42. Although there are some outliers, there was no severe violation of normality on the error terms. The error terms were therefore assumed to follow a normal distribution.

Subsequently, a two-way MANOVA was performed, with the evaluation measures as responses, algorithms as the treatments and diseases as the blocks with an interaction effect between algorithms and diseases. The model used was

$$Y_{ij\ k} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ij\ k}$$

$$i = 1, 2, \dots, 5$$

$$j = 1, 2, 3$$

$$k = 1, 2, \dots, 10$$

where $Y_{ij\ k}$ is a 10×1 vector, where the columns represent sensitivity weeks, sensitivity run, specificity, and PPV for each of the three diseases. The parameter β_j represents

effect of disease j , τ_i represents the effect of algorithm i and γ_{ij} represents the algorithm i disease j interaction effect. We assume for any given measure

$$\sum_{i=1}^5 \tau_i = \sum_{j=1}^3 \beta_j = \sum_{i=1}^5 \gamma_{ij} = \sum_{j=1}^3 \gamma_{ij} = 0$$

and in general $\varepsilon_{ijk} \sim N(0, \Sigma)$

The null hypothesis of no algorithm-disease interaction effect

$$H_0 : \gamma_{11} = \gamma_{12} = \dots \gamma_{53} = 0 \text{ (No algorithm-disease interaction effect)}$$

against

$$H_a : \text{At least one } \gamma_{ij} \neq 0$$

was tested. The approximate F-statistic for the Wilk's Lambda statistic for testing the hypothesis of no algorithm-disease interaction effect was $F = 16.3$ and the p-value of the test was $p < 0.0001$. The null hypothesis of no algorithm-disease interaction effect was therefore rejected at a 5% level of significance. Since the interaction effects were significant, tests for algorithms effects was not performed. Instead, interaction plot analyses and univariate analyses with evaluation measures as responses were conducted to examine the interaction effect.

5.3.4. Interaction Plots between Algorithms and Diseases

The interaction plots on the average of the measures for all the algorithms show that interactions exist between the algorithms and the diseases. This confirms the results from the test of no interaction effects performed using the two-way MANOVA, which was discussed earlier in section 5.3.3.

Figure 26 and Figure 27 respectively display the average sensitivity weeks by disease and algorithm respectively. In Figure 26 each curve represents an algorithm, whereas in Figure 27, each curve represents a disease. An examination of the plots suggests that the EARS-H and RIVM perform similarly in detecting outbreaks in rare and seasonal disease and that there is not much difference in the sensitivity of RIVM and EARS-H algorithms for both diseases. GAM and RIVM algorithms perform similarly with respect to sensitivity weeks for common and rare diseases. The sensitivity of RIVM algorithm was however much higher than that of GAM.

The performance of the GAM algorithm in detecting outbreaks was similar across the three diseases, as was that of the HPA algorithm. The performance of RIVM in detecting outbreaks was similar for common and rare diseases, as was that of the EARS-NH algorithm. The RIVM algorithm had the highest sensitivity for common diseases; EARS-H and RIVM highest sensitivity for rare diseases, and GAM and EARS-NH had the sensitivity for seasonal diseases.

Figure 28 and Figure 29 respectively display the average sensitivity run by disease and algorithm. Each curve in Figure 27 represents an algorithm, whereas each curve in Figure 28 represents a disease. While the patterns in the detection of outbreaks for common and rare diseases were similar in RIVM and HPA algorithms, the patterns in detection of outbreaks for rare and seasonal diseases was similar in EARS-H and RIVM algorithms. However, the sensitivity run of RIVM algorithm was higher than that of HPA

algorithm and the sensitivity of EARS-H algorithm was higher than that of the RIVM algorithm. The EARS-H, GAM and EARS-NH algorithms also performed similarly in terms of the detection of outbreaks in common and rare diseases. However, the sensitivity of EARS-H algorithm was higher than the other two algorithms.

The sensitivity run of all five algorithms for common disease seemed to be similar. For rare disease, the HPA algorithm seemed to be different from the other four algorithms. These four algorithms seemed to have similar sensitivities for rare diseases. The EARS-NH, GAM and EARS-H algorithms performed similarly for seasonal disease. The performance of EARS-H algorithm in detecting outbreaks across the three diseases was similar. However, any such pattern was not clear from the plot for the remaining four algorithms.

Figure 30 and Figure 31 respectively display the average specificity by disease and algorithm. In general, all the algorithms had a high specificity ($> 85\%$) for all three diseases. The specificity of all algorithms for common disease was similar except for the RIVM algorithm. The HPA and GAM algorithms had similar specificities for rare disease; the EARS-H and EARS-NH algorithms had similar specificity for rare diseases. The specificity of GAM and EARS-NH algorithms were similar for seasonal disease, as was the specificity of RIVM and EARS-H algorithms. It was noted that with an exception of EARS-H algorithm, the specificities of the remaining four algorithms across the three diseases were similar. The RIVM algorithm had the lowest specificity of all the five

algorithms. Overall HPA algorithm followed by GAM algorithm had the highest specificity for all the three diseases.

Figure 32 and Figure 33 respectively display the average PPV by disease and algorithm. The EARS-H and RIVM algorithms, and EARS-H, HPA and GAM algorithms performed similarly with respect to the PPV for the common diseases; and RIVM, EARS-H, and EARS-NH; and GAM and HPA algorithms performed similarly for rare; and with an exception of HPA algorithm the other four algorithms performed similarly for seasonal diseases. The PPV of GAM as well as HPA algorithm were similar for the three diseases; and that of EARS-NH algorithm was similar for common and seasonal diseases. No such conclusions could however be made for EARS-H and RIVM algorithms. In general the HPA and GAM algorithms maximized the PPV for the three diseases.

5.3.5. Univariate Analysis of Variance (ANOVA)

To provide further insight of the interaction between algorithms and diseases, individual univariate analysis of variance (ANOVA) was performed for the four evaluation measures with algorithms as treatments and diseases as blocks. Subsequently, Tukey's pair-wise multiple comparisons to test the algorithm-disease interaction effect was conducted. The model with an algorithm-interaction effect used was

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

$$i = 1, 2, \dots, 5$$

$$j = 1, 2, 3$$

$$k = 1, 2, \dots, 10$$

where Y_{ijk} is a 50×1 vector, with values representing sensitivity weeks (or sensitivity run, specificity, PPV) for the three diseases. The β_j represents the effect of disease j , τ_i represents the effect of algorithm i and γ_{ij} represents the algorithm i disease j interaction effect. We assume

$$\sum_{i=1}^5 \tau_i = \sum_{j=1}^3 \beta_j = \sum_{i=1}^5 \gamma_{ij} = \sum_{j=1}^3 \gamma_{ij} = 0$$

$$\text{and } \varepsilon_{ijk} \sim N(0, \sigma^2).$$

The null hypothesis of no algorithm-disease interaction effect

$$H_0 : \gamma_{11} = \gamma_{12} = \dots \gamma_{53} = 0 \text{ (No algorithm-disease interaction effect)}$$

against

$$H_a : \text{At least one } \gamma_{ij} \neq 0$$

was tested. The F-statistic and p-value of the univariate tests are displayed in Table 3. The hypotheses of no algorithm-interaction effect were rejected at a 5% level of significance. Hence we conclude that interaction effects between algorithm and diseases are present.

Measure	F-Statistic	Num DF	Denom DF	P-value
Sensitivity Weeks	24.33	8	135	<0.0001
Sensitivity Run	5.16	8	135	<0.0001
Specificity	12.83	8	135	<0.0001
PPV	12.44	8	135	<0.0001

Table 3: Univariate analysis for testing the interaction effects (Simulation Study)

To investigate the nature of the effects Tukey's pairwise comparisons with family level of significance set at $\alpha = 0.05$ was performed to test the hypotheses of no difference in the average for each evaluation measure for the three diseases.

$$H_0 : \mu_{ij} = \mu_{i'j'}, \quad ij \neq i'j'$$

against

$$H_a : \mu_{ij} \neq \mu_{i'j'}$$

where $i, i' = 1, 2, \dots, 5$ and $j, j' = 1, 2, 3$, and μ_{ij} is the average evaluation measure for algorithm i disease j .

The p-values from the Tukey's pair-wise comparison for sensitivity weeks, sensitivity run, specificity, and PPV are given in Tables 18, 19, 20 and 21. The p-values of the test that are significant at a 5% level of significance are highlighted in the tables.

Table 18 shows the p-values of the Tukey's pair-wise comparison of algorithm disease combination performed on the estimated sensitivity weeks. For common disease, there was a significant difference between the average sensitivity weeks of EARS-NH, GAM, and RIVM algorithms. Also there was a significant difference between GAM, EARS-H, and RIVM algorithms, and between HPA and RIVM algorithms. This shows that the average sensitivity weeks of the RIVM and GAM algorithms are significantly different from that of the other algorithms. An inspection of average sensitivity weeks of these algorithms shows that the sensitivity weeks of the RIVM algorithm followed by the GAM algorithm was higher than the other algorithms for the common disease. For rare disease, the average sensitivity weeks between EARS-NH, EARS-H, and RIVM

algorithms were significantly different. There was also a significant difference between the average sensitivity weeks of GAM, EARS-H, HPA and RIVM algorithms for the rare diseases, and between EARS-H and HPA, and HPA and RIVM algorithms. This suggests that the average sensitivity weeks of EARS-H and RIVM algorithms for rare diseases were significantly different from that of the other algorithms. It was further noted that the average estimated sensitivity weeks of the RIVM and EARS-H algorithms were higher than that of the other algorithms. For seasonal disease there was a significant difference between the sensitivity weeks of EARS-NH, EARS-H, HPA and RIVM algorithms; and between GAM, EARS-H, HPA and RIVM algorithms. However, there was no significant difference between the average sensitivity weeks of EARS-NH and GAM algorithms. An inspection of the estimated average sensitivity weeks showed that the average sensitivity weeks of GAM and EARS-NH algorithms were higher than that of the other three algorithms.

Further it was noted that there was a significant difference between the average sensitivity weeks of EARS-NH algorithm for common and seasonal diseases, EARS-H for common and rare diseases, and RIVM for common and seasonal diseases and for rare and seasonal diseases.

Table 19, shows the p-values of the Tukey's pair-wise comparison of algorithm disease combination performed on estimated sensitivity run. The table suggests that there was no significant difference between the average sensitivity run of all algorithms for common diseases. There was a significant difference between the average estimated sensitivity run

between EARS-H and HPA algorithms for rare diseases; however no difference was noted between the average sensitivity run of the other algorithms. An inspection of the average estimated sensitivity run showed that the EARS-H algorithm had the highest and the HPA had the lowest sensitivity run for rare diseases. For the seasonal disease, there was a significant difference between the average sensitivity run of EARS-NH, RIVM and HPA algorithms for seasonal disease. Also there was a significant difference between the average sensitivity run of GAM, HPA and RIVM algorithms. An inspection of the estimated average sensitivity run showed that the average sensitivity run of EARS-NH, EARS-H and GAM algorithms were higher than that of the other two algorithms for seasonal disease. Further it was noted that there was a significant difference between the average sensitivity run of RIVM algorithm for the three diseases, and GAM for rare and seasonal diseases, and HPA for common and rare diseases. However, there was no significant difference between the average sensitivity run of the other two algorithms for the three diseases.

Table 20 shows the p-values of the Tukey's pair-wise comparison of algorithm disease combination performed on estimated specificity. The table shows that the average estimated specificity of RIVM algorithm was significantly different from the remaining four algorithms for common disease. An examination of the average specificity for the common disease showed that the RIVM algorithm had lowest specificity; however, all algorithms had a specificity of greater than 88%. The average estimated specificity of the EARS-NH algorithm was significantly different from the GAM, RIVM, and HPA algorithms for rare disease. There was also a significant difference between the average

specificity of GAM, EARS-H, and RIVM algorithms, between EARS-H, HPA, and RIVM algorithms, and HPA and RIVM algorithms for rare diseases. However, it was noted that the specificity of all algorithms was higher than 84% for rare diseases. Similarly, although there was a significant difference between the average specificity of some algorithms for seasonal diseases, the specificity of all algorithms was higher than 87% for the seasonal disease.

Also there was a significant difference between the average estimated specificity of EARS-H algorithm for the three diseases, and RIVM for common and rare diseases. However, there was no significant difference between the average specificity of the other three algorithms for the three diseases.

Table 21 show the p-values of the Tukey's pair-wise comparison of algorithm disease combination performed on estimated PPV. There was a significant difference between the average PPV for EARS-NH, EARS-H, GAM, and HPA algorithms for common diseases. However, there was no significant difference between the RIVM and the remaining four algorithms; and between EARS-NH and RIVM algorithms with respect to average estimated PPV for common disease. An inspection of the average estimated PPV for the common diseases showed that the average estimated PPV of GAM, EARS-H, and HPA algorithms and subsequently of RIVM algorithm (since there was no significant difference between these algorithms) was higher than that of the EARS-NH algorithm. For rare disease, the average PPV of EARS-NH, GAM and HPA algorithms were significantly different. There was also a significant difference between the GAM, EARS-

H and RIVM algorithms, between EARS-H and HPA algorithms, and between HPA and RIVM algorithms for rare diseases. It was noted that the estimated average PPV of GAM and HPA algorithms were higher than that of the other three algorithms for rare diseases. For seasonal disease, the average PPV of the HPA algorithm was significantly different from that of the other four algorithms. There was also a significant difference between the PPV of EARS-H, EARS-NH, and GAM algorithms. An inspection of the estimated PPV for the seasonal diseases showed that the average PPV of HPA algorithms was higher than that of the other four algorithms for seasonal diseases.

The tests also showed that there was a significant difference between the average PPV of EARS-NH algorithm for common and rare diseases, and between the average PPV of EARS-H and RIVM algorithms for all three diseases.

The results from the Tukey's pairwise multiple comparison tests were mostly consistent with the results from the interaction plot analysis. The summary of the results from the simulation study is presented in Table 4. RIVM and GAM algorithms had the best sensitivity weeks for common disease; however, there was not much difference between the five algorithms with respect to sensitivity run and specificity for the common disease. GAM, HPA, RIVM, and EARS-H algorithms had overall better PPV than the other algorithms. Therefore it can be concluded based on the evidence from the simulation study, that any of the five algorithms can be used in detecting outbreaks of common diseases.

For the rare disease, RIVM and EARS-H algorithms had higher sensitivity weeks than the other algorithms. Just as in the case of common disease, the sensitivity run and specificity of all algorithms were not very different. The HPA and GAM algorithms however had higher PPV for rare disease than the other algorithms. Based on the evidence from the simulation study it can therefore be concluded that any of the five algorithms may be used in detecting outbreaks in rare diseases.

For seasonal disease, GAM and EARS-NH algorithms had higher sensitivity weeks; GAM, EARS-NH, and EARS-H algorithms had higher sensitivity run; and HPA algorithm had higher PPV. All algorithms had similar specificities for the seasonal diseases. Therefore it can be concluded based on the evidence from the simulation study that GAM, EARS-NH, and EARS-H algorithms may be more suitable in detecting outbreaks in seasonal diseases.

Evaluation Measures	Common Disease	Rare Disease	Seasonal Disease
Sensitivity Weeks	RIVM, GAM	RIVM, EARS-H	GAM, EARS-NH
Sensitivity Run	All algorithms	All algorithms	GAM, EARS-NH, EARS-H
Specificity	All algorithms	All algorithms	All algorithms
PPV	GAM, HPA, EARS-H, RIVM	HPA, GAM	HPA
All Measures	All algorithms	All algorithms	GAM, EARS-NH, EARS-H

Table 4: Algorithms suitable in detecting outbreaks by type of disease (Simulation Study)

CHAPTER 6. CONCLUSION

As discussed in the previous chapters, five algorithms were compared in this study. These algorithms were applied to empirical data sets that depicted common, rare, seasonal, and secular disease patterns and simulated data sets that depicted common, rare, and seasonal disease patterns. The empirical study and the simulation study showed that the performance of the algorithms in detecting outbreaks, in a timely fashion and with less number of false flags, depended on the type of disease to which they were applied. Some algorithms performed well in detecting common disease, some in detecting rare disease, while others in detecting seasonal disease.

As was mentioned earlier, it should be noted that the algorithms were applied to the data sets without changing the parameter settings in the algorithms. In practice, countries that use these algorithms adapt the algorithms to suit the type of disease to which they are applied. It is very likely that if the algorithms were applied to the same data sets, with a different set of parameters, the results obtained may be different. Hence, caution should be taken when interpreting the results presented in this study.

The empirical study showed that any of the five algorithms may have been chosen for detecting outbreaks in common, rare, seasonal and secular type of diseases. These algorithms had an overall high sensitivity, specificity, and PPV and good times to detection when compared to the other algorithms.

To generate data sets for the simulation study the empirical probabilities of disease counts from the empirical data sets were used. Since the algorithms were applied to a greater number of data sets from common, rare and seasonal diseases in the simulation evaluation, there was a potential for performing statistical tests and examining the distribution of the evaluation measures using CDF plots, box plots, and interaction plots. The CDF plot of the evaluation measures by algorithms for each disease showed that RIVM algorithm followed by EARS-NH and EARS-H algorithms performed well in detecting outbreaks in common diseases. The EARS-H algorithm performed well in detecting outbreaks in rare diseases, while GAM and EARS-NH algorithms performed well in detecting outbreaks in seasonal diseases.

In general EARS-NH and EARS-H algorithms had good times to detection. These algorithms had less variability and smaller times to detection than the other algorithms. However there was not a vast difference in the algorithms in terms of times to detection.

An investigation of the evaluation measures of algorithms by diseases showed that there was an interaction between algorithms and diseases. That is to say, the performance of the algorithms depended on the type of disease to which they were applied. The two-way MANOVA and one-way ANOVA suggested that there was an interaction between algorithms and diseases. The interaction plots showed the same results.

The results from the empirical and simulation study were very similar. Both empirical and simulation study showed any of the five algorithms might be used for detecting

outbreaks common and rare diseases. However, there was a difference in the results from empirical and simulation study for seasonal diseases. The empirical study showed that any of the five algorithms might be used for detecting outbreaks in seasonal diseases, while the simulation study showed that some of the algorithms might be preferred for detecting outbreaks in seasonal diseases.

The conclusions drawn from the simulation study are based on the assumption that the simulated data sets capture the pattern of the occurrence of disease counts in the empirical data sets. The results from the simulation study may be extrapolated provided the simulated data sets capture the true variation in the processes of the disease counts in the empirical data sets.

The MANOVA and ANOVA were performed with only 10 data sets per algorithm-disease combination in this study. Comprehensive research with more replication is needed to further support the results from this study. Based on the evidence from the study it may be concluded that any of the five algorithms may be used to detect outbreaks in common, rare, and seasonal diseases.

Bibliography

- Barbujani, G., Calzolari, E. (1984), "Comparison of Two Statistical Techniques for the Surveillance of Birth Defects through a Monte Carlo Simulation", *Statistics In Medicine*, 3, 239-247.
- Barbujani, G., Ceccherini, I., Russo, A. (1986), "Surveillance of Birth Defects: The Multi-community Sets Technique by Computer Simulation", *European Journal of Epidemiology*, 0392-2990, 52-62.
- Barker, W.H., Feldt, K.S., Feibel, J., et al, (1984), "Assessment of hospital admission surveillance of stroke in a metropolitan community", *American Journal of Chronic Diseases*, 37, 609-15
- Belio-Blasco, C. (2000), "Evaluation of two retrospective active surveillance methods for the detection of nosocomial infection in surgical patients". *Infection Control and Hospital Epidemiology*, 21, 24-27.
- Centers for Disease Control and Prevention, (1988), "Guidelines for evaluating surveillance systems". *Morbidity and Mortality Weekly Report*, 37, S-5, 1-18.
Available at: <http://www.cdc.gov/mmwr/preview/mmwrhtml/00001769.htm>
- Centers for Disease Control and Prevention, (1998), "Revising CDC's guidelines for evaluating surveillance systems". *Morbidity and Mortality Weekly Report*, 47, 1083.
- Centers for Disease Control and Prevention, (2001), "Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines working group". *Morbidity and Mortality Weekly Report*, 50, No RR13, 1.
Available at: <http://www.cdc.gov/mmwr/PDF/RR/RR5013.pdf>
- Chen, R., (1978), "A Surveillance System for Congenital Malformations". *Journal of the American Statistical Association*, 73, 323-327.
- Cleveland, W.S., Devlin, S.J. (1988), "Locally-weighted Regression: An Approach to Regression Analysis by Local Fitting", *Journal of American Statistical Association*, Col. 83, pp 596-610.
- Farrington, C.P., Andrews, N.J., et al., (1996), "A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease", *Journal of the Royal Statistical Society Series A*, 159, 547-563.
- Feller, W. (1969), "Introduction to Probability Theory and its Applications", Vol 1, 3rd Ed, Wiley 1968, page 173, eqn 10.9.
- Frisén, M., (1992), "Evaluation of Methods for Statistical Surveillance", *Statistics in Medicine*, 11, 1489-1502.

Fujii, H., Takahashi, H., et al. (2002), "Evaluation of a Sentinel Surveillance System for Influenza, 1995-2000, Kyoto city, Japan", *Japanese Journal of Infectious Diseases*, 55, 23-26.

German, R.R. (2000), "Sensitivity and Predictive Value Positive measurements for Public Health Surveillance Systems. *Epidemiology*, 11(6), 720-727.

Hastie, T., Tibshirani, R. (1990), "Generalized Additive Models.

Hashimoto, S., Murakami, Y., Taniguchi, K., Nagai, M. (2000), "Detection of Epidemics in their Early Stage through Infectious Disease Surveillance. *International Journal of Epidemiology*, 29, 905-910.

Available at: <http://ije.oupjournals.org/cgi/reprint/29/5/905.pdf>

Center for Surveillance Coordination, Health Canada, (2002), "Environmental Scan of Existing Surveillance Systems with Automated Alerts".

Hutwagner, L.C., Maloney, E.K., et al. (1997), "Using Laboratory-Based Surveillance Data for Prevention: An algorithm for Detecting Salmonella Outbreaks" *Emerging Infectious Diseases*, Vol 3 No 3.

Available at: <ftp://ftp.cdc.gov/pub/EID/vol3no3/adobe/hutwagner.pdf>

Klaucke, D.N., Buehler, J.W., Thacker, S.B., Parrish, R.G., Trowbridge, F.L., Berkelman, R.L. (1988), "Guidelines for Evaluating Surveillance Systems". *Morbidity and Mortality Weekly Report*, 37(S-5), 1-18.

Available at: <http://www.cdc.gov/mmwr/preview/mmwrhtml/00001769.htm>

Liverpool School of Tropical Medicine and Ministry of Child Health and Welfare, Government of Zimbabwe. (1997), "Protocol for the Evaluation of Epidemiological Surveillance Systems." World Health Organization.

Available at: <http://www.who.int/emc-documents/surveillance/docs/whoemcdis972E.pdf>

Pediaditis, J. "Evaluation of Surveillance Systems".

Available at: <http://www.mzcp-zoonoses.gr/chania/Lectures/Evaluation/Evaluation.PPT>

Quénel, P., Dab, W., Hannoun, C., Cohen, J.M. (1994), "Sensitivity, Specificity and Predictive values of Health-Service Based Indicators for the Surveillance of Influenza A Epidemics", *International Journal of Epidemiology*, 23 (4), 849-855.

Robotin, M. (2002), "Evaluation of the Australian CJD Surveillance System".

Communicable Diseases Intelligence, Vol. 25, No. 2.

Available at: <http://www.cda.gov.au/pubs/cdi/2002/cdi2602/cdi2602n.htm#pagetop>

Serfling, R.E. (1963), "Methods for Current Statistical Analysis of Excess Pneumonia-Influenza Death", *Public Health Reports*, 78(6), 494-506.

Stroup, D.F., Williamson, G.D., Herndon, J.L. (1989), “Detection of Aberration in the Occurrence of Notifiable Diseases Surveillance Data”, *Statistics in Medicine*, 8, 323-329.

Stroup, D.F., Wharton, M., Kafadar, K., Dean, A.G. (1993), “Evaluation of a Method for Detecting Aberrations in Public Health Surveillance Data”, *American Journal of Epidemiology*, 137 (3), 373-380.

Takahashi, H., Fujii, H., et al., (2001), “Evaluation of the Japanese school health surveillance system for influenza”, *Japanese Journal of Infectious Diseases*, 54, 27-30.

Tsui, F.C, Wagner, M.M, et al., (2002), “Value of ICD-9–Coded Chief Complaints for Detection of Epidemics”, *Journal of the American Medical Informatics Association*, 9, S41-S47.

Available at <http://dev.health.pitt.edu/ahrq/restricted/papers/AMIA01%20tsui.pdf>

Wagner, M.M., Tsui, F.C., et al. (2001), “The Emerging Science of Very Early Detection of Disease Outbreaks”, *Journal of Public Health Management and Practice*, 7(6), 51–59.

Weatherall, J.A.C., Haskey, J.C. (1976), “Surveillance of Malformations”, *British Medical Bulletin*, 32, 39-44.

White, R. (2002), Personal Communication.

Widdowson, M.A, Bosman, A., et al. (2003), “Automated, Laboratory-based System Using the Internet for Disease Outbreak Detection, the Netherlands”, *Emerging Infectious Diseases*, Vol. 9, No 9.

Available at: <http://www.cdc.gov/ncidod/EID/vol9no9/pdfs/02-0450.pdf>

World Health Organization. (2001), “Protocol for the Assessment of National Communicable Disease Surveillance and Response Systems”. Available at: <http://www.who.int/emc-documents/surveillance/docs/whocdscsr20012.pdf>

Appendix A: Tables

Algorithm	Sensitivity (Std dev)		Specificity (Std dev)	Positive Predictive Value	Times to Detection (ttd) (Weeks)										Mean ttd	Median ttd
	Weeks	Run	Weeks	Weeks	Outbrk1 1997/ wk37	Outbrk2 1998/ wk16	Outbrk3 1998/ wk32	Outbrk4 1999/ wk7	Outbrk5 1999/ wk29	Outbrk6 2000/ wk10	Outbrk7 2000/ wk32	Outbrk8 2002/ wk5	Outbrk9 2002/ wk15			
EARS-NH	14.4 (0.037)	77.7 (0.139)	98.2 (0.009)	76.5 (0.008)	0	0	ND	0	9	0	0	0	ND	1.29	0	
GAM	20 (0.042)	44.4 (0.166)	100 (0)	100 (0)	5	0	ND	ND	ND	ND	ND	3	0	2	1.5	
EARS-H	16.6 (0.039)	55.5 (0.166)	99.1 (0.006)	88.2 (0.006)	5	3	-1	0	1	ND	ND	ND	ND	1.6	1	
HPA	1.1 (0.011)	11.1 (0.105)	100 (0)	100 (0)	5	ND	ND	ND	ND	ND	ND	ND	ND	5	5	
RIVM	31.1 (0.049)	55.5 (0.166)	94.1 (0.016)	68.3 (0.005)	0	0	1	6	ND	ND	ND	1	ND	1.6	1	

Table 5: Estimates of measures for Hepatitis A data set (Empirical Study)

Algorithm	Sensitivity (Std dev)		Specificity (Std dev)	Positive Predictive Value (Std dev)	Times to Detection (ttd) (Weeks)												Mean ttd	Median ttd
	Weeks	Run	Weeks	Weeks	Outbrk1 2000/ wk49	Outbrk2 2001/ wk16	Outbrk3 2001/ wk28	Outbrk4 2001/ wk30	Outbrk5 2001/ wk45	Outbrk6 2001/ wk50	Outbrk7 2002/ wk17	Outbrk8 2002/ wk22	Outbrk9 2002/ wk31	Outbrk10 2002/ wk48	Outbrk11 2002/ wk50			
EARS-NH	81.8 (0.116)	81.8 (0.116)	100 (0)	100 (0)	0	0	0	0	0	ND	0	ND	0	0	0	0	0	
GAM	36.4 (0.145)	36.4 (0.145)	100 (0)	100 (0)	ND	0	0	0	0	ND	ND	ND	ND	ND	ND	0	0	
EARS-H	72.7 (0.134)	72.7 (0.134)	100 (0)	100 (0)	0	0	0	0	0	ND	0	ND	0	0	ND	0	0	
HPA	0 (0)	0 (0)	100 (0)	0 (0)	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	-	-	
RIVM	18.1 (0.116)	18.1 (0.116)	97.2 (0.014)	33.3 (0.044)	ND	ND	0	0	ND	ND	ND	ND	ND	ND	ND	0	0	

Table 6: Estimates of measures for Hanta Virus data set (Empirical Study)

Algorithm	Sensitivity (Std dev)		Specificity (Std dev)	Positive Predictive Value (Std dev)	Times to Detection (ttd) (Weeks)			
	Weeks	Run			Outbrk1 2000/ wk26	Outbrk2 2000/ wk40	Mean ttd	Median ttd
EARS-NH	55.5 (0.166)	100 (0)	95.3 (0.032)	55.5 (0.166)	0	0	0	0
GAM	66.6 (0.157)	100 (0)	93 (0.038)	66.6 (0.157)	2	0	1	1
EARS-H	88.8 (0.104)	100 (0)	86 (0.053)	88.8 (0.104)	0	0	0	0
HPA	44.4 (0.166)	100 (0)	100 (0)	44.4 (0.166)	2	0	1	1
RIVM	66.6 (0.157)	50 (0.354)	67.4 (0.071)	66.6 (0.157)	ND	-9	-9	-9

Table 7: Estimates of measures for Salmonella Dublin data set (Empirical Study)

Algorithm	Sensitivity (Std dev)		Specificity (Std dev)	Positive Predictive Value	Times to Detection (ttd) (Weeks)										Mean ttd	Median ttd
	Weeks	Run	Weeks	Weeks	Outbrk1 1998/ wk3	Outbrk2 1999/ wk1	Outbrk3 1999/ wk18	Outbrk4 1999/ wk44	Outbrk5 2000/ wk7	Outbrk6 2000/ wk11	Outbrk7 2000/ wk18	Outbrk8 2000/ wk47	Outbrk9 2002/ wk5			
EARS-NH	39.5 (0.053)	66.6 (0.157)	93.1 (0.019)	73.9 (0.005)	3	0	ND	-3	ND	0	ND	-2	-2	-0.67	-1	
GAM	56.9 (0.053)	66.6 (0.157)	95.9 (0.015)	87.5 (0.004)	1	1	0	1	ND	ND	ND	1	1	0.83	1	
EARS-H	62.8 (0.052)	66.6 (0.157)	91.9 (0.021)	79.4 (0.004)	0	1	0	-3	ND	ND	ND	5	4	1.17	0.5	
HPA	38.4 (0.052)	55.5 (0.166)	99.4 (0.058)	97.1 (0.002)	4	1	1	2	ND	ND	ND	ND	5	2.6	2	
RIVM	52.3 (0.054)	55.5 (0.166)	85.5 (0.0027)	64.3 (0.004)	2	2	2	-3	ND	ND	ND	ND	5	1.6	2	

Table 8: Estimates of measures for Influenza A & B data set (Empirical Study)

Algorithm	Sensitivity (Std dev)		Specificity (Std dev)	Positive Predictive Value (Std dev)	Times to Detection (ttd) (Weeks)										
	Weeks	Run	Weeks	Weeks	Outbrk1 1996/ wk2	Outbrk2 1996/ wk7	Outbrk3 1996/ wk19	Outbrk4 1997/ wk37	Outbrk5 1998/ wk2	Outbrk6 1998/ wk7	Outbrk7 2000/ wk12	Outbrk8 2001/ wk5	Outbrk9 2002/ wk45	Mean ttd	Median ttd
EARS-NH	29.4 (0.111)	33.3 (0.157)	86.5 (0.018)	9.6 (0.008)	ND	ND	0	0	ND	ND	0	ND	ND	0	0
GAM	23.5 (0.102)	22.2 (0.139)	98.3 (0.007)	40 (0.026)	0	ND	ND	0	ND	ND	ND	ND	ND	0	0
EARS-H	41.2 (0.119)	66.6 (0.157)	96.3 (0.018)	35 (0.02)	0	0	0	0	2	0	ND	ND	ND	0.333	0
HPA	17.6 (0.092)	22.2 (0.139)	98.5 (0.064)	37.5 (0.03)	0	ND	ND	ND	ND	ND	0	ND	ND	0	0
RIVM	41.2 (0.119)	55.5 (0.166)	91.1 (0.015)	18.4 (0.011)	0	ND	0	ND	2	0	0	ND	ND	0.4	0

Table 9: Estimates of measures for Salmonella Enteritidis data set (Empirical Study)

Algorithm	Times to Detection (ttd) (Weeks)												
	Sensitivity (Std Dev)		Specificity (Std Dev)	Positive Predictive Value (Std Dev)									
	Weeks	Run	Weeks	Weeks	Outbrk1 2000/ wk24	Outbrk2 2000/ wk40	Outbrk3 2001/ wk30	Outbrk4 2002/ wk8	Outbrk5 2002/ wk19	Outbrk6 2002/ wk32	Outbrk7 2002/ wk49	Mean ttd	Median ttd
EARS-NH	18.2 (0.067)	57.1 (0.187)	98.4 (0.011)	75 (0.017)	0	ND	2	ND	ND	0	0	0.5	0
GAM	63.6 (0.084)	57.1 (0.187)	93.5 (0.022)	72.4 (0.008)	ND	ND	0	0	0	ND	0	0	0
EARS-H	48.5 (0.087)	85.7 (0.132)	95.1 (0.019)	72.7 (0.009)	0	ND	0	0	0	0	0	0	0
HPA	30.3 (0.08)	57.1 (0.187)	97.6 (0.014)	76.9 (0.012)	ND	ND	0	0	0	ND	0	0	0
RIVM	93.9 (0.042)	71.4 (0.171)	71.5 (0.041)	47 (0.004)	ND	ND	0	0	0	-2	0	-0.4	0

Table 10: Estimates of measures for Gonorrhea data set (Empirical Study)

Algorithm	Sensitivity (Std dev)		Specificity (Std dev)	Positive Predictive Value	Times to Detection (ttd) (Weeks)								
	Weeks	Run			Outbrk1 1999/ wk35	Outbrk2 1999/ wk52	Outbrk3 2000/ wk36	Outbrk4 2000/ wk50	Outbrk5 2001/ wk16	Outbrk6 2002/ wk1	Outbrk7 2002/ wk14	Mean ttd	Median ttd
	EARS-NH	22.9 (0.061)	85.7 (0.132)	96.1 (0.016)	Weeks	0	0	0	6	ND	0	0	1
GAM	43.8 (0.072)	57.1 (0.187)	96.7 (0.014)	80.8 (0.007)	0	0	0	7	ND	ND	ND	1.75	0
EARS-H	62.5 (0.07)	100 (0)	85.5 (0.029)	57.7 (0.006)	0	0	3	0	0	0	0	0	0
HPA	18.8 (0.056)	42.9 (0.187)	99.3 (0.007)	90 (0.008)	ND	3	3	6	ND	ND	ND	4	3
RIVM	66.6 (0.068)	57.1 (0.187)	80.9 (0.032)	52.5 (0.005)	0	0	0	0	ND	ND	ND	0	0

Table 11: Estimates of measures for Meningococcal data set (Empirical Study)

Measures	Statistic	EARS-NH	GAM	EARS-H	HPA	RIVM
Sensitivity Weeks	Min	5.263158	6.542056	14.10256	2.777778	37.68116
	P25	16.4557	29.85075	20	24.17582	58.42697
	Mean	20.28884	43.98796	25.53451	32.36934	67.96376
	Median	20.4416	45.05495	22.78481	31.64313	68.73303
	P75	23.80952	52.32558	28.37838	39.24051	77.77778
	Max	34.54545	83.33333	64.91228	62.7451	94.66667
	Std Dev	5.669235	18.08357	9.399194	12.21953	13.84301
Sensitivity Run	Min	33.33333	37.5	42.85714	33.33333	50
	P25	75	54.54545	71.42857	62.5	83.33333
	Mean	81.61068	70.8164	78.85591	77.05433	90.21162
	Median	83.33333	71.42857	80	80	100
	P75	100	85.71429	85.71429	100	100
	Max	100	100	100	100	100
	Std Dev	17.04881	19.60176	15.29936	18.76621	12.90268
Specificity	Min	90.21739	79.08497	93.30144	93.75	79.61165
	P25	93.75	97.35099	98.42105	97.15909	85.37736
	Mean	94.83192	97.43609	98.66388	97.86601	88.71517
	Median	94.78145	98.07692	98.82005	97.84847	88.79121
	P75	95.93023	99.05213	99.4152	98.87006	92.52874
	Max	98.36066	100	100	100	95.93909
	Std Dev	1.752076	3.308237	1.154609	1.400119	4.149291
PPV	Min	18.75	17.94872	56.25	6.666667	45.45455
	P25	52.63158	85.71429	85.71429	81.81818	65
	Mean	61.30094	88.02709	88.9372	84.99113	71.11787
	Median	62.03346	90	90.90909	87.02665	72.58813
	P75	71.42857	93.47826	93.33333	91.66667	80.32787
	Max	93.75	100	100	100	92.22222
	Std Dev	14.89301	13.27468	8.930549	13.9248	11.15239
SD Sensitivity Weeks	Min	0.029576	0.023904	0.03377	0.027389	0.025946
	P25	0.039615	0.049411	0.042578	0.045145	0.046594
	Mean	0.046168	0.053679	0.049279	0.052143	0.050899
	Median	0.043216	0.053595	0.047568	0.050732	0.050513
	P75	0.052432	0.060747	0.054757	0.061568	0.054415
	Max	0.066902	0.072106	0.072169	0.071603	0.070001
	Std Dev	0.009249	0.008752	0.008743	0.010283	0.008311
SD Sensitivity Run	Min	0	0	0	0	0
	P25	0	0.13226	0.13226	0	0
	Mean	0.108602	0.141328	0.125573	0.124742	0.067345
	Median	0.152145	0.165635	0.152145	0.152145	0
	P75	0.153093	0.187044	0.171163	0.176777	0.13226
	Max	0.272166	0.25	0.219089	0.272166	0.219089
	Std Dev	0.07939	0.07235	0.07093	0.080269	0.077224
SD Specificity	Min	0.010927	0	0	0	0.014063
	P25	0.014974	0.006671	0.005831	0.007945	0.020177
	Mean	0.016145	0.009979	0.007605	0.009814	0.022975
	Median	0.016054	0.010528	0.008186	0.010569	0.023309
	P75	0.017863	0.012504	0.009065	0.012329	0.025733
	Max	0.021901	0.03288	0.017293	0.016173	0.030464
	Std Dev	0.00253	0.005981	0.00371	0.003962	0.003723
SD PPV	Min	0.000344	0.000122	0.000146	0.000179	0.000203
	P25	0.000654	0.000261	0.000311	0.000348	0.000252
	Mean	0.000872	0.000342	0.000521	0.000479	0.000331
	Median	0.000796	0.000313	0.00042	0.000411	0.000292
	P75	0.001062	0.000404	0.000605	0.000505	0.000376

Measures	Statistic	EARS-NH	GAM	EARS-H	HPA	RIVM
	Max	0.00211	0.000864	0.002015	0.001706	0.000723
	Std Dev	0.000325	0.000136	0.000331	0.000258	0.000113

Table 12: Summary measures for Common disease (Simulation study)

Variable	Number Detected	Total Number of Outbreaks	Min	25th Pctl	Mean	Median	75th Pctl	Std Dev	Max
EARS-NH TTD	268	332	-2	0	1.66	1	2	2.37	15
GAM TTD	183	332	-25	1	2.62	2	3	3.83	19
EARS-H TTD	230	332	0	1	3.60	3	5	3.41	17
HPA TTD	224	332	-1	1	2.91	2	4	3.48	22
RIVM TTD	230	332	-34	0	0.13	1	2	5.82	15

Table 13: Times to detection for Common disease (Simulation study)

Measures	Statistic	EARS-NH	GAM	EARS-H	HPA	RIVM
Sensitivity Weeks	Min	14.516129	19.04762	29.72973	15.09434	30.64516
	P25	27.659574	32.6087	55.81395	19.67213	50.81967
	Mean	34.837938	44.21393	62.03627	26.18746	61.30445
	Median	35.294118	44.56265	62.18202	24.99	62.5
	P75	39.655172	53.44828	69.56522	30.55556	73.13433
	Max	57.894737	75	91.66667	42.85714	95.83333
	Std Dev	9.420249	12.22685	10.69512	7.332215	14.32922
Sensitivity Run	Min	33.333333	38.46154	55.55556	30.76923	33.33333
	P25	50	55.55556	70	46.15385	57.14286
	Mean	64.857295	65.49954	78.92121	59.18944	67.67166
	Median	63.636364	66.66667	77.77778	55.55556	66.66667
	P75	75	75	88.88889	70	81.81818
	Max	100	100	100	100	100
	Std Dev	15.96557	12.9273	12.84406	17.25373	16.70754
Specificity	Min	89.63964	93.92523	85.42714	97.60766	76.19048
	P25	91.071429	96.65272	89.25234	99.04762	81.28079
	Mean	92.629001	97.61082	90.5329	99.34797	84.37034
	Median	92.559498	97.74775	90.65148	99.51567	85.06412
	P75	94.063927	98.54369	91.74757	99.55157	86.87783
	Max	96.534653	99.54545	94.30052	100	91.58879
	Std Dev	1.8137197	1.216521	2.055207	0.546696	3.53365
PPV	Min	28.571429	58.33333	28.20513	63.63636	25.64103
	P25	39.393939	71.42857	52.83019	85.71429	38.88889
	Mean	50.616753	79.67214	58.81682	89.69456	46.34712

	Median	51.576355	81.01673	58.94673	91.28788	47.40686
	P75	57.5	88.88889	65.38462	95.2381	53.125
	Max	79.411765	96.2963	81.96721	100	70
	Std Dev	12.909197	10.40873	9.663543	8.742411	10.98528
SD Sensitivity Weeks	Min	0.0447375	0.054414	0.053162	0.047401	0.040789
	P25	0.062038	0.064974	0.064333	0.056189	0.062348
	Mean	0.0690535	0.071021	0.069426	0.064106	0.068225
	Median	0.0666287	0.070031	0.068434	0.062909	0.0665
	P75	0.0730297	0.075232	0.072916	0.067843	0.07303
	Max	0.1059712	0.105971	0.092296	0.10799	0.108985
	Std Dev	0.0112096	0.00907	0.007591	0.012027	0.010859
SD Sensitivity Run	Min	0	0	0	0	0
	P25	0.1280077	0.13226	0.104757	0.128812	0.116927
	Mean	0.1363639	0.140991	0.110083	0.141777	0.129699
	Median	0.1385799	0.143328	0.116927	0.144338	0.136083
	P75	0.1571348	0.157135	0.134932	0.157135	0.153093
	Max	0.1924501	0.176777	0.171163	0.19245	0.187044
	Std Dev	0.0359039	0.026112	0.044843	0.030561	0.038487
SD Specificity	Min	0.0128688	0.004535	0.016553	0	0.018973
	P25	0.0163189	0.008347	0.018926	0.004474	0.023301
	Mean	0.0177622	0.010087	0.019963	0.004718	0.024707
	Median	0.0178014	0.010069	0.019969	0.004832	0.024751
	P75	0.0194059	0.012329	0.021172	0.006702	0.026182
	Max	0.0207607	0.016329	0.025012	0.01057	0.029391
	Std Dev	0.0020107	0.002681	0.001957	0.002906	0.002211
SD PPV	Min	0.0004913	0.000199	0.000274	0.000217	0.000323
	P25	0.0008271	0.000385	0.000507	0.000459	0.000473
	Mean	0.0011465	0.00074	0.000665	0.000793	0.000659
	Median	0.001095	0.000616	0.000602	0.000617	0.000583
	P75	0.0013397	0.000894	0.000748	0.001068	0.00077
	Max	0.0027095	0.002879	0.001393	0.002468	0.002083
	Std Dev	0.0004405	0.000512	0.000249	0.000492	0.000285

Table 14: Summary measures for Rare disease (Simulation study)

Variable	Number Detected	Number of Outbreaks	Min	25th Pctl	Mean	Median	75th Pctl	Max	Std Dev
EARS-NH TTD	334	527	-2.00	0.00	1.17	1.00	2.00	11.00	1.51
GAM TTD	308	527	-1.00	1.00	1.81	2.00	3.00	9.00	1.48
EARS-H TTD	351	527	-1.00	0.00	1.01	1.00	1.00	5.00	1.16
HPA TTD	293	527	0.00	1.00	2.28	2.00	3.00	7.00	1.39
RIVM TTD	309	527	-13.00	0.00	0.78	1.00	2.00	8.00	2.39

Table 15: Times to detection summary for Rare Disease (Simulation Study)

Measures	Statistic	EARS-NH	GAM	EARS-H	HPA	RIVM
Sensitivity Weeks	Min	38.157895	28.37838	14.89362	1.785714	12.04819
	P25	44.642857	43.42105	26.78571	13.72549	31.81818
	Mean	49.223636	51.18239	35.37202	21.85732	42.33411
	Median	49.193339	51.55075	35.94225	21.71429	42.9972
	P75	52.857143	58.33333	44.23077	29.16667	51.6129
	Max	62.5	75.60976	60.97561	43.90244	73.17073
	Std Dev	6.064864	11.78875	11.11213	10.31789	14.22526
Sensitivity Run	Min	66.666667	66.66667	42.85714	20	20
	P25	80	80	80	50	60
	Mean	90.961905	90.22857	86.25714	63.14286	74.00952
	Median	100	100	83.33333	60	80
	P75	100	100	100	80	100
	Max	100	100	100	100	100
	Std Dev	10.784186	11.28688	15.40813	18.6888	21.04166
Specificity	Min	88.235294	88.29268	81.9209	95.09804	77.71739
	P25	91.219512	92.69406	85.85859	99.5122	85.78431
	Mean	92.165726	94.46344	87.48376	99.64488	87.32401
	Median	92.270531	94.84413	87.71491	100	87.50259
	P75	93.362832	96.46465	89.3401	100	89.85507
	Max	95.431472	100	92.95154	100	93.8326
	Std Dev	1.7321686	2.610651	2.50225	0.749747	3.550212
PPV	Min	44.736842	53.19149	20	9.090909	30.50847
	P25	59.090909	67.92453	39.02439	92.30769	40.90909
	Mean	63.851592	72.89971	43.68534	94.26683	47.72105
	Median	62.767898	72.72727	44.77041	100	47.29532
	P75	70.37037	79.16667	50	100	54.38596
	Max	76.744186	100	62.7451	100	66.66667
	Std Dev	7.3031256	10.24187	9.640208	13.55531	9.144259
SD Sensitivity Weeks	Min	0.0534248	0.049764	0.045235	0.017697	0.035731
	P25	0.0614326	0.059679	0.056103	0.044237	0.057647
	Mean	0.0664531	0.065025	0.062152	0.053039	0.063465
	Median	0.0662828	0.063813	0.061157	0.053915	0.063749
	P75	0.0700005	0.068079	0.067141	0.062826	0.069196
	Max	0.0893811	0.089756	0.088627	0.083959	0.089756
	Std Dev	0.007924	0.007976	0.009529	0.014833	0.010478
SD Sensitivity Run	Min	0	0	0	0	0
	P25	0	0	0	0.178885	0
	Mean	0.0753458	0.07973	0.097248	0.188341	0.143616
	Median	0	0	0.152145	0.198287	0.178885
	P75	0.1788854	0.178885	0.178885	0.219089	0.204124
	Max	0.1924501	0.19245	0.219089	0.219089	0.219089
	Std Dev	0.0863567	0.087751	0.091616	0.051647	0.087474

SD Specificity	Min	0.0146956	0	0.016989	0	0.015967
	P25	0.0174537	0.013525	0.022094	0	0.021108
	Mean	0.0187997	0.01547	0.023172	0.002507	0.023148
	Median	0.018696	0.015719	0.022928	0	0.023506
	P75	0.0198144	0.018019	0.02437	0.004866	0.024828
	Max	0.0232673	0.022455	0.028927	0.015117	0.030678
	Std Dev	0.0020174	0.004089	0.002249	0.003341	0.002785
SD PPV	Min	0.0004151	0.000254	0.000406	0.000208	0.000448
	P25	0.0004987	0.000413	0.000677	0.000369	0.00061
	Mean	0.0006412	0.000571	0.000846	0.000635	0.000736
	Median	0.0006117	0.000492	0.000785	0.000461	0.000702
	P75	0.0006915	0.000606	0.000912	0.000819	0.000824
	Max	0.0013191	0.001727	0.00223	0.001486	0.001203
	Std Dev	0.0001965	0.000272	0.000293	0.000385	0.000172

Table 16: Summary measures for Seasonal Disease (Simulation Study)

Variable	Number Detected	Total Number of Outbreaks	Min	25th Pctl	Mean	Median	75th Pctl	Max	Std Dev
EARS-NH TTD	236	261	-3.00	0.00	0.38	0.00	0.00	8.00	1.59
GAM TTD	234	261	-3.00	0.00	2.14	1.00	4.00	14.00	2.53
EARS-H TTD	223	261	-4.00	0.00	2.94	2.00	5.00	14.00	3.30
HPA TTD	164	261	0.00	3.00	5.65	5.00	8.00	20.00	3.82
RIVM TTD	191	261	-5.00	2.00	4.46	4.00	7.00	20.00	3.63

Table 17: Times to detection summary for Seasonal Disease (Simulation Study)

i/j	EARS-NH-Common	EARS-NH-Rare	EARS-NH-Seasonal	GAM-Common	GAM-Rare	GAM-Seasonal	EARS-H-Common	EARS-H-Rare	EARS-H-Seasonal	HPA-Common	HPA-Rare	HPA-Seasonal	RIVM-Common	RIVM-Rare	RIVM-Seasonal
EARS-NH-Common															
EARS-NH-Rare	0.0858														
EARS-NH-Seasonal	<0001	0.0199													
GAM-Common	<0001	0.6638	0.9657												
GAM-Rare	0.0002	0.925	0.7722	1											
GAM-Seasonal	<0001	0.0429	1	0.9928	0.8972										
EARS-H-Common	0.9997	0.6003	<0001	0.0014	0.0085	<0001									
EARS-H-Rare	<0001	<0001	0.5126	0.0098	0.0016	0.3444	<0001								
EARS-H-Seasonal	0.4158	1		0.2016	0.4958	0.0038	0.962	<0001							
HPA-Common	0.141	1	0.0104	0.5243	0.8454	0.0236	0.7348	<0001	1						
HPA-Rare	1	0.3828	<0001	0.0004	0.0028	<0001	1	<0001	0.8601	0.5175					
HPA-Seasonal	1	0.1637	<0001	<0001	0.0005	<0001	1	<0001	0.5987	0.2511	1				
RIVM-Common	<0001	<0001	0.0483	0.0002	<0001	0.0226	<0001	0.9992	<0001	<0001	<0001	<0001			
RIVM-Rare	<0001	<0001					<0001	1	<0001	<0001	<0001	<0001	0.9999		
RIVM-Seasonal	0.2371	1	0.3713	0.0048	0.0007	0.2301	<0001	<0001					<0001	<0001	
				0.3661	0.7095	0.0111	0.8639	<0001	1	1	0.6817	0.3866	<0001	<0001	

Table 18: Tests of hypothesis for interaction effects for Sensitivity Weeks (Simulation Study)

i/j	EARS-NH- Common	EARS-NH- Rare	EARS-NH- Seasonal	GAM- Common	GAM- Rare	GAM- Seasonal	EARS-H- Common	EARS-H- Rare	EARS-H- Seasonal	HPA- Common	HPA- Rare	HPA- Seasonal	RIVM- Common	RIVM- Rare	RIVM- Seasonal
EARS-NH- Common															
EARS-NH- Rare	0.9459														
EARS-NH- Seasonal	0.9812	0.1351													
GAM- Common	1	0.9999	0.6632												
GAM- Rare	0.8497	1	0.0686	0.9981											
GAM- Seasonal	0.5155	0.0073	0.9997	0.1128	0.0028										
EARS-H- Common	0.9804	0.1327	1	0.658	0.0672	0.9998									
EARS-H- Rare	1	0.5545	1	0.9791	0.374	0.931	1								
EARS-H- Seasonal	1	0.7172	0.9997	0.9958	0.5341	0.8367	0.9997	1							
HPA- Common	1	0.4807	1	0.9629	0.3104	0.9573	1	1	1						
HPA- Rare	0.1403	0.983		0.5791	0.9975	<0.0001	0.0012	0.018	0.0372	0.0128					
HPA- Seasonal	0.563	1	0.0178	0.9591	1	0.0005	0.0173	0.1485	0.2486	0.1153	1				
RIVM- Common	0.844	0.0374	1	0.3379	0.0165	1	1	0.9969	0.9832	0.9988	0.0002	0.0035			
RIVM- Rare	0.7777	1	0.0472	0.9943	1	0.0017	0.0462	0.294	0.4409	0.239	0.9993	1	0.0106		
RIVM- Seasonal	0.7629	1	0.0439	0.9932	1	0.0016	0.043	0.2804	0.4242	0.2271	0.9995	1	0.0098	1	

Table 19: Tests of hypothesis for interaction effects for Sensitivity Run (Simulation Study)

i/j	EARS-NH-Common	EARS-NH-Rare	EARS-NH-Seasonal	GAM-Common	GAM-Rare	GAM-Seasonal	EARS-H-Common	EARS-H-Rare	EARS-H-Seasonal	HPA-Common	HPA-Rare	HPA-Seasonal	RIVM-Common	RIVM-Rare	RIVM-Seasonal
EARS-NH-Common															
EARS-NH-Rare	0.4956														
EARS-NH-Seasonal	0.6807	1													
GAM-Common	0.7369	0.0012	0.0032												
GAM-Rare	0.848	0.0025	0.0065	1											
GAM-Seasonal	1	0.8833	0.9622	0.3164	0.4414										
EARS-H-Common	0.1391	<0001	<0001	0.9996	0.9975	0.0248									
EARS-H-Rare	0.0057	0.9355	0.8303	<0001	<0001	0.043	<0001								
EARS-H-Seasonal	<0001	0.0103	0.0041	<0001	<0001	<0001	<0001	0.6161							
HPA-Common	0.8495	0.0026	0.0066	1	1	0.4436	0.9974	<0001	<0001						
HPA-Rare	0.0229	<0001	<0001	0.9525	0.8915	0.0027	1	<0001	<0001	0.8902					
HPA-Seasonal	0.0045	<0001	<0001	0.7421	0.6095	0.0004	0.9987	<0001	<0001	0.6072	1				
RIVM-Common	<0001	0.0309	0.0134	<0001	<0001	<0001	<0001	0.8315	1	<0001	<0001	<0001			
RIVM-Rare	<0001	<0001	<0001	<0001	<0001	<0001	<0001	<0001	0.1095	<0001	<0001	<0001	0.0432		
RIVM-Seasonal	<0001	0.1099	0.0544	<0001	<0001	0.0002	<0001	0.9757	1	<0001	<0001	<0001	1	0.0103	

Table 20: Tests of hypothesis for interaction effects for Specificity (Simulation Study)

i/j	EARS-NH-Common	EARS-NH-Rare	EARS-NH-Seasonal	GAM-Common	GAM-Rare	GAM-Seasonal	EARS-H-Common	EARS-H-Rare	EARS-H-Seasonal	HPA-Common	HPA-Rare	HPA-Seasonal	RIVM-Common	RIVM-Rare	RIVM-Seasonal
EARS-NH-Common															
EARS-NH-Rare	0.0073														
EARS-NH-Seasonal	1	0.0238													
GAM-Common	0.0004	<0001	0.0001												
GAM-Rare	0.27	<0001	0.121	0.8009											
GAM-Seasonal	0.4843	<0001	0.2598	0.5763	1										
EARS-H-Common	0.001	<0001	0.0002		0.9	0.7208									
EARS-H-Rare	0.9999	0.1127	1	<0001	0.0261	0.0705	<0001								
EARS-H-Seasonal	0.0008	1		<0001	<0001	<0001	<0001	0.0207							
HPA-Common	0.0364	<0001	0.0116	0.9966	1	0.9986	0.9995	0.0016	<0001						
HPA-Rare	<0001	<0001	<0001	1	0.5482	0.3207	1	<0001	<0001	0.9601					
HPA-Seasonal	<0001	<0001	<0001	0.8788	0.0156	0.0049	0.7703	<0001	<0001	0.1499	0.979				
RIVM-Common	0.9452	<0001		0.1232	0.9983	1	0.2004	0.4253	<0001	0.841	0.0443	0.0002			
RIVM-Rare	0.2678	0.9948	0.495	<0001	<0001	<0001	<0001	0.85	0.8764	<0001	<0001	<0001	0.0018		
RIVM-Seasonal	0.0274	1	0.0775	<0001	<0001	<0001	<0001	0.2785	0.9997	<0001	<0001	<0001	<0001	0.9999	

Table 21: Tests of hypothesis for interaction effects for PPV (Simulation Study)

Appendix B: Figures

Figure 2: Flags of algorithms on Hepatitis A data set (Empirical Study)

Figure 3: Flags of algorithms on Hanta Virus data set (Empirical Study)

Figure 4: Flags of algorithms on Salmonella Dublin data set (Empirical Study)

Figure 5: Flags of algorithms on Influenza A & B data set (Empirical Evaluation)

Figure 6: Flags of algorithms on Salmonella Enteritidis data set (Empirical Study)

Figure 7: Flags of algorithms on Gonorrhea data set (Empirical Study)

Figure 8: Flags of algorithms on Meningococcal data set (Empirical Study)

Figure 9 Example of simulation (Common disease)

Figure 10 Example of Simulation (Rare disease)

Figure 11 Example of Simulation (Seasonal disease)

Figure 12: Box plot of Times to detection for Common disease (Simulation Study)

Figure 13: Box Plot of Times to detection for Rare Disease (Simulation Study)

Figure 14: Box Plot of Times to detection for Seasonal Disease (Simulation Study)

Figure 15: CDF Plot of Sensitivity run for Common Disease (Simulation Study)

Figure 16: CDF Plot of Specificity for Common Disease (Simulation Study)

Figure 17: CDF Plot of PPV for Common Disease (Simulation Study)

Figure 18: CDF Plot of Sensitivity Weeks for Rare Disease (Simulation Study)

Figure 19: CDF plot for Sensitivity run for Rare disease (Simulation Study)

Figure 20: CDF plot of Specificity for Rare disease (Simulation Study)

Figure 21: CDF plot of PPV for Rare disease (Simulation Study)

Figure 22: CDF plot of Sensitivity weeks for Seasonal disease (Simulation Study)

Figure 23: CDF plot of Sensitivity run for Seasonal disease (Simulation Study)

Figure 24: CDF plot of Specificity for Seasonal disease (Simulation Study)

Figure 25: CDF plot of PPV for Seasonal disease (Simulation Study)

Figure 26: Average Sensitivity Weeks for Algorithms by Disease (Simulation Study)

Figure 27: Average Sensitivity Weeks by Disease for each Algorithm (Simulation Study)

Figure 28: Average Sensitivity Run for Algorithms by Disease (Simulation Study)

Figure 29: Average Sensitivity Run by Disease for each Algorithm (Simulation Study)

Figure 30: Average Specificity for Algorithms by Disease (Simulation Study)

Figure 31: Average Specificity by Disease for each Algorithm (Simulation Study)

Figure 32: Average PPV for Algorithms by Disease (Simulation Study)

Figure 33: Average PPV by Disease for each Algorithm (Simulation Study)

Figure 34: Normal Probability Plot for Sensitivity Weeks (Simulation Study)

Figure 35: Normal Probability Plot for Sensitivity Run (Simulation Study)

Figure 36: Normal Probability Plot for Specificity (Simulation Study)

Figure 37: Normal Probability Plot for PPV (Simulation Study)

Figure 38: Box Plot of Residuals for Sensitivity Weeks (Simulation Study)

Figure 39: Box Plot of Residuals for Sensitivity Run (Simulation Study)

Figure 40: Box Plot of Residuals for Specificity (Simulation Study)

Figure 41: Box Plot of residuals for PPV (Simulation Study)

Figure 42: Box Plot of the residuals for the measures (Simulation Study)

Appendix C: R Codes for Simulation

R Codes for Simulating Data Sets for Common Diseases

```

#Reading Hepatitis A baseline data (daily series)
hepa.data <- read.table(file = "c:/thesis/hepa_data_97_02.txt", header = T)

# Reading Hepatitis A gold standard data (daily series)
hepa.gs <- read.table(file = "c:/thesis/hepa_gs_97_02.txt", header = T)

#Reading Hepatitis A data set (weekly series)
hepa.gs.wkly = read.table( file = "c:/data for alerts/bcdata/hepa_gs_97_02.txt", header =
T)

# Specifying n.obs, the number of observations to be generated in the simulation
# and n.years, the number of years of data available in the Hepatitis A data set
n.obs = 3660; n.years = 6

# Initializing x1, the baseline counts to be generated
x1 = rep(0,n.obs)

# Assigning x to the Hepatitis A baseline data (daily series)
x <- hepa.data$count

## Initializing the number of outbreaks to be generated n.outb
n.outb = 0

## Determining the number of outbreaks in the Hepatitis A data set (weekly series)
for (i in 1 :length(hepa.gs.wkly$GOLDSTD)) {
# the gold standard in the data set was indicated by 2
  if (hepa.gs.wkly$GOLDSTD[i] ==2) {
    n.outb = n.outb + 1
  }
}

## Loop to calculate the length of each outbreak in the Hepatitis A data

## Initializing the outbreak size, outb.size for each outbreak outb.size = rep(0,n.outb)

i = 1
j = 1

while (i <= length(hepa.gs.wkly$GOLDSTD))
{
  if (hepa.gs.wkly$GOLDSTD[i] == 2)
  {
    outb.size[j] = 1 ## if goldstd = 2 then initialize outb.size to be 1
  }
}

```

```

        while (hepa.gs.wkly$GOLDSTD[i+1] == 1)
        {
            outb.size[j] = outb.size[j] + 1
            i = i + 1
        }
        j = j + 1
    }
    i = i + 1
}

```

```

# Since the outbreak size outb.size was obtained from weekly series of Hepatitis A data
# while daily data is generated, therefore the outb.size should be multiplied by 7
outb.size = outb.size * 7

```

```

# Determining the number of outbreaks to be generated in a 10 years series
n.outb = ceiling(n.outb * 10 / n.years)

```

```

# Generating the first section of data using the empirical probability distribution of the
# baseline Hepatitis A data
i = 1
n = n.obs / 3

```

```

p1 = table(x[ i : n ])/ length(x[ i : n ])
x1[ i : n ] <- sample( c(0 : 5, 7 ) , length(i : n), replace = T, prob = p1 )

```

```

# Generating the second section of data
i = (n.obs / 3) + 1
n = (2 / 3) * n.obs

```

```

p2 = table(x[ i : n ])/ length(x[ i : n ])
x1[ i : n ] <- sample( c(0 : 2 ) , length(i : n), replace = T, prob = p2 )

```

```

# Generating the third section of data
i = ( (2 / 3) * n.obs ) + 1
n = n.obs

```

```

p3 = table(x[ i : n ])/ length(x[ i : n ])
x1[ i : n ] <- sample( 0 : 4 , length(i : n), replace = T, prob = p3 )

```

```

# Assigning the outbreak series of Hepatitis A data set to t.outbreaks
t.outbreaks <- hepa.gs$count

```

```

# Initializing the outbreak counts to be generated
outbreaks <- rep(0, n.obs)

```



```

## Randomly generating points where outbreaks have to be inserted

ind <- rep( 0, n.obs )

i = 1
while ( i < n.obs)
{
    ind[i] <- 30 + i
    i = 30 + i
}

ind <- ind[ ind > 0]

# Selecting the number of outbreaks randomly
n.outbk <- sample(((n.outb-2):(n.outb+2)),1)
indic <- sample(ind, n.outbk)

## change the value of the outbreaks at these points to some value (here 100)
outbreaks[indic] <- 100

## Initializing the gold standard
goldstd <- rep(0, 3660)

## Generating the first section of the outbreaks

i = 1
n = n.obs/3

p1 = table(t.outbreaks[ i : n ])/ length(t.outbreaks[ i : n ])
while ( i < n)
{
    if (outbreaks[i] > 0)
    {
        k = sample(min(outb.size):max(outb.size), 1)
        outbreaks[ i: (i+k) ] <- sample(c( 0:6, 8 ), k+1, replace = T, prob = p1 )
        x1[ i : (i + k) ] <- outbreaks[ i: (i+k) ]
        goldstd[ i : (i + k) ] <- 1
        i = i + k + 1
    }
    else
    {
        i = i + 1
    }
}

```

```
## Generating the second section of the outbreaks
```

```
i = (n.obs / 3) + 1
n = (2 / 3) * n.obs

p2 = table(t.outbreaks[ i : n ])/ length(t.outbreaks[ i : n ])
while ( i < n)
{
  if (outbreaks[i] > 0)
  {
    k = sample(min(outb.size):max(outb.size), 1)
    outbreaks[ i: (i+k) ] <- sample(0:6, k+1, replace = T, prob = p2 )
    x1[ i : (i + k) ] <- outbreaks[ i: (i+k) ]
    goldstd[ i : (i + k) ] <- 1
    i = i + k + 1
  }
  else
  {
    i = i + 1
  }
}
```

```
## Generating the third section of the outbreaks
```

```
i = ( (2 / 3) * n.obs ) + 1
n = n.obs

p3 = table(t.outbreaks[ i : n ])/ length(t.outbreaks[ i : n ])
while ( i < n)
{
  if (outbreaks[i] > 0)
  {
    k = sample(min(outb.size):max(outb.size), 1)
    outbreaks[ i: (i+k) ] <- sample(0:6, k+1, replace = T, prob = p3 )
    x1[ i : (i + k) ] <- outbreaks[ i: (i+k) ]
    goldstd[ i : (i + k) ] <- 1
    i = i + k + 1
  }
  else
  {
    i = i + 1
  }
}
```

```
## Finalizing simulated series by adding the baseline and outbreak series
```

```
x1 <- x1[1:n.obs]
```

```
outbreaks <- outbreaks[1:n.obs]
goldstd <- goldstd[1:n.obs]
```

R Codes for Simulating Data Sets for Rare Diseases

```
# Reading the Salmonella Dublin data (daily series)
dublin.data = read.table( file = "c:/thesis/dublin_count_2000.txt", header = T)

# Reading the Salmonella Dublin gold standard data (daily series)
dublin.gs <- read.table(file = "c:/thesis/dublin_gs_2000.txt", header = T)

# Reading the Salmonella Dublin gold standard data (weekly series)
dublin.gs.wkly = read.table( file = "c:/thesis/salm_dublin_gs_2000.txt", header = T)

# Specifying n.obs, the number of observations to be generated in the simulation
# and n.years, the number of years of data available in the Hepatitis A data set
n.obs = 3660; n.years = 1

# Initializing x1, the baseline counts to be generated
x1 = rep(0,n.obs)

# Assigning x to the Hepatitis A baseline data (daily series)
x <- dublin.data$Count

## Initializing the number of outbreaks to be generated n.outb
n.outb = 0

for (i in 1 :length(dublin.gs.wkly$GOLDSTD)) {
  if (dublin.gs.wkly$GOLDSTD[i] == 2) {
    n.outb = n.outb + 1
  }
}

## Loop to calculate the length of each outbreak in the Hepatitis A data

## Initializing the outbreak size, outb.size for each outbreak
outb.size = rep(0,n.outb)

i = 1
j = 1

while (i <= length(dublin.gs.wkly$GOLDSTD))
{
```

```

    if (dublin.gs.wkly$GOLDSTD[i] == 2)
    {
        outb.size[j] = 1
        while (dublin.gs.wkly$GOLDSTD[i+1] == 1)
        {
            outb.size[j] = outb.size[j] + 1
            i = i + 1
        }
        j = j + 1
    }
    i = i + 1
}

# Since the outbreak size outb.size was obtained from weekly series of Hepatitis A data
# while daily data is generated, therefore the outb.size should be multiplied by 7
outb.size = outb.size * 7

# Determining the number of outbreaks to be generated in a 10 years series
n.outb = ceiling(n.outb * 10 / n.years)

# Generating the baseline counts
p = table(x[ i : n ])/ length(x[ i : n ])
x1 <- sample( 0:3 , n.obs, replace = T, prob = p)

# Assigning the outbreak series of Hepatitis A data set to t.outbreaks
t.outbreaks <- dublin.gs$Count

# Initializing the outbreak counts to be generated
outbreaks <- rep(0, n.obs)

## Randomly generating points where outbreaks have to be inserted
n.outbk <- sample(((n.outb-2):(n.outb+2)),1)
indic <- sample(1 : n.obs, n.outbk)

## change the value of the outbreaks at these points to some value (here 100)
outbreaks[indic] <- 100

## Initializing the gold standard
goldstd <- rep(0, n.obs)

# Generating the outbreak counts
i = 1

```

```

n = n.obs

p = table(x[ i : n ])/ length(x[ i : n ])
while ( i < n)
{
  if (outbreaks[i] > 0)
  {
    k = sample(min(outb.size):max(outb.size), 1)
    outbreaks[ i: (i+k) ] <- sample(0:2, k+1, replace = T, prob = p )
    x1[ i : (i + k) ] <- outbreaks[ i: (i+k) ]
    goldstd[ i : (i + k) ] <- 1
    i = i + k + 1
  }
  else
  {
    i = i + 1
  }
}

## Finalizing simulated series by adding the baseline and outbreak series
outbreaks <- outbreaks[1:n.obs]
x1 <- x1[1:n.obs]
goldstd <- goldstd[1:n.obs]

```

R Codes for Simulating Data Sets for Seasonal Diseases

```

#Reading Influenza A & B baseline data (daily series)
flu.ab = read.table( file = "c:/thesis/flu_daily_cases.txt", header = T)
count.d = flu.ab$COUNT

# Reading Influenza A & B gold standard data (daily series)
flu.gs <- read.table( file = "c:/thesis/flu_daily_gs.txt", header = T)

#Reading Influenza A & B data set (weekly series)
flu.gs.wkly = read.table( file = "c:/data for alerts/alta_data/flu_ab/flu_gs_98_02.txt",
header = T)

# Specifying n.obs, the number of observations to be generated in the simulation
# and n.years, the number of years of data available in the Hepatitis A data set
n.obs = 3660; x = flu.ab$count; n.years = 5

## Determining the number of outbreaks in the Hepatitis A data set (weekly series)
n.outb = 0

for (i in 1 :length(flu.gs.wkly$goldstd)) {

```

```

        if (flu.gs.wkly$goldstd[i] == 2) {
            n.outb = n.outb + 1
        }
    }

## Loop to calculate the length of each outbreak in the Hepatitis A data
outb.size = rep(0,n.outb)

i = 1
j = 1

while (i <= length(flu.gs.wkly$goldstd))
{
    if (flu.gs.wkly$goldstd[i] == 2)
    {
        outb.size[j] = 1      ## if goldstd = 2 then initialize outb.size to be 1
        while (flu.gs.wkly$goldstd[i+1] == 1)
        {
            outb.size[j] = outb.size[j] + 1 ## increment till the end of outbreak
            is reached
            i = i + 1
        }
        j = j + 1
    }
    i = i + 1
}

# Since the outbreak size outb.size was obtained from weekly series of Hepatitis A data
# while daily data is generated, therefore the outb.size should be multiplied by 7
outb.size = outb.size * 7

# Determining the number of outbreaks to be generated in a 10 years series
n.outb = ceiling(n.outb * 10 / n.years)

## Creating random uniform variables
u = runif(n.obs, 0, max(p1))

# Initializing x1, the baseline counts to be generated
x1 = rep(0,n.obs)

# Determining the number of observations and probability of occurrence of each
# observation in the Influenza A & B data sets
n = length(x)
p1 = table(x)/n

```

```

# Generating the baseline counts
x1[ 1 : n.obs ] <- sample( c(0 : 2 ) ,replace = T, length(1 : n.obs), prob = p1 )

## Randomly selecting the number of outbreaks to be inserted in the simulated data set
n.outbk <- sample(((n.outb-2):(n.outb+2)),1)

# Assigning the outbreak series of Hepatitis A data set to t.outbreaks
t.outbreaks <- flu.gs$count

# Initializing the outbreak counts to be generated
outbreaks <- rep(0, n.obs)

## Randomly generating points where outbreaks have to be inserted in 10 years of data
ind1 <- sample(1:35, 1)
ind2 <- sample(730:765 ,1)
ind3 <- sample(366:401 ,1)
ind4 <- sample(1096:1131 ,1)
ind5 <- sample(1461:1496 ,1)
ind6 <- sample(1826:1861 ,1)
ind7 <- sample(2191:2226 ,1)
ind8 <- sample(2556:2591 ,1)
ind9 <- sample(2921:2956 ,1)
ind10 <- sample(3286:3321 ,1)

indic <- c(ind1, ind2, ind3, ind4, ind5, ind6, ind7, ind8, ind9, ind10)

## change the value of the outbreaks at these points to some value (here 100)
outbreaks[indic] <- 100

## Determining the first and third quartile of the outbreak counts in the Influenza A & B
## data set
q1.tout <- quantile(t.outbreaks, 0.25)
q3.tout <- quantile(t.outbreaks, 0.75)

## Initializing the gold standard
goldstd <- rep(0,n.obs)

# Determining the probabilities of occurrence of the outbreak counts
p.q3 <- table(t.outbreaks[t.outbreaks < q3.tout])/length(t.outbreaks[t.outbreaks < q3.tout])
p.q1 <- table(t.outbreaks[t.outbreaks > q1.tout])/length(t.outbreaks[t.outbreaks > q1.tout])

# Generating the outbreak counts
i = 1
while (i < n.obs)
{

```

```

    if (outbreaks[i] > 0 )
    {
        k = sample(min(outb.size):max(outb.size),1)
        t.outbreaks[t.outbreaks < q3.tout]
        outbreaks[ i : (i + k/3)] <- sample(0:6, length(i : (i + k/3)), replace = T,
prob = p.q3)
        t.outbreaks[t.outbreaks > q1.tout]
        outbreaks[( (i + k/ 3) + 1 ) : (( (i + 2* k/ 3)) + 1) ] <- sample(c(2 : 29,
32:33, 37, 39 ), length(( (i + k/ 3) + 1 ) : (( (i + 2* k/ 3)) + 1)), replace = T, prob = p.q1)
        t.outbreaks[ t.outbreaks < q3.tout ]
        outbreaks[((( (i + 2* k / 3)) + 2) : (i + k) ] <- sample(0:6, length((( (i + 2* k
/ 3)) + 2) : (i + k)), replace = T, prob = p.q3)
        x1[ i : (i + k) ] <- outbreaks[ i : (i+k) ]
        goldstd[ i : (i + k) ] <- 1
        i = i + k + 1
    }
    else
    {
        i = i + 1
    }
}

## Finalizing simulated series by adding the baseline and outbreak series
outbreaks <- outbreaks[1:n.obs]
goldstd <- goldstd[1:n.obs]
x1 <- x1[1:n.obs]

```


**Appendix D: Exact Binomial Confidence Interval for
Empirical Study**

Lower CI	Upper CI	Disease	Algorithm	Measure
0.013	0.987	Salmonella Dublin	RIVM	Sensitivity Run
0.158	1.000	Salmonella Dublin	HPA	Sensitivity Run
0.158	1.000	Salmonella Dublin	EARS-H	Sensitivity Run
0.158	1.000	Salmonella Dublin	GAM	Sensitivity Run
0.158	1.000	Salmonella Dublin	EARS-NH	Sensitivity Run
0.299	0.925	Salmonella Dublin	RIVM	Sensitivity Weeks
0.137	0.788	Salmonella Dublin	HPA	Sensitivity Weeks
0.518	0.997	Salmonella Dublin	EARS-H	Sensitivity Weeks
0.299	0.925	Salmonella Dublin	GAM	Sensitivity Weeks
0.212	0.863	Salmonella Dublin	EARS-NH	Sensitivity Weeks
0.515	0.809	Salmonella Dublin	RIVM	Specificity
0.918	1.000	Salmonella Dublin	HPA	Specificity
0.721	0.947	Salmonella Dublin	EARS-H	Specificity
0.809	0.985	Salmonella Dublin	GAM	Specificity
0.842	0.994	Salmonella Dublin	EARS-NH	Specificity
0.119	0.543	Salmonella Dublin	RIVM	PPV
0.398	1.000	Salmonella Dublin	HPA	PPV
0.289	0.823	Salmonella Dublin	EARS-H	PPV
0.299	0.925	Salmonella Dublin	GAM	PPV
0.290	0.963	Salmonella Dublin	EARS-NH	PPV
0.212	0.863	Salmonella Enteritidis	RIVM	Sensitivity Run
0.028	0.600	Salmonella Enteritidis	HPA	Sensitivity Run
0.299	0.925	Salmonella Enteritidis	EARS-H	Sensitivity Run
0.028	0.600	Salmonella Enteritidis	GAM	Sensitivity Run
0.075	0.701	Salmonella Enteritidis	EARS-NH	Sensitivity Run
0.184	0.671	Salmonella Enteritidis	RIVM	Sensitivity Weeks
0.038	0.434	Salmonella Enteritidis	HPA	Sensitivity Weeks
0.184	0.671	Salmonella Enteritidis	EARS-H	Sensitivity Weeks
0.068	0.499	Salmonella Enteritidis	GAM	Sensitivity Weeks
0.103	0.560	Salmonella Enteritidis	EARS-NH	Sensitivity Weeks
0.876	0.938	Salmonella Enteritidis	RIVM	Specificity
0.967	0.995	Salmonella Enteritidis	HPA	Specificity
0.937	0.980	Salmonella Enteritidis	EARS-H	Specificity
0.963	0.994	Salmonella Enteritidis	GAM	Specificity
0.824	0.899	Salmonella Enteritidis	EARS-NH	Specificity
0.077	0.343	Salmonella Enteritidis	RIVM	PPV
0.085	0.755	Salmonella Enteritidis	HPA	PPV
0.154	0.592	Salmonella Enteritidis	EARS-H	PPV
0.122	0.738	Salmonella Enteritidis	GAM	PPV
0.032	0.210	Salmonella Enteritidis	EARS-NH	PPV
0.212	0.863	Influenza A & B	RIVM	Sensitivity Run
0.212	0.863	Influenza A & B	HPA	Sensitivity Run
0.299	0.925	Influenza A & B	EARS-H	Sensitivity Run

Lower CI	Upper CI	Disease	Algorithm	Measure
0.299	0.925	Influenza A & B	GAM	Sensitivity Run
0.299	0.925	Influenza A & B	EARS-NH	Sensitivity Run
0.413	0.632	Influenza A & B	RIVM	Sensitivity Weeks
0.281	0.495	Influenza A & B	HPA	Sensitivity Weeks
0.517	0.730	Influenza A & B	EARS-H	Sensitivity Weeks
0.458	0.676	Influenza A & B	GAM	Sensitivity Weeks
0.292	0.507	Influenza A & B	EARS-NH	Sensitivity Weeks
0.794	0.904	Influenza A & B	RIVM	Specificity
0.968	1.000	Influenza A & B	HPA	Specificity
0.868	0.955	Influenza A & B	EARS-H	Specificity
0.918	0.984	Influenza A & B	GAM	Specificity
0.882	0.964	Influenza A & B	EARS-NH	Specificity
0.519	0.754	Influenza A & B	RIVM	PPV
0.847	0.999	Influenza A & B	HPA	PPV
0.679	0.883	Influenza A & B	EARS-H	PPV
0.759	0.948	Influenza A & B	GAM	PPV
0.589	0.857	Influenza A & B	EARS-NH	PPV
0.290	0.963	Gonorrhea	RIVM	Sensitivity Run
0.184	0.901	Gonorrhea	HPA	Sensitivity Run
0.421	0.996	Gonorrhea	EARS-H	Sensitivity Run
0.184	0.901	Gonorrhea	GAM	Sensitivity Run
0.184	0.901	Gonorrhea	EARS-NH	Sensitivity Run
0.798	0.993	Gonorrhea	RIVM	Sensitivity Weeks
0.156	0.487	Gonorrhea	HPA	Sensitivity Weeks
0.308	0.665	Gonorrhea	EARS-H	Sensitivity Weeks
0.451	0.796	Gonorrhea	GAM	Sensitivity Weeks
0.070	0.355	Gonorrhea	EARS-NH	Sensitivity Weeks
0.627	0.793	Gonorrhea	RIVM	Specificity
0.930	0.995	Gonorrhea	HPA	Specificity
0.897	0.982	Gonorrhea	EARS-H	Specificity
0.876	0.972	Gonorrhea	GAM	Specificity
0.942	0.998	Gonorrhea	EARS-NH	Specificity
0.346	0.597	Gonorrhea	RIVM	PPV
0.462	0.950	Gonorrhea	HPA	PPV
0.498	0.893	Gonorrhea	EARS-H	PPV
0.528	0.873	Gonorrhea	GAM	PPV
0.349	0.968	Gonorrhea	EARS-NH	PPV
0.023	0.518	Hanta Virus	RIVM	Sensitivity Run
0.000	0.285	Hanta Virus	HPA	Sensitivity Run
0.390	0.940	Hanta Virus	EARS-H	Sensitivity Run
0.109	0.692	Hanta Virus	GAM	Sensitivity Run
0.482	0.977	Hanta Virus	EARS-NH	Sensitivity Run
0.023	0.518	Hanta Virus	RIVM	Sensitivity Weeks

Lower CI	Upper CI	Disease	Algorithm	Measure
0.000	0.285	Hanta Virus	HPA	Sensitivity Weeks
0.390	0.940	Hanta Virus	EARS-H	Sensitivity Weeks
0.109	0.692	Hanta Virus	GAM	Sensitivity Weeks
0.482	0.977	Hanta Virus	EARS-NH	Sensitivity Weeks
0.930	0.992	Hanta Virus	RIVM	Specificity
0.975	1.000	Hanta Virus	HPA	Specificity
0.975	1.000	Hanta Virus	EARS-H	Specificity
0.975	1.000	Hanta Virus	GAM	Specificity
0.975	1.000	Hanta Virus	EARS-NH	Specificity
0.043	0.777	Hanta Virus	RIVM	PPV
0.000	1.000	Hanta Virus	HPA	PPV
0.631	1.000	Hanta Virus	EARS-H	PPV
0.398	1.000	Hanta Virus	GAM	PPV
0.664	1.000	Hanta Virus	EARS-NH	PPV
0.212	0.863	Hepatitis A	RIVM	Sensitivity Run
0.003	0.482	Hepatitis A	HPA	Sensitivity Run
0.212	0.863	Hepatitis A	EARS-H	Sensitivity Run
0.137	0.788	Hepatitis A	GAM	Sensitivity Run
0.400	0.972	Hepatitis A	EARS-NH	Sensitivity Run
0.218	0.417	Hepatitis A	RIVM	Sensitivity Weeks
0.000	0.060	Hepatitis A	HPA	Sensitivity Weeks
0.096	0.260	Hepatitis A	EARS-H	Sensitivity Weeks
0.123	0.298	Hepatitis A	GAM	Sensitivity Weeks
0.079	0.234	Hepatitis A	EARS-NH	Sensitivity Weeks
0.902	0.968	Hepatitis A	RIVM	Specificity
0.984	1.000	Hepatitis A	HPA	Specificity
0.968	0.999	Hepatitis A	EARS-H	Specificity
0.984	1.000	Hepatitis A	GAM	Specificity
0.955	0.995	Hepatitis A	EARS-NH	Specificity
0.519	0.819	Hepatitis A	RIVM	PPV
0.025	1.000	Hepatitis A	HPA	PPV
0.636	0.985	Hepatitis A	EARS-H	PPV
0.815	1.000	Hepatitis A	GAM	PPV
0.501	0.932	Hepatitis A	EARS-NH	PPV
0.184	0.901	Meningococcal Infections	RIVM	Sensitivity Run
0.099	0.816	Meningococcal Infections	HPA	Sensitivity Run
0.590	1.000	Meningococcal Infections	EARS-H	Sensitivity Run
0.184	0.901	Meningococcal Infections	GAM	Sensitivity Run
0.421	0.996	Meningococcal Infections	EARS-NH	Sensitivity Run
0.516	0.796	Meningococcal Infections	RIVM	Sensitivity Weeks
0.089	0.326	Meningococcal Infections	HPA	Sensitivity Weeks
0.474	0.760	Meningococcal Infections	EARS-H	Sensitivity Weeks
0.295	0.588	Meningococcal Infections	GAM	Sensitivity Weeks

Lower CI	Upper CI	Disease	Algorithm	Measure
0.120	0.373	Meningococcal Infections	EARS-NH	Sensitivity Weeks
0.738	0.868	Meningococcal Infections	RIVM	Specificity
0.964	1.000	Meningococcal Infections	HPA	Specificity
0.789	0.907	Meningococcal Infections	EARS-H	Specificity
0.925	0.989	Meningococcal Infections	GAM	Specificity
0.916	0.985	Meningococcal Infections	EARS-NH	Specificity
0.393	0.654	Meningococcal Infections	RIVM	PPV
0.555	0.997	Meningococcal Infections	HPA	PPV
0.432	0.713	Meningococcal Infections	EARS-H	PPV
0.606	0.934	Meningococcal Infections	GAM	PPV
0.383	0.858	Meningococcal Infections	EARS-NH	PPV

Table 22: Exact Binomial Confidence Intervals (Empirical Study)

Exact Binomial Confidence Intervals

Hepatitis A

It was noticed that the confidence intervals of the sensitivity weeks of all the algorithms except HPA, had an overlap for Hepatitis A data set. As was mentioned earlier the sensitivity weeks of HPA algorithm was lowest for Hepatitis A data set. There was however an overlap of confidence intervals of sensitivity run, specificity, and PPV for all algorithms for this data set. It may therefore be concluded the performance of algorithms in detecting outbreaks in Hepatitis A data set may not be different.

Hanta Virus

There was an overlap in the confidence intervals of sensitivity weeks for RIVM, GAM, EARS-H, and EARS-NH algorithms; and for RIVM, HPA, and GAM algorithms. The same pattern was noticed for sensitivity run. The confidence intervals of specificity and PPV for all algorithms overlapped for Hanta Virus data set. Therefore, it may be

concluded that there is no difference in the performance of all algorithms, except for HPA algorithm, for Hanta Virus data set

Salmonella Dublin

For Salmonella Dublin data set the confidence interval for all evaluation measures overlapped for all algorithms. It may be concluded that there is no difference in the performance of the algorithms for Salmonella Dublin data set. Hence, any of the five algorithms may be preferred for detecting outbreaks in Salmonella Dublin data set.

Salmonella Enteritidis and Influenza A & B

For Salmonella Enteritidis and Influenza A & B data sets the confidence interval for all evaluation measures overlapped for all algorithms. It may be concluded that there is no difference in the performance of the algorithms for Salmonella Enteritidis and Influenza A & B data sets. Hence, any of the five algorithms may be preferred for detecting outbreaks in Salmonella Enteritidis and Influenza A & B data sets.

Gonorrhea

For Gonorrhea the confidence interval of GAM, EARS-H, and HPA algorithms; and EARS-NH, EARS-H, and HPA algorithms overlapped for sensitivity weeks. Earlier, it was noted that the sensitivity weeks of RIVM was highest for Gonorrhea data set. The confidence intervals for the sensitivity run, specificity, and PPV overlapped for all algorithms. Therefore, it may be concluded that all algorithms perform similarly in detecting outbreaks in Gonorrhea data set.

Meningococcal Infections

There was an overlap in the sensitivity weeks of GAM, EARS-NH, and HPA algorithms, and RIVM, EARS-H, and GAM algorithms for Meningococcal data set. The confidence

intervals of sensitivity run, specificity, and PPV overlapped for all algorithms for the Meningococcal data set. Therefore, it may be concluded that all algorithms perform similarly in detecting outbreaks in Meningococcal data set.

It appears from the estimated evaluation measures and exact binomial confidence intervals that any of the five algorithms may be used in detecting outbreaks in common, rare, seasonal, and secular type of diseases.