Stockholms universitet

Changing epidemiology of Invasive Pneumococcal Disease in the Stockholm area due to the introduction of a 7-valent conjugate vaccine in 2007

Ilias Galanis

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2011:5
Matematisk statistik
Juni 2011

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2011:5**
http://www.math.su.se

Changing epidemiology of Invasive Pneumococcal Disease in the Stockholm area due to the introduction of a 7-valent conjugate vaccine in 2007

Ilias Galanis*

June 2011

Abstract

The use of the heptavalent conjugate vaccine (PCV-7) against invasive pneumococcal disease (IPD) which initiated in 2007 was bound to bring changes in the incidence of IPD in the Stockholm area. A vaccine like this, including 7 of the 90 known serotypes that cause the disease, has been known to lead to significant decrease in the incidence of those seven serotypes, but in some cases this result is leveled off by the increase in the incidence of some serotypes not included in the vaccine, a phenomenon known as serotype replacement. An analytical description of the serotype behavior in each age group, a meta-analysis and a Poisson model were used in this study in order to identify signs of this replacement and to estimate the vaccine efficacy. The rate of IPD decreased significantly for the vaccinated group (0-2 years), with relative risk (RR): 0.54 and p-value ($p<0.01$), while herd immunity effect was also apparent. Males also proved to be more prone to the disease than females, especially for the younger age group (RR): 1.81 ($p<0.01$). Serotypes 3, 7F, 19A, 22F and 38 not included in the vaccine, have increased their incidence significantly in the post-vaccine period, but the extent of serotype replacement is not as clear as one would expect. The reasons behind this are the narrow 3-year post-vaccine period, the initiation of a new 13-valent vaccine in 2010 which included serotypes 3, 7F and 19A and the peculiar distribution serotype 1 presents which acts as an outlier in our results.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: gillan13@yahoo.com. Supervisor: Åke Svensson.

Acknowledgments

There are many people that have contributed to the implementation of this study which i feel the need to thank. First of all, i want to express my gratitude to my supervisor Åke Svensson for his advice and guidance, but most of all for the fact that he agreed to join me on my quest for a thesis at SMI. Respectively, i want to thank Birgitta Henriques-Normark, Jessica Darenberg and Pontus Nauclér from SMI, for their help in the formulation of the subject and their assistance during these past months. Special thanks go to Jan-Olov Persson and Rolf Sundberg for their valuable advices and last, but not least, i want to thank my family and friends for their support and encouragement.

Contents

List of Figures

List of Tables

❖ 1

Introduction

1.1 About the disease

Streptococcus Pneumoniae is a dangerous family of bacteria colonizing the nasopharyngeal flora and is responsible for a wide range of illnesses such as otitis, pneumonia, meningitis and bacteremia [1] [2]. Even though someone can be a carrier of the bacteria without being infected by them, in some cases people can develop from a mild otitis to an invasive disease. The infection is characterized as invasive if the bacteria are isolated from a normally sterile site, like blood or cerebrospinal fluid [2] [3]. People most likely to be infected are very young children, the elderly, members of indigenous populations (similar to the American natives or the Aboriginals) and individuals with underlying conditions such as HIV, alcoholism or diabetes mellitus [2] [4]. Although it has been more than a century since the U. S. Army physician George Sternberg isolated the bacteria from a rabbit, they are still a very important reason of morbidity and mortality worldwide, causing an estimated 1.6 million deaths annually, most of which occur in the developing countries [5] [6]. In Sweden the overall incidence of invasive pneumococcal disease (IPD) is 15 cases in 100.000 people per year [7].

Until today over 90 different serotypes of *Streptococcus Pneumoniae* have been found, which are being identified by the differences in the immunochemistry of their capsules. These serotypes vary a lot in their prevalence, virulence and antibiotic resistance and it is thought that the chemical structure of their capsule is a significant factor of this different behavior [8]. In February 2000, a conjugate pneumococcal vaccine was introduced in the United States covering for the seven more prevalent serotypes causing invasive disease at the time. The vaccine was licensed for use in children younger than two years old and to those up to five years who had some high-risk conditions [4]. Being vaccinated provided protection against both disease and nasopharyngeal carriage of the bacteria, so a vaccinated person was not only protecting him/herself, but at the same time was helping to lower the transmission rates among the population - hence protecting the unvaccinated. This is described by

the term *herd immunity* and previous studies have proven that conjugate vaccines provide us with such an effect [3] [9]. In Sweden the conjugate vaccine was first introduced in Stockholm in July 2007 and gradually spread in all the regions of the country until the beginning of 2009.

After the introduction of the 7-valent conjugate vaccine (PCV7) in the United States in 2000, there was a remarkable decrease in the prevalence of the 7 vaccine serotypes and at the same time a significant decrease in the incidence of invasive pneumococcal disease, both for the vaccinated and unvaccinated people [10] [11]. Similar results were concluded from studies in other countries using the vaccine, e.g. Canada or Portugal [12] [13]. Later studies, though, proved that while the incidence rates of IPD caused by the vaccine serotypes were decreasing, at the same time the rates of the serotypes not included in the vaccine (hereafter called non-vaccine serotypes) were increasing significantly [1]. At some cases, after an initial decrease in IPD rates, the effect of the vaccine remained stable, since the decrease of the vaccine serotype incidence was leveled off by the increase of that of the non-vaccine ones [14]. This phenomenon is called *serotype replacement* and is a result of the serotype specificity of the conjugate vaccine.

1.2 Serotype replacement

Prior to the implementation of the conjugate vaccine in 2000, there had been several concerns as to whether the expected reduction in the carriage of the vaccine serotypes will create an ecological niche to be covered finally by the non-vaccine serotypes. In 1997, Marc Lipsitch designed a mathematical model which actually verified these predictions [9]. By using his transmission dynamic model he managed to evaluate the effect a conjugate vaccine, specifically "aiming" only at certain serotypes, would have in the prevalence of both vaccine and non-vaccine serotypes.

In the scenario where only two serotypes were present at the population and the vaccine was covering just for one of them, lets call it serotype 1, he estimated that the increase in the prevalence of the other one (serotype 2) would always be less than the decrease in the carriage of serotype 1. To make it simpler, if the prevalence of the two serotypes before vaccination was 10% and 20% for serotype 1 and 2 respectively, then the prevalence of serotype 2 after vaccination would never exceed 30%. This means, that the number of people carrying either of the two serotypes after the vaccination will be always less compared to the pre-vaccine period [3] [9].

Unfortunately, this protective effect does not necessarily apply when we have three or more different serotypes present among the population. In that case, we might see a non-vaccine serotype becoming so prevalent, that this will overcome the decrease we will gain from the vaccine type(s). Some of the factors that influence the

size of the increase are the basic reproductive number[1] of the serotypes, the degree of competition between them and the vaccine coverage. Under certain conditions, hence, vaccination could lead to a decrease of the vaccine serotypes, but at the same time to an increase in the total serotype prevalence [3] [9].

At this point, we should make clear the two ways serotype replacement can take place. One form of replacement is when two or more serotypes are competing to colonize a host and by vaccinating against one of them, we result in making the other more prevalent. The other form happens when a novel serotype, that was absent from the population because it could not compete with the vaccine serotype(s) previously, finds the opportunity to take over, develop and spread among the population. The major problem in this latter case is that we can not predict the virulence of this new serotype and how invasive it can get [3].

Serotype replacement does not need to be something solely harmful. On the contrary, as Lipsitch also commented in his paper [9], the competition between the serotypes to colonize a host is actually in favor of the work of the vaccine to diminish the vaccine serotypes. It would be ideal if we could include all the harmful serotypes in a vaccine and gain from this competition to make sure that the least invasive end up to be prevalent. There are three obstacles, though, that make it more complicated than this. First of all, there are some clinical limitations regarding the number of serotypes that can be included in a vaccine and second, as mentioned above, there is a possibility that some new serotypes might appear, whose invasiveness we ignore. Last, but maybe most important, these species of colonizing bacteria tend to be highly transformable and so interfering to their prevalence with a conjugate vaccine could lead to a change in the serotype-virulence relationship as we know it [3] [9]. Marc Lipsitch had concluded that the use of a conjugate vaccine against bacteria like *Streptococcus Pneumoniae* with multiple serotypes will provide us with very good results at start, but in order to remain effective we should always monitor the prevalence for possible changes in the serotype-virulence relation and adjust the vaccine respectively.

From our experience nowadays, ten years after the initiation of the vaccine program in the U.S.A, we can confirm that serotype replacement is a fact and has been reported from several countries where the vaccine was applied [10] [11] [12] [13]. In all cases the prevalence of the vaccine serotypes reduced significantly and along with that the rates of invasive disease cases. Some non-vaccine serotypes, on the other hand, thrived and became the main reason for disease occurrence, but the incidence rates of the disease were always lower than the pre-vaccine period. The main reason behind this fact was that, respectively to serotype prevalence at times,

[1] Basic reproductive number, denoted as R_0, is the number of secondary hosts who will acquire the organism directly from one infectious host, placed in an unexposed population at equilibrium.

the vaccine was updated to 10-valent and 13-valent recently, containing hence 13 serotypes. The 13-valent conjugate vaccine was introduced to Stockholm in June 2010.

1.3 Purpose of the study

The main adjective of this study is to try and identify signs of serotype replacement among the people of Stockholm, due to the initiation of the vaccine program in 2007. We will check if the incidence of the non-vaccine serotypes has increased significantly in the post-vaccine period and at the same time we will evaluate the vaccine efficacy by observing the decrease in the vaccine serotype's incidence. Finally, a Poisson model will be constructed in order to see how factors such as age and sex affect a person's risk of getting infected by an invasive pneumococcal disease.

1.4 Data

Starting in 1997 and still ongoing, all hospitals and laboratories in the Stockholm area have initiated a population-based surveillance over invasive *S. pneumoniae cases*, obtaining specimens from patients and sending these specimens to the Swedish Institute for Infectious Disease Control (SMI) in order to be analyzed. An invasive pneumococcal disease case was defined when *S. pneumoniae* bacteria were recovered from culture of a normally sterile body fluid, such as blood or cerebrospinal fluid. The methods used by the SMI to categorize the isolates into serotypes were gel diffusion and the capsular reaction test (i.e. the Quellung test), with the use of type specific serum samples from the World Health Organization Collaborative Center for Reference and Research on Pneumococci.

In this study we included all the cases reported to the SMI between 1 January 1997 and 31 December 2010, giving a total of 3476 cases of IPD. The age of the patients was available in 3442 (99%) cases, the sex in 3348 (96%) cases, the serotype in 3466 (99%) cases and the date a sample was taken was available in all cases. There were several double entries to be cleared out, which had occurred because in some cases two samples were taken from the same patient (usually from the blood and the liver) in order to identify the serotype or because the same patient had entered the hospital again in the same month and the reason for that was the serotype initially labeled. Following the structure of the data set, the patients were divided into four age groups and more specifically infants (0-2 years), younger children and adolescents (3-17 years), adults aged from 18 to 64 years and finally people older than 65 years, accounting for 5.3%, 3.7%, 50% and 41% of the total cases respectively. The annual incidence of IPD was calculated as the total number of cases

divided by the population of the Stockholm area and then expressed per 100.000 people. The same procedure was followed when we wanted to estimate the incidence of specific serotypes. The data for the total population of Stockholm each year and for the age groups defined above were obtained from Statistics Sweden [15].

As we have already mentioned earlier, a conjugate vaccine (PCV-7) containing the seven more prevalent serotypes at the time (2000) was introduced in the United States and was followed by a large decrease in the number of IPD incidence. This exact vaccine, containing the same seven serotypes, started being used in Stockholm in July 2007 and was given to children up to two years old in a 3-doses schedule. The seven serotypes contained in the vaccine were: *4, 6B, 9V, 14, 18C, 19F* and *23F*, while all the rest we consider as non-vaccine serotypes in our analysis. In June 2010 this vaccine was substituted by a 13-valent conjugate vaccine containing extra the following serotypes: *1, 3, 5, 6A, 7F* and *19A*.

Since the vaccination program initiated in Stockholm in the middle of 2007 we have decided to consider this year as an intermediate and hence not take it under consideration when we compare rates before and after vaccination. Thus, as pre-vaccine period we define the time from 1 January 1997 to 31 December 2006 and as post-vaccine the period from 1 January 2008 to 31 December 2010.

Closing this data description we should refer to serotype *6C* which seems to appear suddenly in our data in 2005. The truth behind this is that up to 2005 the laboratories were not able to distinguish between serotypes *6B* and *6C* considering, thus, all the cases as *6B* responsible. It would be safe to assume, consequently, that the prevalence of *6B* might have been a bit inflated in the pre-vaccine years due to this reason.

1.5 Disposition

In the introductory first chapter of this report the reader has already found some general information regarding *Streptococcus Pneumoniae* and a presentation of the characteristics of serotype replacement. In addition, a description of the data was given along with the aim of the study. In Chapter 2 a statistical theoretical background is presented, focusing on Poisson modeling, the phenomenon of overdispersion and methods of overcoming it. Chapter 3 provides the reader with a descriptive analysis of the results for each age group, a meta-analysis of the serotype's behavior for the whole population as well as the outcome of the fitted Poisson model. Finally, in Chapter 4 the conclusions derived from this study are being discussed, while in the Appendix that follows we present some basic theory and additional auxiliary outcomes.

❖ 2

Background

2.1 Poisson regression

When we are dealing with count data, such as number of deaths due to a disease, number of car accidents on a highway or in our case the incidence of IPD cases in a specified time interval, and we assume that these cases are independent between them, we usually use the Poisson distribution to model for these data. So, if Y_1,Y_2,...,Y_n is a set of independent random variables expressing counts the way defined above then we would expect them to follow a distribution of the form

$$f(y_i;\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \tag{2.1}$$

where y_i is the non-negative number of counts observed and μ_i represents the expected value $E(Y_i) = \mu_i$. Since the Poisson distribution belongs to the exponential family, a generalized linear model with a Poisson distributed response variable will be given as follows:

$$\log \mu_i = x_i^T \beta = \sum_{j=1}^{p} x_{ij}\beta_j \tag{2.2}$$

This is called a Poisson regression model where (β_1,β_2,...,β_p) is the set of the parameters, x_i^T is the i'th row of the design matrix and as link function we use the log-link, since this is the canonical link for the Poisson distribution. A reader not familiar with this terminology is recommended to go through parts 5.1-5.3 in the appendix. A very important property of the Poisson distribution is that the expected value and the variance of a Poisson random variable are equal and more analytically

$$E(Y) = var(Y) = \mu \tag{2.3}$$

Although this stands in theory, in practice we often come across situations where this property is being violated. More details on this statistical issue and ways to overcome it are given in the following parts.

2.2 Overdispersion

The phenomenon of overdispersion is observed when our data exhibit larger variability than would be expected by the suggested distribution function. In practice and when dealing with count data, we observe overdispersion in the case where

$$var(Y) > \mu \tag{2.4}$$

The reasons that lead to that can be from an incorrect choice of link function or the presence of outliers to more difficult issues like the lack of independence among the observations and subject heterogeneity. By the latter we mean that there are some explanatory variables missing from our model regarding the subject's status, maybe because they are unknown to us, leading, thus, to some excessive variance in the end that our model could not predict.

One easy way to check for overdispersion is to divide the Pearson's Chi-square statistic by the corresponding degrees of freedom. This ratio gives us an estimate of the dispersion parameter and if it has a value over one then we conclude that overdispersion is present. If we ignore this fact and apply a Poisson regression model anyway, in which it is assumed that the dispersion parameter is equal to one, then, although this will result to the correct parameter estimates, the standard errors will be erroneously smaller leading us to falsely narrower confidence intervals.

We understand, thus, that using a Poisson model under these circumstances would not be the optimal thing to do. Two methods that are widely used when dealing with overdispersion and are presented here are the Negative Binomial and Quasi-Poisson regression models.

2.3 Negative Binomial regression

Negative Binomial regression is an extension to Poisson regression allowing for greater variability than that in the Poisson model and is based on the Negative Binomial distribution. To make the connection between the two distributions clearer suppose we have a random variable that is conditionally Poisson distributed given the parameter λ ($\lambda > 0$) where

$$E(Y) = \lambda \tag{2.5}$$

and that λ follows a Gamma distribution *Gamma*(k,μ).

From this we come to the conclusion that

$$E(\lambda) = \mu$$

and

$$var(\lambda) = \mu^2/k \text{ where } \mu, k > 0 \tag{2.6}$$

The Negative Binomial distribution can be considered then as gamma mixture of Poisson distributions which marginally yields

$$P(Y = y) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(1 - \frac{k}{\mu+k}\right)^y, y = 0, 1, 2... \tag{2.7}$$

where Γ is the gamma function. The mean and variance of the Negative Binomial distribution are:

$$E(Y) = \mu \text{ and } var(Y) = \mu + \frac{\mu^2}{k} \tag{2.8}$$

We observe that the variance of the Negative Binomial distribution is not the same as the mean, as it would occur in a Poisson distribution Po(μ) , but instead the variance is the mean multiplied by the term $1+\mu/k$, which expresses the overdispersion. The term k^{-1} is called the dispersion parameter and we observe that as $k^{-1} \to 0$ the Negative Binomial distribution converges to the Poisson distribution since

$$var(Y) \to \mu \tag{2.9}$$

If k^{-1} is not given then maximum likelihood estimation can be used to estimate it, although some iterative method like Newton-Raphson will be necessary. When given or estimated then the Negative Binomial distribution belongs to the exponential family and hence can be used to create a generalized linear model with the log-link as the most natural choice.

2.4 Quasi-Poisson model

The likelihood equations in a generalized linear model, as it is explicitly shown in part 5.3 of the Appendix, have the following form:

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \tag{2.10}$$

where the term $\frac{\partial \mu_i}{\partial \eta_i}$ depends on the chosen link function each time. What is apparent from the above formula is that only the knowledge of the mean and variance of the

Y_i's is enough to estimate the likelihood equations and that the distribution they follow is unimportant as long as it belongs to the exponential family.

In fact, the variance itself is a function of the mean and based on this Wedderburn proposed the so called quasi-likelihood estimation in 1974 . We only consider a mean-variance relationship of the form $var(Y_i) = v(\mu_i)$, a link function and linear predictor, but no assumption regarding the distribution of the Y_i 's. The quasi-likelihood equations, thus, are not likelihood equations in the traditional way, since we are lacking the necessity of the random component belonging to an exponential family. They will, though, give the same results if, for example, we assume a mean-variance relationship of the form $var(Y_i) = \mu_i$, consider that the Y_i 's follow a distribution in the exponential family and compare it with the case that $Y_i \sim Po(\mu_i)$.

The quasi-likelihood estimation becomes a lot more interesting, though, when we use it for mean-variance relationships that we don't come across in the exponential family of distributions. In particular, the quasi-Poisson model that adjusts for overdispersion has a mean-variance relationship of the form $v(\mu_i) = \phi\mu_i$, where in case $\phi > 1$ then ϕ would capture the overdispersion and of course a log-link. If we replace this variance form in the likelihood equations we gave in the very beginning we will actually end up with the same estimates as we would have gotten by using the Poisson model, because the term ϕ will drop out. The same is not valid, though, concerning the variance of the parameter estimates. Under the quasi-Poisson model they have larger variance, which leads to larger standard errors and hence to bigger p-values and broader confidence intervals than the ones the standard Poisson model would have falsely given us.

In an attempt to make a comparison between the quasi-Poisson model and the Negative Binomial one, we should say that there is no golden rule about which to favor and quite often they actually give similar results. Their difference lies on the fact that the two models treat overdispersion differently. While in the quasi-Poisson model the overdispersion is constant, in the Negative Binomial it is itself a function of the mean. This leads us back to the cause of overdispersion which usually is subject heterogeneity and hence to the question how the factors not included in our model inflate the variance. Of course, different problems would give different answers to this question and respectively one of the two methods would be more preferable. In practice, choosing one of these models results in a different way of weighing our observations. This is due to the fact that we use the weighted least squares method when fitting these models and these weights are inversely proportional to the variance.

❖ 3

Results

In this chapter we are going to present the results of the analysis done on the dataset, which we have divided into three parts. In the first part we have found the distribution of the serotypes throughout our study period and checked the incidence of vaccine and non-vaccine serotypes in general and for the four predetermined age groups. Supplementary, we decided to take a more thorough look individually for the main non-vaccine serotypes within each age group, in an effort to identify signs of serotype replacement. In the second part, we made a meta-analysis where we estimated the relative risk of getting infected by each of the vaccine serotypes in the post-vaccine period compared to the pre-vaccine one and the same procedure was followed for the non-vaccine serotypes. Finally, we applied a Poisson model to our data in order to see more clearly the age and sex effect, the vaccine efficacy and to check the significance of the interactions of the main effects. The statistical software used for all the above purposes was R (R version 2.8.1), while some initial exploratory work was done using Microsoft Excel 2003 to get better acquainted with the data.

3.1 Serotype behavior

We will start this part by a presentation of our results concerning the whole population and regardless of the patient's age. Our goal is to obtain a general idea of what has occurred in the previous 14 years (1997-2010) in the Stockholm area, which serotypes were more prevalent and how the vaccine has altered the natural occurrence of IPD. After that, we will take a closer look at the changes within each age group regarding the vaccine's effect and serotype replacement.

The rates of both vaccine and non-vaccine serotypes throughout our entire study period can be seen in Figure 3.1. The instability observed is normal when dealing with diseases caused by bacteria, as in our case, where years with outbreaks as occurred in 2008 and more recessive years like 2002 are to be expected.

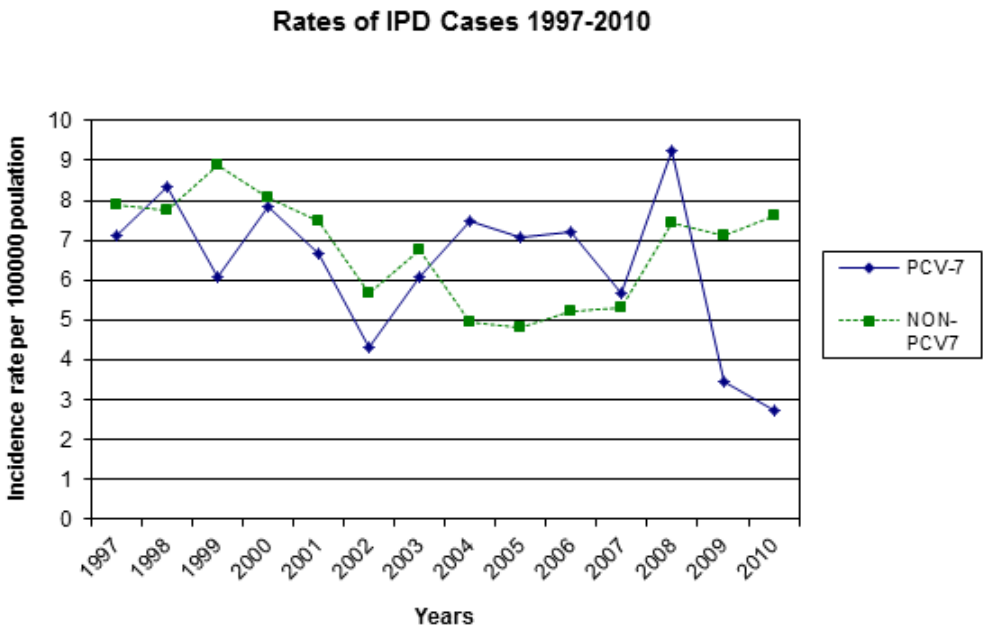

Fig. 3.1. Incidence rates per 100.000 population for vaccine (PCV-7) and non-vaccine serotypes (Non-PCV7) through years 1997-2010.

What is highly remarkable is that the outbreak observed in 2008, which was mostly caused by vaccine serotypes, occurred after the vaccination program initiation in July 2007, which poses the natural question about the extent of the vaccine serotype outbreak in the absence of the vaccine. The vaccine's effect became indisputably apparent in the year 2009, where the incidence of the vaccine serotypes dropped from 9.24 cases per 100.000 population to 3.46 cases respectively, to become only 2.73 cases/100.000 population the next year.

One could have expected that the vaccine serotypes would have been a lot more prevalent in the pre-vaccine period than the ones not included in the vaccine. We should at this point remind that the Prevnar(PCV-7) vaccine used in Stockholm is the same as the one given in the United States in 2000, which was manufactured to deal with the serotypes that domestically had the biggest prevalence. As it can be seen in Figures 3.2 and 3.3, where the distribution of vaccine and non-vaccine serotypes is described respectively, serotypes *18C* and *19F* included in the vaccine are not as dangerous as the others included. Some non-vaccine serotypes like serotype *1,3* and *7F* have been giving larger numbers of incidence throughout the years and most probably a vaccine manufactured in Sweden would have contained a different selection of serotypes. We should mention that in the graph below we only give the distribution of the main 9 non-vaccine serotypes, the ones that gave at least 50 cases in the past 14 years. In total, we observed 58 serotypes this past period which

leaves us with 42 serotypes, which from time to time had some low incidence rates. Keeping all these in mind helps us understand better how the incidence of the non-vaccine serotypes can be as large, in some years even larger, as the vaccine serotypes' incidence.

Using figures 3.2 and 3.3 we can get a clearer view of what exactly took place in 2008 and how the vaccine effected on the serotypes. Regarding the vaccine serotypes, which had a more dramatic increase in 2008, we see that all of them had higher incidence than the previous year while serotypes *9V, 14* and *23F* reached their peak for the past decade. We can say for sure that serotype *14* has been the most virulent serotype in Stockholm the last 14 years. Starting in 2009 and continuing the next year, we can see the obvious decrease in the incidence of all vaccine serotypes, with the exception of serotype *4*, but it is expected that the vaccine won't have the same effect on all serotypes. The increase of the non-vaccine serotypes in the 2008 outbreak was a bit milder than the vaccine ones with only serotype *22F* standing out. In our effort to identify signs of serotype replacement, we would expect considerably increased incidence for the following years also. This is definitely the case for serotypes *19A, 22F* and *6C* and seems to stand also for serotypes *3* and

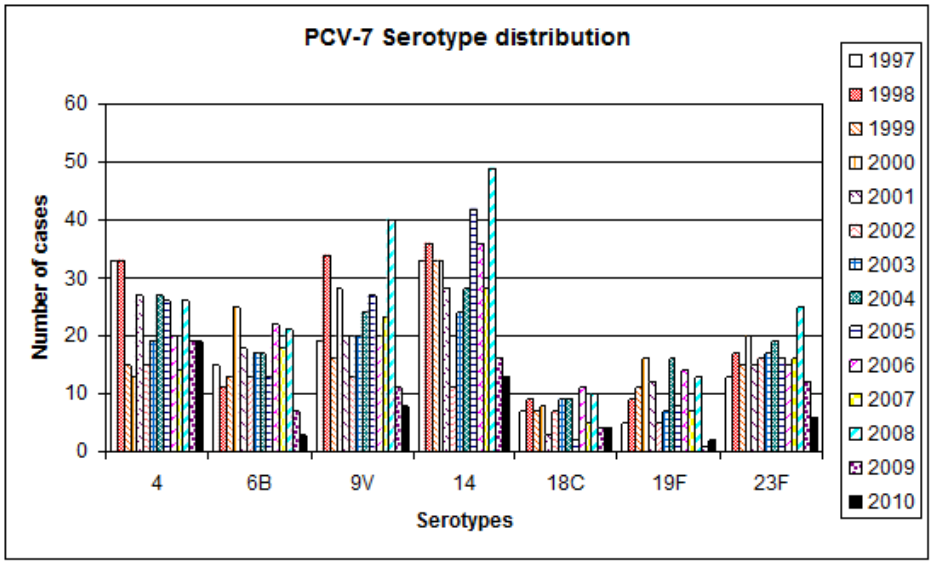

Fig. 3.2. The distribution of the seven vaccine serotypes (PCV7) through the period 1997-2010.

7F, but of course a more detailed statistical analysis is required. A fact that we should keep in mind and explains why some non-vaccine serotypes remained stable or showed reduced incidence in year 2010 is the initiation of the new vaccine in

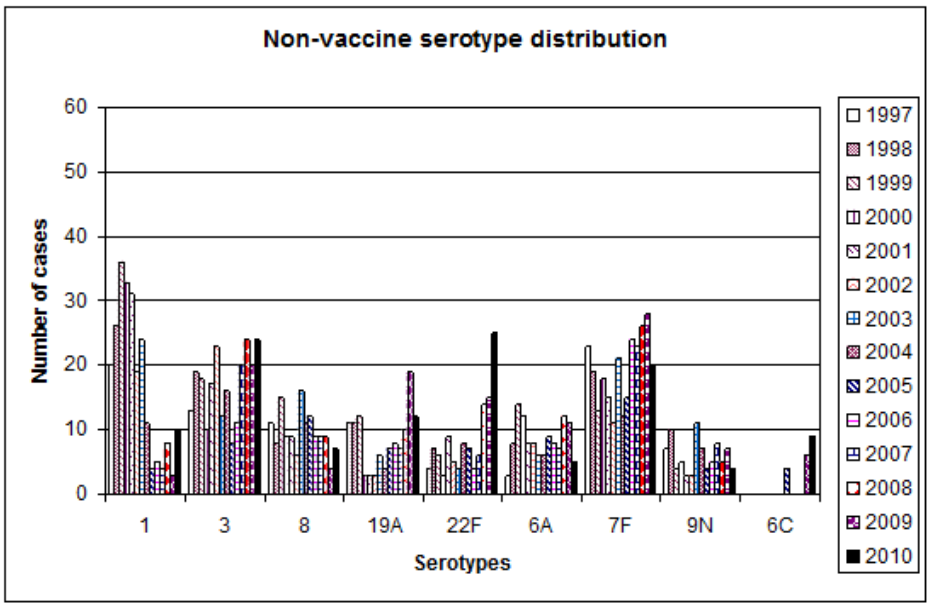

Fig. 3.3. The distribution of the non-vaccine serotypes (Non-PCV7) through the period 1997-2010.

June that year, containing 13 serotypes now, which were the 7 of the first vaccine and in addition serotypes *1, 3, 5, 6A, 7F, 19A*.

Closing this first part of general comments on the data, we feel that we should make a particular reference to serotype *1* and the special distribution it presents. We cannot state exactly what might has happened, but only speculate to explain why, while being one of the most prevalent serotypes until 2003, it dropped significantly afterwards. We know that the distribution of a serotype varies over time, but Serotype *1* is also known to have two more characteristic features. The first is that it is associated with large outbreaks, although rarely carried for long and the second is that it is unlikely to be found on a healthy carrier, which shows that it is highly invasive. The combination of high attacking rate and history of outbreaks have made serotype *1* a great danger when considering the serotype replacement that often follows a vaccination. For that reason, including it in the 13-valent vaccine that was composed later was a necessity [16].

Age group 0-2 years

As it can be seen in Figure 3.4 that follows, this age group has had the second highest incidence of IPD cases throughout the whole period under study, being surpassed only by the elderly, giving an average of 22.04 cases/100.000 population each year in the pre-vaccine period and reaching a maximum of 35.42 cases/100.000 population in 2001. The age group was not affected at all during the 2008 outbreak, on the

contrary it showed a large decrease, which is reasonable if we consider that we are talking about the age group that started receiving the vaccine in the middle of 2007.

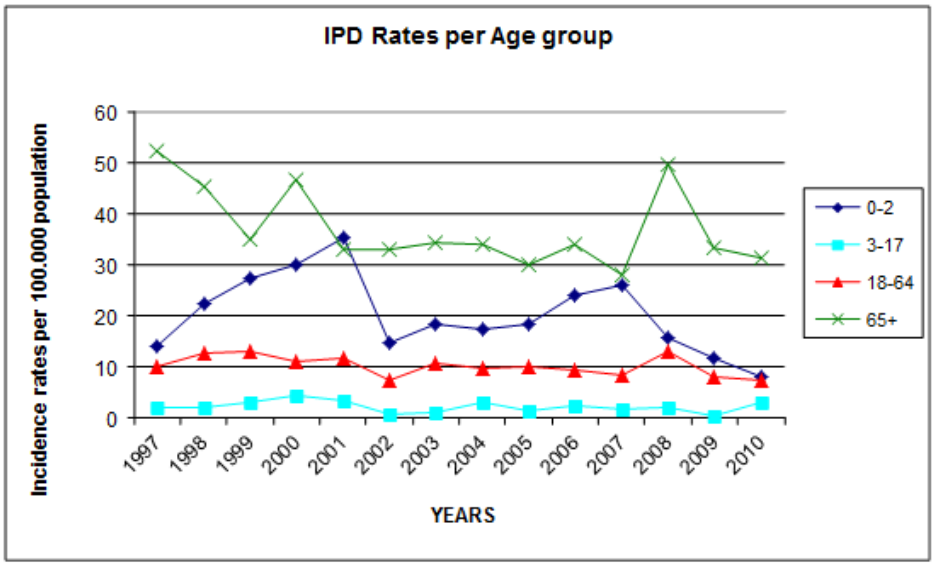

Fig. 3.4. Incidence rates per 100.000 population for the four age groups through the period 1997-2010.

In Table 3.1 that follows we see that the overall annual rate of IPD for this age group decreased from an average of 22.04 cases/100.000 population in the pre-vaccine period to 11.84 cases/100.000 population afterwards. The relative risk of getting sick with IPD in the post-vaccine period was 0.54, which means that it was 46% less likely for someone to get the disease in the after period than before the vaccination. A 95% confidence interval for this relative risk ranges from 0.36 to 0.80, a wide range due to the relatively small number of cases, not including the value 1 though, which means that it is a significant effect as we can also see from the p-value ($p < 0.01$). This result was mostly due to the significant decrease in the rate of disease caused by vaccine serotypes, from an annual average of 14.45 cases/100.000 population in the period 1997-2006 to 4.34 cases/100.000 population in the period 2008-2010 ($p < 0.01$). The annual rate of the non-vaccine serotypes remained almost identical, from an average of 7.59 cases/100.000 population in the pre-vaccine period to 7.50 cases/100.000 population in the after period ($p = 0.97$), so there were no signs of serotype replacement taking place. Indeed, after checking for each of the main non-vaccine serotypes for this age group we found out that none of them presented a statistically significant increase. We should mention at this point that in the table below we have included the non-vaccine serotypes that had at least 10 cases in the past 13 years (excluding year 2007), since it would be very

difficult to get any significant results with even lower incidence. In the paragraph 5.5 of the Appendix one can see analytically the distribution of the vaccine and the main non-vaccine serotypes for this age group.

Table 3.1. Invasive pneumococcal disease among young people (0-2 years) in the Stockholm area, before (1997-2006) and after (2008-2010) the vaccination program initiated.

	No. Of Cases		Cases/100.000 People		Rel.Risk(95% CI)	P-value
Serotype	1997-2006	2008-2010	1997-2006	2008-2010		
Overall	151	30	22.04	11.84	0.54(0.36,0.80)	<0.01
PCV7	99	11	14.45	4.34	0.30(0.16,0.56)	<0.01
Non-PCV7	52	19	7.59	7.50	0.99(0.58,1.67)	0.97
7F	15	8	2.19	3.16	1.44(0.61,3.40)	0.48
19A	10	3	1.46	1.18	0.81(0.22,2.95)	1.00
SEX						
M	96	22	27.38	16.85	0.62(0.39,0.98)	0.04
F	55	7	16.45	5.70	0.35(0.16,0.76)	<0.01

Finally, another important result lies in the very end of Table 3.1 where we see that the vaccine effect was significant for both males and females. In addition, we see that the incidence rate of males getting sick with IPD is a lot higher than that of females, both in the pre- and post-vaccine period. This did not come as a big surprise since it had been already noted in previous studies that males are more likely to get infected than females [4] [17]. We see that before the vaccination the incidence rate for males was 22.38 cases/100.000 population, while for females it was 16.45 cases/100.000 population. We have checked if this difference was significant and it proved to be, since we found a value of 0.60 for the relative risk for girls getting infected as compared to boys, with the 95% confidence interval for it being (0.43,0.84) and a very small p-value ($p<0.01$). The same result of significant difference stands for the post-vaccine period, where the relative risk was now 0.34 (0.14, 0.79) and again $p<0.01$. Thus, at least for this age group, there is a significant difference between the two sexes regarding the way they get infected with an IPD.

As a final comment we need to point out that the vaccine seems to have a different effect on both sexes, with females being more affected by it. More particularly, the relative risk for males in the post-vaccine period was 0.62, while for females it was 0.35. Due to the small number of cases, the confidence intervals for both sexes are rather wide. In order to check if this difference was significant we calculated the ratio of these relative risks to find a point estimate of 1.77 with a 95% confidence interval being (0.51, 6.18) and p-value ($p = 0.37$). Consequently, there seems to be no difference of the vaccine's effect between males and females.

Age group 3-17 years

We will start again our analysis by looking at Figure 3.4 to see that this age group has had steadily the lowest incident rates throughout our whole study period. In Table 3.2 that follows, we see that the overall incidence decreased from an average of 2.43 cases/100.000 population in the pre-vaccine period to 1.88 cases/100.000 population in the post-vaccine one, a decrease that was not significant (p = 0.30). The results were similar when we looked more closely at the vaccine and non-vaccine serotypes separately.

Table 3.2. Invasive pneumococcal disease among young people (3-17 years) in the Stockholm area, before (1997-2006) and after (2008-2010) the vaccination program initiated.

	No. Of Cases		Cases/100.000 People		Rel.Risk(95% CI)	P-value
Serotype	1997-2006	2008-2010	1997-2006	2008-2010		
Overall	81	20	2.43	1.88	0.77(0.47,1.26)	0.30
PCV7	33	8	0.99	0.75	0.76(0.35,1.64)	0.48
Non-Vaccine	48	12	1.44	1.13	0.78(0.42,1.47)	0.45
1	15	1	0.45	0.09	0.21(0.03,1.58)	0.14
7F	11	2	0.33	0.19	0.57(0.13,2.57)	0.75
22F	2	3	0.06	0.28	4.70(0.79,28.12)	0.1
SEX						
M	62	16	1.86	1.50	0.81(0.47,1.4)	0.45
F	19	3	0.57	0.28	0.50(0.15,1.67)	0.25

More specifically, the incidence of the vaccine serotypes decreased from 0.99 cases/100.000 population in the period 1997-2006 to 0.75 cases/100.000 population in the period 2008-2010 and that of the non-vaccine serotypes from 1.44 cases/100.000 population to 1.13 cases/100.000 population. Both changes were not significant giving p-values p=0.48 and p=0.45 respectively. When looking at the main non-vaccine serotypes individually, we observe that none of them presents a significant change in a 5% significance level. By the term main non-vaccine serotypes we mean again those that had at least 10 cases in the period under study, the distribution of which can be seen in paragraph 5.5 of the Appendix. What stands out and is worth noting, is of course the increased risk of getting infected by serotype *22F*, which for the post-vaccine period is 4.7 times higher than before, but the number of cases is too low leading to a very broad confidence interval (0.79, 28.12) and a not significant p-value (p = 0.1). Finally, closing our description of results for this age group, which is characterized by a small number of cases, we should mention that, similarly to the previous age group, the incidence of males was again significantly

higher than that of females for both the pre- and post-vaccine period ($p<0.01$ on both occasions).

Age group 18-64 years

We can start our analysis of this age group by looking at Figure 3.4 to obtain a general idea of what has occurred the previous years. We see that this age group has had steadily lower incidence rates than both the group of the youngest children and that of the elderly, having an average of 10.28 cases/100.000 population each year.

In Table 3.3 that follows we see that the overall annual rate of IPD decreased from an average of 10.58 cases/100.000 population in the pre-vaccine period to 9.54 cases/100.000 population in post-vaccine period. This is translated to a 0.90 relative risk for the after period with a 95% confidence interval being (0.80, 1.01), meaning that after the vaccination program had started, people of this age group were 10% less likely to get infected. The decrease was not significant, though, in a 5% significance level giving us a p-value of ($p = 0.08$).

Table 3.3. Invasive pneumococcal disease among middle-aged people (18-64 years) in the Stockholm area, before (1997-2006) and after (2008-2010) the vaccination program initiated.

	No. Of Cases		Cases/100.000 People		Rel.Risk(95% CI)	P-value
Serotype	1997-2006	2008-2010	1997-2006	2008-2010		
Overall	1245	367	10.58	9.54	0.90(0.80,1.01)	0.08
PCV7	587	154	4.99	4.00	0.80(0.67,0.96)	0.015
Non-Vaccine	658	213	5.59	5.53	0.99(0.85,1.16)	0.9
1	150	16	1.27	0.42	0.33(0.19,0.55)	<0.01
3	63	28	0.54	0.73	1.36(0.87,2.12)	0.17
6A	27	10	0.23	0.26	1.13(0.55,2.34)	0.74
7F	98	51	0.83	1.33	1.59(1.13,2.23)	<0.01
8	74	14	0.63	0.36	0.58(0.33,1.02)	0.06
9N	25	9	0.21	0.23	1.10(0.51,2.36)	0.8
19A	24	11	0.20	0.29	1.40(0.69,2.86)	0.35
22F	27	25	0.23	0.65	2.83(1.64,4.88)	<0.01
SEX						
M	640	188	10.86	9.73	0.90(0.76,1.05)	0.18
F	560	176	9.53	9.18	0.96(0.81,1.14)	<0.01

This result was mostly a consequence of the significant decrease in the rate of disease caused by vaccine serotypes, from an annual average of 4.99 cases/100.000 population before the initiation of the vaccine to 4.0 cases/100.000 population afterwards ($p= 0.015$). We can say, thus, that the signs of herd immunity as discussed

in the introduction are here apparent. The non-vaccine serotypes, which are the main reason of disease in this age group, remained almost stable by changing from 5.59 cases/100.000 population to 5.53 cases/100.000 population in the pre- and post-vaccine period respectively (p= 0.9). The reasons behind this result become more obvious if we check for each of the main non-vaccine serotypes individually. We see, then, that the first and third more incidental serotype of the pre-vaccine period (serotypes *1* and *8*) have a significant and almost significant decrease of their rates with p-values being p <0.01 and p = 0.06 respectively, while all the rest have increased their incidence rates and especially for serotypes *7F* and *22F* this increase was significant (p<0.01 for both). In paragraph 5.5 of the Appendix, one can see the distribution of the vaccine and the main non-vaccine serotypes for this age group in order to visualize the results of Table 3.3.

Following, we resulted that the rate of IPD cases among males is again larger than that for females as in the previous groups and, more particularly, for the period before the vaccination we had an annual average of 10.86 cases/100.000 population for men and 9.53 cases/100.000 population for women. This difference proved once more to be significant giving a relative risk of 0.88 for women to get infected as opposed to men, with a 95% confidence interval being (0.78, 0.98) and p-value p = 0.023. So, similarly to the previous age groups, there is again a significant difference between the two sexes, with males being more prone to get an IPD infection than women.

Age group 65+

The group of elderly people is by far the one having the highest incident rates when compared to the other age groups, as we can see from Figure 3.4. The incidence rate's peak lies in 1997 with the impressive number of 52.26 cases/100.000 population. We see that the outbreak of 2008 affected mostly this age group, leading from an incidence rate of 28.18 cases/100.000 population in 2007 to 49.72 cases/100.000 population the next year.

In Table 3.4 that follows, we can see that the overall annual rate of IPD remained unchanged from an average of 37.82 cases/100.000 population in the pre-vaccine period to 37.78 cases/100.000 population in post-vaccine period. This means that the elderly were almost equally likely to be infected from an IPD (Relative Risk = 0.99) after the initiation of the vaccine program (p = 0.98). The meaning of this result, though, is far from indicating a stable situation. On the contrary, we see that the incidence due to vaccine serotypes decreased significantly from 19.98 cases/100.000 population to 15.18 cases/100.000 population (p<0.01) in the after period, while at the same time the non-vaccine serotypes increased their incidence also significantly from 17.84 cases/100.000 population in the pre-vaccine period to 22.60 cases/100.000 population afterwards (p <0.01).

Table 3.4. Invasive pneumococcal disease among elder people (>65 years) in the Stockholm area, before (1997-2006) and after (2008-2010) the vaccination program initiated.

	No. Of Cases		Cases/100.000 People		Rel.Risk(95% CI)	P-value
Serotype	1997-2006	2008-2010	1997-2006	2008-2010		
Overall	990	336	37.82	37.78	0.99(0.88,1.13)	0.98
PCV7	523	135	19.98	15.18	0.76(0.63,0.92)	<0.01
Non-Vaccine	467	201	17.84	22.60	1.27(1.07,1.49)	<0.01
1	40	4	1.53	0.45	0.29(0.11,0.82)	0.013
3	78	35	2.98	3.93	1.32(0.88,1.97)	0.17
6A	37	18	1.41	2.02	1.43(0.82,2.51)	0.21
7F	51	17	1.95	1.91	0.98(0.57,1.70)	0.95
8	29	6	1.11	0.67	0.61(0.25,1.47)	0.26
9N	31	6	1.18	0.67	0.57(0.24,1.37)	0.20
11A	34	7	1.3	0.79	0.61(0.27,1.37)	0.22
19A	28	22	1.07	2.47	2.31(1.32,4.04)	<0.01
22F	27	26	1.03	2.92	2.83(1.65,4.86)	<0.01
SEX						
M	409	146	38.41	38.01	0.99(0.82,1.20)	0.91
F	535	185	34.46	36.60	1.06(0.90,1.26)	0.48

Looking more analytically at each non-vaccine's serotype behavior we understand that they actually form three groups. In the first is serotype *1* which was the only serotype to show significant decrease when comparing the pre- and post - vaccine period. In the second group we include serotypes *19A* and *22F* whose incidence rates increased significantly between the two periods and the third one contains all the rest which present either an increasing or decreasing trend but never any significant change. We should mention at this point that in the above table we have included those non-vaccine serotypes that had at least 30 cases in the past 14 years for this age group. There is a large number of other non-vaccine serotypes that had low incidence in the pre-vaccine period, but their post-vaccine incidence rates are already higher than before. Their small number of cases prevents us from seeing any significant difference individually, but in total they are a very important factor for the overall non-vaccine serotype's increase. The distribution of the vaccine and the main non-vaccine serotypes for this age group in order to give a clearer picture of what has occurred is given in paragraph 5.5 of the Appendix.

Closing this presentation of results regarding the group of elderly people, we see that once more men were more likely to be infected than women. In the pre-vaccine years, the average annual incidence rate for men was 38.41 cases/100.000 population, while for women it was 34.46 cases/100.000 population. The relative risk, thus, of a woman getting infected as opposed to a man came out to be 0.9 with

a 95% confidence interval for that value being (0.79,1.02) and giving a p-value p = 0.0978. So in a 5% significance level we can say that there is no difference in the way the two sexes get infected by IPD in this age group.

3.2 Meta-anaysis

In an effort to sum up the results from all the age-groups mentioned above and to have a clearer overall view of the changes in the serotype behavior due to the initiation of the vaccine program in 2007, we decided to make a meta-analysis. In the forest plot (Figure 3.5) one can see the relative risk of getting infected from an IPD in the years after the vaccination (2008-2010) compared to the previous years (1997-2006) for each vaccine serotype individually. In addition, we found the Mantel-Haenszel summary relative risk of getting infected from an IPD due to a vaccine serotype in the after period in general. There are two different approaches to estimating the common pooled effect in cases like this. The first is based on a fixed effects model, where we only consider the within-serotype variance and the other is based on a random effects model, where besides the within-serotype variance we acknowledge for between-serotype heterogeneity. A common prerequisite in both scenarios, which is often neglected, is the independency of the serotypes. Although we have no indications that the latter is not true, it is a subtle issue that we should keep in mind and maybe some future studies can provide us with more insight.

We started by checking for heterogeneity between the vaccine serotypes which resulted in a p-value of (p=0.357), so we can accept the null hypothesis of homogeneity and hence a fixed effects model is adequate. The pooled relative risk is thus estimated as a weighted average of each serotype's relative risk individually with the weights being inversely proportional to the serotypes variances. More analytically, the pooled relative risk is found by $\bar{Y} = \sum_{k=1}^{7} W_i Y_i / \sum_{k=1}^{7} W_i$ where Y_i is the relative risk of each vaccine serotype and the weights W_i are inversely proportional to each serotypes variance $W_i = 1/V_i$. To make the forest plot more interpretable we have to explain that the boxes represent the point estimates and their height is inversely proportional to the standard error of the estimate, while the horizontal lines express a 95% confidence interval of this value. The diamond, on the other hand, is the point estimate of the summary relative risk, with the horizontal limits defining a 95% confidence interval and the width being again inversely proportional to the standard error of the estimate. In Table 3.5 that follows the graph, we give the values of the point estimates for each serotype along with their 95% confidence intervals as derived from the meta-analysis, in order to have a more definite view of the results. A first comment by looking at Figure 3.5 is that all serotypes have point estimates that are lower than 1. Two of them, serotypes *6B* and *19F*, have confidence intervals that

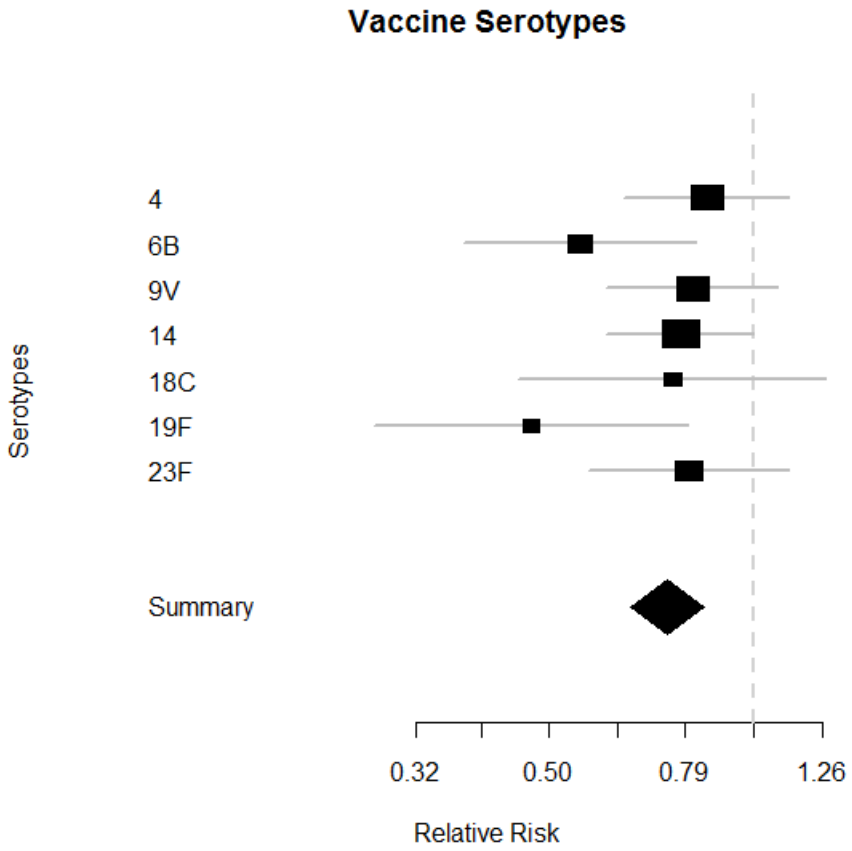

Fig. 3.5. The relative risk of getting infected with IPD in the post-vaccine period (2008-2010) from a vaccine serotype compared to the pre-vaccine period (1997-2006).

do not cross the border of value 1, suggesting a significant decrease in their incidence rates between the pre- and post-vaccine period, while serotype *14* has an upper limit which is exactly 1, being hence on the verge of significance ($p = 0.0504$). All the other serotypes have confidence intervals that include the value 1 and consequently their incidence rate's decrease can not be considered significant. To gain a higher understanding of these results, a look at Figure 3.2 is recommended, which presents the distribution of the vaccine serotypes. It is obvious that the serotype outbreak of 2008 distorts what we would expect as a vaccine effect, which finally becomes apparent in 2009 and afterwards. The small incidence rates of serotypes *18C* and *19F* also justify the large confidence intervals that they present in Figure 3.5. The pooled conclusion for the vaccine serotypes as a group showed a clear decrease in the incidence rates between the two periods, with the point estimate for the relative risk being 0.75 and the 95% confidence interval having limits (0.66, 0.84).

We understand, thus, that the risk of getting infected with an IPD by a vaccine serotype is 25% less in the post-vaccine period and that this effect ranges between 16% and 34%.

Table 3.5. Analytical results of the meta-analysis on the vaccine serotypes.

Serotype	Relative Risk (Point estimate)	95% Confidence Interval
4	0.85	(0.65,1.13)
6B	0.56	(0.38,0.82)
9V	0.81	(0.61,1.08)
14	0.78	(0.61,1.00)
18C	0.76	(0.45,1.27)
19F	0.47	(0.28,0.80)
23F	0.80	(0.57,1.12)
Overall	**0.75**	**(0.66,0.84)**

The same procedure was followed in order to investigate the effect of the vaccine in the behavior of the non-vaccine serotypes. We started by checking the null hypothesis of homogeneity among the non-vaccine serotypes, which resulted in a very small p-value ($p<0.001$). The use of a random effects model to derive the pooled relative risk is thus more appropriate, which, as we have already mentioned, takes under consideration the between serotypes variance too. The pooled relative risk is found in the exact way as in the fixed effects model, with the difference that the weights are now of the form $W_i = 1/(V_i + \tau^2)$, where the term τ^2 is the estimate of the between-serotype variance. Due to this different form of the weights the confidence interval derived from this method for the pooled effect is always more conservative compared to the one coming from the fixed effects model. In the forest plot that follows (Figure 3.6) we can see the relative risk of getting infected in the post-vaccine period by each non-vaccine serotype individually as opposed to the pre-vaccine era, while at the bottom we see the Mantel-Haenszel summary relative risk for all the non-vaccine serotypes pooled together. The characteristics of the graph, concerning the boxes, the horizontal lines and the diamond, are exactly the same as explained in the graph of the vaccine serotypes previously. We should at this point mention that in an effort to capture even small changes we have included all those non-vaccine serotypes that had at least 20 cases in the past 13 years, since we have excluded the year 2007. An exception to this rule was made for serotype *6C*, which we decided not to include in our analysis, since, as we have already pointed out earlier, until the year 2005 the laboratories would not distinguish it from serotype *6B*. In Table 3.6 that follows the graph, one can see more analytically the point estimates and the

95% confidence intervals for the relative risk of every serotype and for the overall summary relative risk.

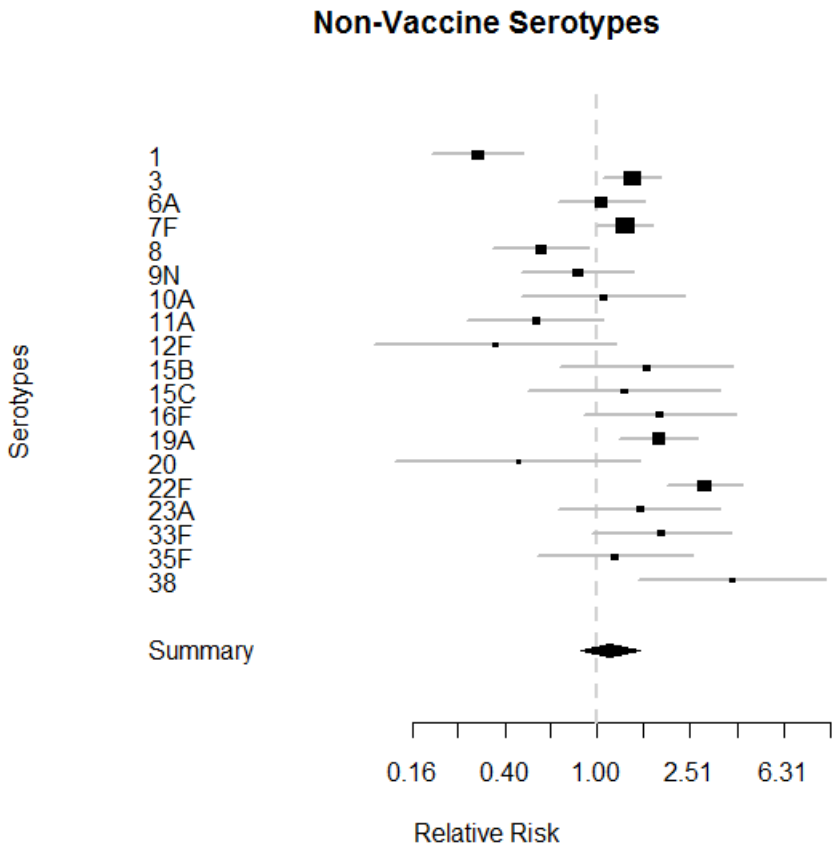

Fig. 3.6. The relative risk of getting infected with an IPD in the post-vaccine period (2008-2010) by non-vaccine serotypes compared to the pre-vaccine period (1997-2006).

Starting our analysis of the results we observe immediately the tendency of most of the non-vaccine serotypes to shift to the right side of the graph, hence being more prevalent in the post-vaccine period. Serotype *1* is the only one that has so clearly shifted to the left side ($p<0.01$), for reasons, though, that as we have already talked about have no relation to the vaccine, while serotype *8* is the only other serotype that had actually a significant decrease of incidence ($p = 0.02$) when comparing the two periods, with this being most probably attributed to natural variation. No other serotype showed a significant decrease, while on the other hand serotypes *3, 7F, 19A, 22F* and *38* presented a significant increase of their incidence with p-values: $p=0.02$, $p=0.048$, $p<0.01$, $p<0.01$, $p<0.01$ respectively. The overall summary relative risk

Table 3.6. Analytical results of the meta-analysis on the Non-vaccine serotypes.

Serotype	Relative Risk (Point estimate)	95% Confidence Interval
1	0.31	(0.19,0.48)
3	1.41	(1.05,1.87)
6A	1.04	(0.68,1.59)
7F	1.32	(1.00,1.73)
8	0.57	(0.36,0.92)
9N	0.82	(0.47,1.43)
10A	1.06	(0.47,2.36)
11A	0.54	(0.28,1.06)
12F	0.36	(0.11,1.21)
15B	1.62	(0.69,3.82)
15C	1.30	(0.50,3.39)
16F	1.86	(0.88,3.93)
19A	1.83	(1.24,2.70)
20	0.46	(0.14,1.53)
22F	2.88	(1.98,4.18)
23A	1.52	(0.68,3.38)
33F	1.88	(0.94,3.76)
35F	1.19	(0.55,2.57)
38	3.80	(1.50,9.63)
Overall	**1.13**	**(0.85,1.52)**

had a point estimate of 1.13 with a 95% confidence interval for this value being (0.85,1.52). There is, thus, a 13% increased risk of getting an IPD due to a non-vaccine serotype in the post-vaccine period, but this difference was not significant as the confidence interval reveals.

Before closing this part, some comments are necessary regarding the presence of heterogeneity among the non-vaccine serotypes. A big factor, is of course, serotype *1* with the very unique behavior it presents throughout our study period. Interestingly enough, though, when we checked what the situation would be in the absence of serotype *1* we found out that heterogeneity was still present and for the record the pooled relative risk was still not significantly larger in the post-vaccine period having a point estimate of 1.25, while (0.98, 1.6) was a 95% confidence interval. Another reason behind this heterogeneity could be the fact that vaccines like the one under study often have a partial effect on some serotypes not included in the vaccine. Serotypes *6A* and *9N* belong in the so-called vaccine related group and this could be a reason for the fact that they don't present the increase someone might expect. In addition to all the factors stated above, we should of course consider the initiation of the 13-valent vaccine,which started in June 2010, including serotypes *1, 3, 5, 6A,*

7F and *19A* affecting, hence, up to a point the incidence these serotypes would have otherwise presented in the last year of our study.

3.3 Poisson model

In an effort to understand better our dataset, the factors that are more influential and the way they relate to each other, we have fitted a Poisson regression model by using the *glm* function in R. The response variable is the number of IPD cases and we had four explanatory variables, all of which were categorical. These variables were: "Period" divided into pre- and post-vaccine, "Serotype" divided between vaccine and non-vaccine serotypes, "Sex" divided into male and female and finally "Age". The latter we decided to split into three age groups which were "0-2", "3-64" and "65+". The reasons for merging the two middle age groups as presented in our analysis before, were the very low incidence of the group "3-17", as well as, the fact that they appear to have a similar behavior as it can be seen in Figure 3.4.

In order to find the model that explains our data in an optimal way we used the stepwise method. More analytically, the variables were included in the model one at a time and their significance was judged by a chi-square test in a 5% significance level. The most significant variable was then included and the same procedure was repeated for the rest. If, by including a variable, one that was already in the model became not significant it was removed. Finally, this procedure was followed until all the terms in the model were significant, keeping in mind, though, that we are talking about hierarchical models and the inclusion of no other term made a significant impact. The model we concluded is the following:

$$\begin{aligned}\log(C_{ijkh}) = \beta_0 + \beta_i^P + \beta_j^{SX} + \beta_k^A + \beta_h^S + \beta_{ik}^{PA} + \beta_{ih}^{PS} \\ + \beta_{jk}^{SXA} + \beta_{jh}^{SXS} + \beta_{kh}^{AS} + \beta_{ikh}^{PAS} + offset(\log(T_{ijk}))\end{aligned} \tag{3.1}$$

where β_0 is the intercept, β_i^P the effect of the period (i=0,1 where 0 refers to the pre-vaccine period and 1 to the post-vaccine), β_j^{SX} denotes the sex effect (j=M,F), β_k^A expresses the age effect (k=0,1,2 where 0 refers to the younger age group and respectively the rest) and finally β_h^S represents the serotype effect (with h=0,1 where 0 refers to the vaccine serotypes and 1 to the non-vaccine). The term C_{ijkh} is the estimated number of cases for every combination of the explanatory variables, but since the corresponding population (T_{ijk}) was each time different depending on period, age and sex the inclusion of an offset was necessary.

In the paragraph 5.4 of the Appendix one can see the resulted estimates, standard errors and p-values along with some characteristics of the model. Judging from the fact that the residual deviance is almost identical to the degrees of freedom we understand that no overdispersion is present in our model. This need not be a surprise,

since our problem fits very well the classic “car accident” Poisson example. The bacteria are spread among the population, having many carriers, but for some people they become invasive. Continuing with the model characteristics, a likelihood ratio test verified the goodness of fit of the model ($p = 0.45$), something that is also apparent from the QQ-Plot, while the Residuals vs Fitted graph shows that our model fits better on the large values of the data. As we can see, the model we resulted with is rather complicated, but the three way interaction was a very significant term whose exclusion was distorting the model's efficiency so it had to be maintained.

A first conclusion that can be made from the model is that the sex effect differs among the age groups. For the younger age group, the relative risk of males getting infected with IPD is 1.81 times higher than that of women with a 95% confidence interval being (1.33,2.46) ($p<0.01$). The absence of the Sex-Period interaction in our model indicates that the above relative risk did not change significantly by the initiation of the vaccine program and hence verifies our result given earlier that the vaccine had the same effect on both sexes. The relative risk for the middle age group was 1.18 (1.07, 1.3) ($p<0.01$) and for the elderly 1.09 (0.98, 1.22) ($p=0.11$). We can see, thus, that males are in general more prone to infection than women and only after the age of 65 the two sexes get infected in a similar way.

From the Serotype-Sex interaction we see that the relative risk of females getting infected by non-vaccine serotypes compared to vaccine serotypes was 0.86 (0.75, 0.98) ($p=0.036$) times that of men, which means that men were getting significantly more infected by non-vaccine serotypes than women. Again, since no Serotype-Sex-Period is present in our model, we can say that this conclusion is not a result of men having higher incidence in general and non-vaccine serotypes being more prevalent in the post-vaccine period, as we will soon point out. It does on the other hand explain our result in Table 3.1 where we see that females had a larger decrease in their incidence due to the vaccination program compared to men.

The presence of the three-way interaction tells us that the Period-Serotype effect is different between the age groups. More analytically for the younger age group, the relative risk of getting infected by a non-vaccine serotype in the post-vaccine period compared to a vaccine serotype is 1.9 (0.88, 4.09) ($p=0.14$), where the confidence interval is rather wide due to the small number of cases. According to the model this relative risk is approximately 3.6 times higher compared to the pre-vaccine period. Similarly for the middle age group, the relative risk was 1.41 (1.15, 1.73) ($p<0.01$) and for the elderly 1.49 (1.19, 1.85) ($p<0.01$), increased by 1.27 and 1.7 times compared to the pre-vaccine period respectively. With the help of Tables 3.1, 3.2, 3.3 and 3.4 previously presented in our study, we can understand better how the above results were formulated. For the younger and middle age group it is exclusively a result of the decrease in the incidence of the vaccine serotypes due to vaccination and herd immunity respectively, since the incidence of the non-vaccine

serotypes remained stable between the two periods. On the contrary, the increase of the relative risk for the elderly was the combined result of significant decrease of the vaccine serotypes and increase of the non-vaccine ones. Hence, although it is more common to get infected by non-vaccine serotypes in the post-vaccine period due to the vaccine, a significant increase occurred only among the elderly.

Finally, another piece of information that can be obtained from the model is that the two older age groups have relative risks that are 2.95 (1.51, 5.8) ($p<0.01$) times for the middle age group and 2.85 (1.45, 5.6) ($p<0.01$) times larger for the elderly of getting infected by vaccine serotypes in the post-vaccine period than the younger age group. Comparing the elderly with the middle aged group gives us a rate of 0.96 (0.74, 1.25) ($p=0.78$), which indicates, as expected, the uniform effect of herd immunity among the population.

❖ 4

Conclusion

The initiation of the vaccine program against *Streptococcus Pneumoniae* family of bacteria in 2007 by using a 7-valent conjugate vaccine (PCV-7) was definite to bring changes to the way infectious pneumococcal disease occurs in the Stockholm area. This is not only due to the expected immediate effect on the vaccinated group and the herd immunity that follows it, but also because, as it has already been proven in previous studies, vaccinating against certain serotypes leads to the so-called serotype replacement phenomenon.

The goal of this study was to derive conclusions concerning both these aspects along with some general information relative to the way IPD incidence occurs among the population. Starting by the latter, we saw that the elderly (over 65 years) were the main risk group, followed by children up to 2 years old. Moreover, we resulted that males have a significantly higher risk of disease than females, especially for the youngest group, where the relative risk (RR) was RR: 1.81 with a 95% confidence interval being (1.33, 2.46) ($p<0.01$). The difference becomes smaller but still significant for the middle age group (RR: 1.18 (1.07, 1.3) ($p<0.01$)), while for the elderly we found out that the two sexes get infected similarly in a 5% significance level (RR: 1.09 (0.98, 1.22) ($p=0.11$)).

Regarding the vaccine efficacy, we concluded that the relative risk of the vaccinated group to get infected by a vaccine serotype in the post-vaccine period was RR: 0.3 (0.16, 0.56) ($p<0.01$) compared to the period before the vaccine, while no difference was noted in the vaccine effect among the two sexes. A point estimate of 70% less incidence of vaccine serotypes ranging between 44% and 84%, a large interval due to the small number of cases, indicates the strong effect of the vaccine. The same relative risk, but for the whole population this time, as found by the meta-analysis was RR: 0.75 (0.66, 0.84) ($p<0.01$), a highly significant decrease revealing the herd immunity effect.

As a consequence of the above, the non-vaccine serotypes became the main reason of IPD in the post-vaccine period, but this should not at any case be confused

as serotype replacement. The only age group, where the non-vaccine serotypes increased their incidence significantly between the two periods, was that of the elderly, while in the other age groups their incidence remained stable in total. After checking for each non-vaccine serotype individually we found out that there were some serotypes in the age group "18-64" that had actually increased their incidence significantly, but this change was not obvious when looking at them as a group.

A better understanding of the way the non-vaccine serotypes changed their incidence is given by the meta-analysis graph provided earlier in the study. The obvious shift of the majority of the serotypes towards the right side of the graph indicates their tendency to increase their incidence in the post-vaccine period and it is too apparent as a pattern to be attributed to seasonality. The serotypes that showed clear signs of replacement were *3, 7F, 19A, 22F* and *38*. The reasons that serotype replacement is not clearer, besides in the elderly group, are firstly, the short post-vaccine period, secondly, the initiation of the 13-valent vaccine in 2010 and finally, the "special" incidence of serotype *1*, which is dominant in the middle age group. Regarding the first, we should note that in the majority of relevant studies, a five year post-vaccine period was taken in order to have a clearer image, but unfortunately we had the limitation of a 3-year after period, so it is very likely that the phenomenon of replacement hasn't had the time to fully develop yet. As for the new vaccine, it could have influenced a bit our results of the last year, lowering up to a point the incidence of serotypes *1, 3, 5, 6A, 7F* and *19A*. Serotypes *22F* and *38*, hence, are the only two serotypes to have had significant increase of their incidence in the post-vaccine period that are also not included in the new vaccine. They were also the two serotypes with the highest relative risks, with the difference that *22F* is responsible for a larger number of cases, making it, thus, even more dangerous. If the new vaccine was composed in Stockholm, then the inclusion of at least this serotype would have been necessary in order to be controlled.

The natural question to pose at this point is "What needs to be done now"? The only thing necessary is that the continued evaluation of the incidence of the non-vaccine serotypes should be maintained, since it is very likely that some serotypes, even some novel ones, along with *22F* and *38* that we are already aware of, will take advantage of the new ecological niche created and become significantly more prevalent.

❖ 5

Appendix

5.1 Exponential family of distributions

Suppose we have a single random variable Y which has a probability distribution function depending on a parameter θ. If the distribution can be written in the following form:

$$f(y;\theta) = h(y)g(\theta)exp[a(y)b(\theta)] \tag{5.1}$$

then we say that it belongs to the exponential distribution family where of course we assume that all the functions a, b, g, h are known. We should mention that $a(y)$ is referred to as the canonical statistic, $b(\theta)$ is called the natural parameter and in the case where $a(y) = y$, the distribution is considered to obtain the canonical form. There are many distributions that belong in the exponential family, like the Poisson, Gamma, Normal and Binomial distribution. In some cases, as in the Normal and Gamma distribution, there is the need for one more parameter to be included. For these situations the general form of exponential distribution, assuming they are in canonical form, becomes:

$$f(y_i;\theta_i,\phi) = \exp\frac{y_i\theta_i - c(\theta_i)}{d(\phi)}h(y_i;\phi) \tag{5.2}$$

and the joint distribution is derived by multiplying the above for all the N observations $(y_1, y_2, \ldots, y_n)$ that we have of Y, assuming that they are independent. The above distribution form is called the *exponential dispersion family* and ϕ is known as the dispersion parameter. It is easily shown that for a known value of ϕ the dispersion exponential distribution and the "classic" form given in the very beginning become identical. This is done by setting $g(\theta) = exp[-c(\theta)/d(\phi)]$ and also $b(\theta)$ equal to $\theta/d(\phi)$. The term $d(\phi)$ is in general of the form $d(\phi) = \phi/\omega_i$, where the weights in the denominator are considered as known.

5.2 Generalized Linear Models (GLM)

We consider Generalized linear models (GLM) as an extension of the linear regression models that are widely known, where the response variable follows a Normal distribution and has a strictly linear relationship with the explanatory variables. Generalized linear models have the following general structure:

$$g(\mu_i) = x_i^T \beta \tag{5.3}$$

and consist actually of three parts. A random component, which is a set of independent random variables $Y_1, Y_2, \ldots, Y_n$ from an exponential distribution family where $E(Y_i) = \mu_i$, a linear predictor of the form $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, where $i = 1, 2, ..N$ is the number of observations and $j = 1, 2, \ldots p$ is the number of parameters in the model and finally a link function that connects the linear predictor with the mean value $\eta_i = g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j$. To make things clearer we should say that β is the vector of parameters $(\beta_1, \beta_2, \ldots, \beta_p)^T$, $\mathbf{X}$ is the so-called design matrix of which x_i^T is the i'th row and g is a monotonic and differentiable function. Regarding the link function we especially distinguish two cases. The first is the identity link where $g(\mu) = \mu$, which leads us back to the ordinary linear model if our response variable is normally distributed and the second is the canonical link in which the link function equals the natural parameter $g(\mu_i) = b(\theta_i)$.

5.3 Likelihood equations of GLM

Before getting down showing how the likelihood equations can be derived it is useful to present the general form of the mean and variance in a GLM. By using the exponential dispersion model defined previously the log-likelihood of one observation would be:

$$f(y_i; \theta_i, \phi) = exp\frac{y_i\theta_i - c(\theta_i)}{d(\phi)}h(y_i; \phi) \tag{5.4}$$

If we differentiate the above with respect to θ_i we get the score function U. The mean value of the score function is known to be zero and from that we get: $\mu_i = E(Y_i) = c\prime(\theta_i)$. By differentiating once more and using the property $E(U^2) = -E(U\prime)$ we get: $var(Y_i) = c\prime(\theta_i)d(\phi)$. We can now obtain the likelihood equations which are: $\sum_{i=1}^{N} \frac{\partial Ł_i}{\partial \beta_j} = 0$ and in order to calculate them we will need to use the chain rule

$$\frac{\partial Ł_i}{\partial \beta_j} = \frac{\partial Ł_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{5.5}$$

Using the relations connecting the above variables, as they have been given throughout the appendix, we conclude that the likelihood equations are

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \tag{5.6}$$

where the term $\frac{\partial \mu_i}{\partial \eta_i}$ can be calculated as long as we know the link function. The above equations are usually non-linear and the use of an iterative method, such as Newton-Raphson or Fisher scoring method is necessary for solutions to be obtained. One important conclusion that should be pointed out is that the likelihood equations depend on the distribution of Y_i only through the mean μ_i and the variance of it $var(Y_i)$, where actually the variance itself is a function of the mean.

5.4 Results and graphs of the Poisson model

Table 5.1. Estimates and characteristics of the fitted Poisson model.

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-8.54709	0.059	-143.161	$<$ 2e-16 ***
PERIOD1	-0.25319	0.09750	-2.597	0.00941 **
SEROTYPE1	-0.04898	0.07732	-0.634	0.52640
AGE0	-0.06949	0.12845	-0.541	0.58851
AGE1	-1.53434	0.07283	-21.068	$<$ 2e-16 ***
SEXF	-0.01365	0.06739	-0.203	0.83945
PERIOD1:SEROTYPE1	0.52987	0.12973	4.084	4.42e-05 ***
PERIOD1:AGE0	-1.04617	0.34584	-3.025	0.00249 **
PERIOD1:AGE1	0.03658	0.13207	0.277	0.78182
SEROTYPE1:AGE0	-0.54305	0.18415	-2.949	0.00319 **
SEROTYPE1:AGE1	0.22316	0.08639	2.583	0.00979 **
SEROTYPE1:SEXF	-0.15195	0.07250	-2.096	0.03610 *
AGE0:SEXF	-0.52080	0.16707	-3.117	0.00183 **
AGE1:SEXF	-0.06925	0.07503	-0.923	0.35599
PERIOD1:SEROTYPE1:AGE0	0.75538	0.44586	1.694	0.09023 .
PERIOD1:SEROTYPE1:AGE1	-0.29199	0.17525	-1.665	0.09569 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 1500.0339 on 23 degrees of freedom
Residual deviance: 7.8479 on 8 degrees of freedom
AIC: 185.92

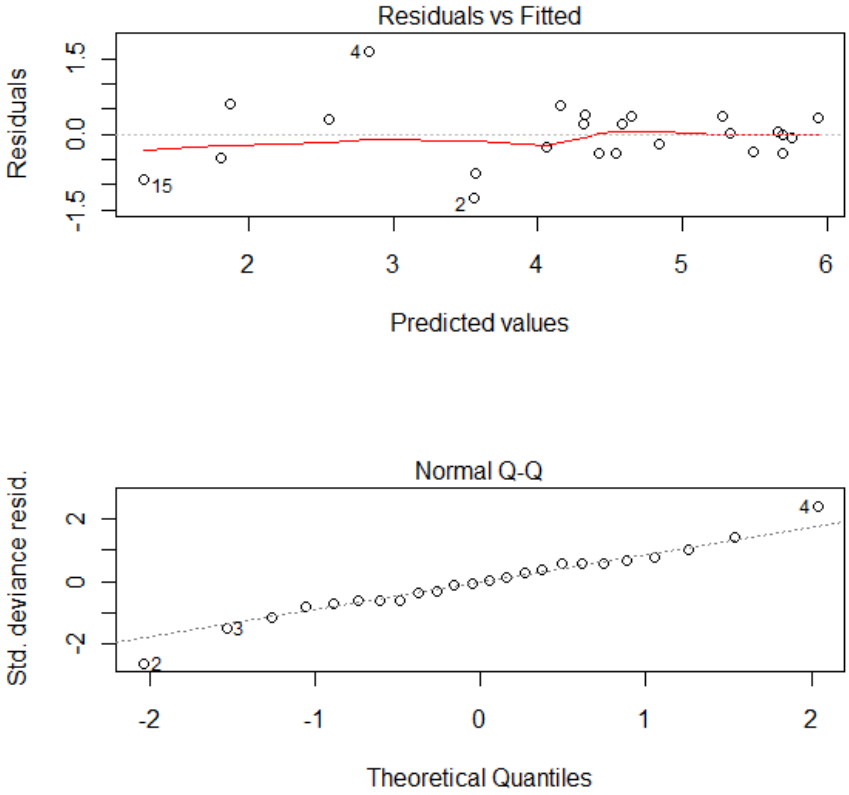

Fig. 5.1. The residuals plotted against the fitted values on a log scale and the quantile-quantile plot.

5.5 Distribution of serotypes per age group

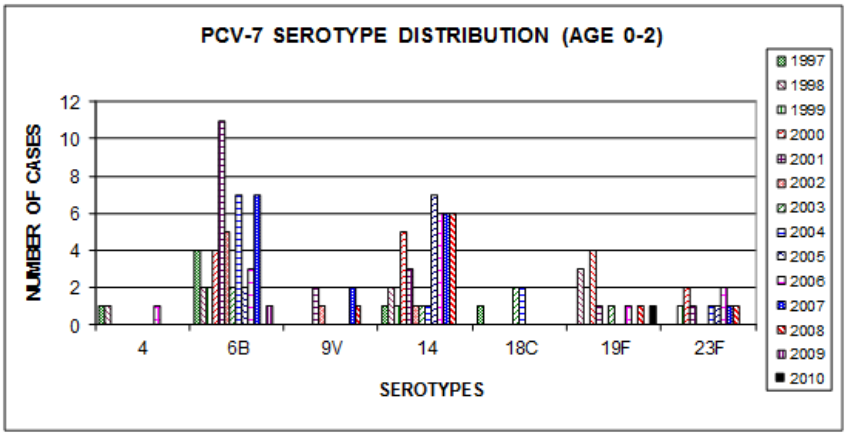

Fig. 5.2. The distribution of the seven vaccine serotypes (PCV7) for the age group 0-2 years, through the period 1997-2010.

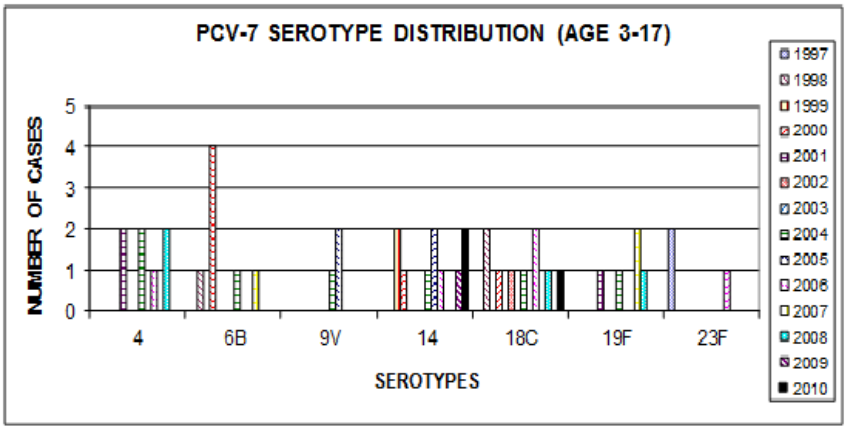

Fig. 5.3. The distribution of the seven vaccine serotypes (PCV7) for the age group 3-17 years, through the period 1997-2010.

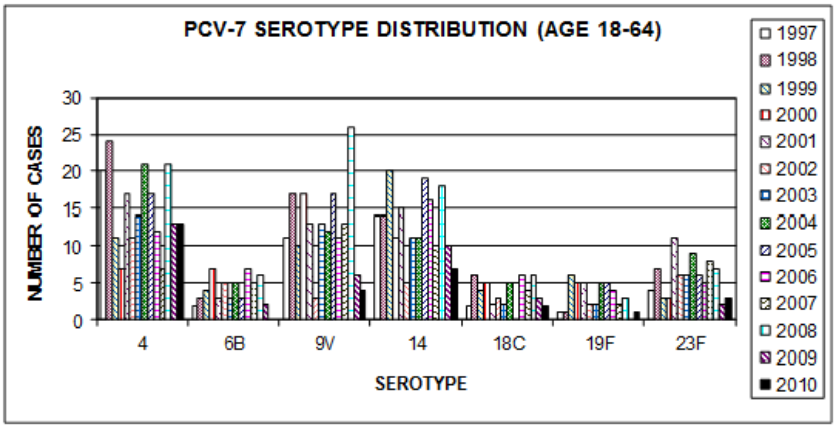

Fig. 5.4. The distribution of the seven vaccine serotypes (PCV7) for the age group 18-64 years, through the period 1997-2010.

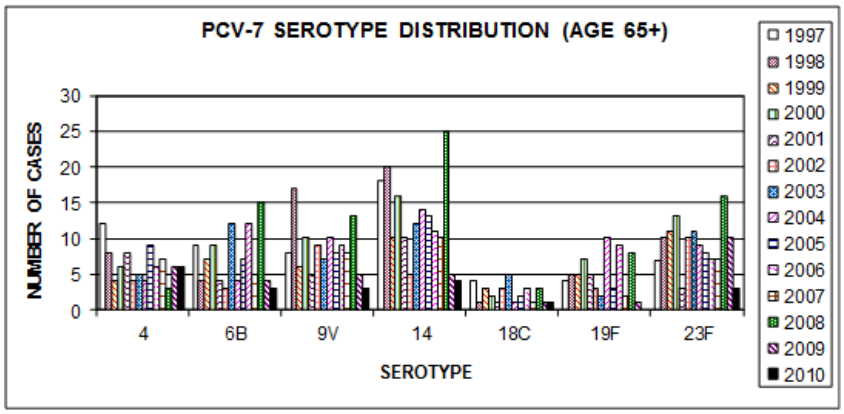

Fig. 5.5. The distribution of the seven vaccine serotypes (PCV7) for people aged over 65 years, through the period 1997-2010.

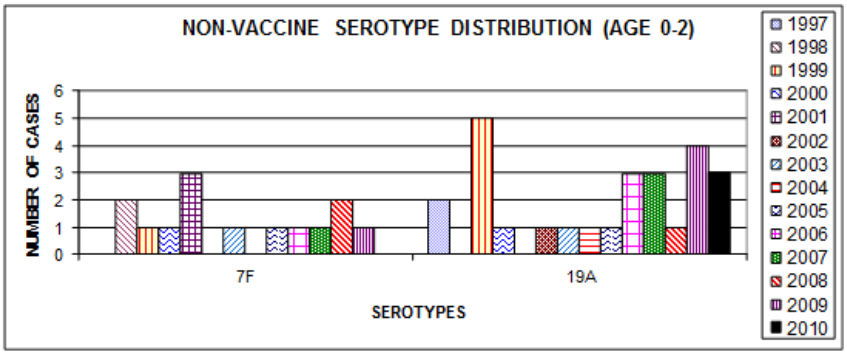

Fig. 5.6. The distribution of the non-vaccine serotypes (Non-PCV7) for the age group 0-2 years, through the period 1997-2010.

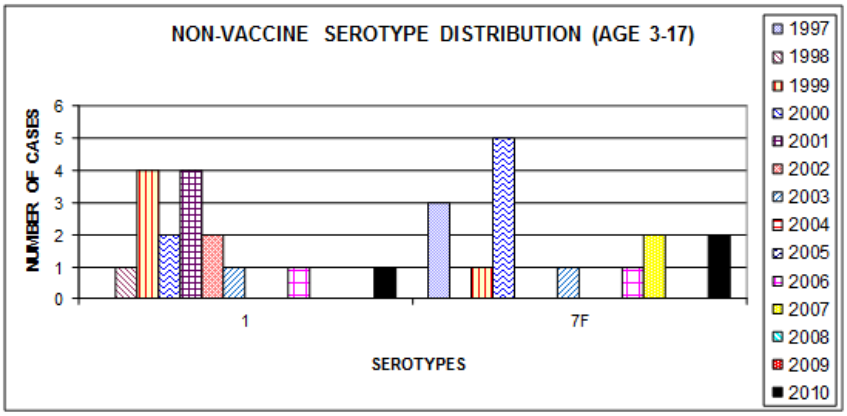

Fig. 5.7. The distribution of the non-vaccine serotypes (Non-PCV7) for the age group 3-17 years, through the period 1997-2010.

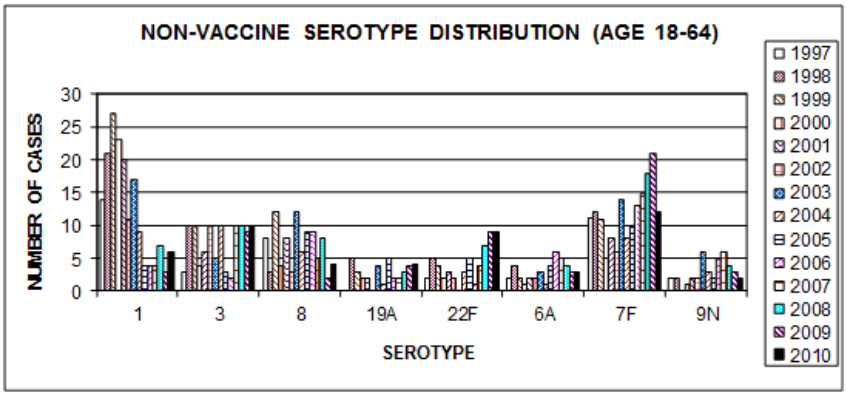

Fig. 5.8. The distribution of the non-vaccine serotypes (Non-PCV7) for the age group 18-64 years, through the period 1997-2010.

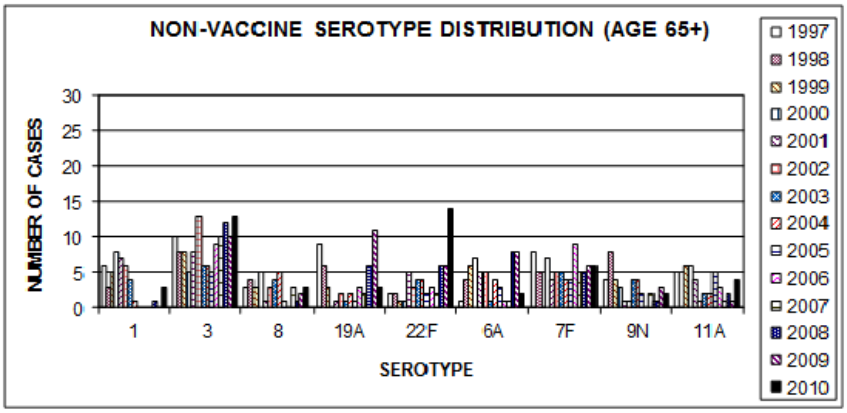

Fig. 5.9. The distribution of the non-vaccine serotypes (Non-PCV7) for people aged over 65 years, through the period 1997-2010.

❖ 6

Bibliography

[1] Hicks, Lauri A./Harrison, Lee H./Flannery, Brendan et al. (2007): “Incidence of pneumococcal disease due to non-pneumococcal conjugate vaccine (PCV7) serotypes in the United States during the era of widespread PCV7 vaccination, 1998-2004”. In:*The Journal of Infectious Diseases* , Oct. 2007, Vol: 196: p. 1346-54.

[2] Henriques Normark, B./Sandgren, A./Kronvall, G. et al. (2004): “Effect of clonal and serotype-specific properties on the invasive capacity of Streptococcus Pneumoniae”.”. *In: The Journal of Infectious Diseases*,Feb. 2004, Vol: 189: p. 785-96.

[3] Lipsitch, Marc (1999): “Bacterial Vaccines and Serotype Replacement: Lessons from Haemophilus influenzae and Prospects for Streptococcus pneumoniae”. In: *Emerging Infectious Diseases*,May 1999, Vol: 5: p. 336-45.

[4] Albrich, W. C./Baughman, W. /Schmotzer, B./Farley, M. M. (2007): “Changing characteristics of invasive pneumococcal disease in Metropolitan Atlanta, Georgia, after introduction of a 7-valent pneumococcal conjugate vaccine”. In: *Clinical Infectious Diseases*,May 2007, Vol: 44: p. 1569-76.

[5] Obaro, Steven K. (2000): “Confronting the pneumococcus: a target shift or bullet change?”. In: *Medical Reasearch Council Laboratories*,Vol: 19: p.1211-17.

[6] Lehmann, Deborah/Willis, Judith/Moore, Hannah C. et al. (2010): “The changing epidemiology of invasive pneumococcal disease in aboriginal and non-aboriginal

western Australians from 1997 through 2007 and emergence of nonvaccine serotypes". In: *Clinical Infectious Diseases*, Apr.2010, Vol: 50: p.1477-86.

[7] Swedish Institute for Infectious Disease Control ("http://www.smi.se/")

[8] Brueggemann, Angela B./Griffiths, David T./Meats, Emma et al (2003): " Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential". In: *The Journal of Infectious Diseases*, Apr. 2003, Vol: 187: p. 1424-32.

[9] Lipsitch, Marc (1997): " Vaccination against colonizing bacteria with multiple serotypes". In: *Proceedings of the National Academy of Sciences of the United States of America* , June 1997, Vol: 94: p.6571-76.

[10] Reingold, A./Hadler, J./Farley, MM et al. (2005): "Direct and Indirect Effects of Routine Vaccination of Children with 7-Valent Pneumococcal Conjugate Vaccine on Incidence of Invasive Pneumococcal Disease - United States, 1998-2003". In: *The Journal of the American Medical Association*, Sept. 2005, Vol: 54: p. 893-7.

[11] Huang, Susan S./Platt, Richard /Rifas-Shiman, Sheryl L. et al. (2005): " Post-PCV7 Changes in Colonizing Pneumococcal Serotypes in 16 Massachusetts Communities, 2001 and 2004" . In: *Pediatrics*,Vol: 116: p.408-13

[12] Kellner, James D./Vanderkooi, Otto G./MacDonald, Judy et al. (2009): "Changing Epidemiology of Invasive Pneumococcal Disease in Canada, 1998–2007: Update from the Calgary-Area Streptococcus pneumoniae Research (CASPER) Study". In: *Clinical Infectious Diseases*, June 2009, Vol:49: p.205-12.

[13] Nunes, S./Paulo, A.C.S/Saldanha, J. et al. (2009): "Changes in pneumococcal serotypes and antibiotypes carried by vaccinated and unvaccinated day-care centre attendees in Portugal, a country with widespread use of the seven-valent pneumococcal conjugate vaccine". In: *European Society of Clinical Microbiology and Infectious Diseases* , Apr. 2009, Vol:15: p.1002-1007.

[14] Reingold, A./Hadler, J./Farley, MM et al. (2008): “Invasive Pneumococcal Disease in Children 5 Years After Conjugate Vaccine Introduction – Eight States, 1998-2005”. In: *The Journal of the American Medical Association* , Feb. 2008, Vol: 57: p. 144-148.

[15] Statistics Sweden (“http://www.scb.se”)

[16] Brueggemann, Angela B./Spratt, Brian G. (2003): “Geographic Distribution and Clonal Diversity of Streptococcus pneumoniae Serotype 1 Isolates”. In: *Journal of Clinical Microbiology* , Nov. 2003, p. 4966–4970.

[17] Nuorti, J.Pekka/Butler Jay C./Farley Monica M. et al. (2000):”Cigarette smoking and invasive pneumococcal disease”. In: *The New England Journal of Medicine* , Mar. 2000, Vol: 342: p. 681-9.

[18]Everitt, Brian S./Hothorn, Torsten (2009): *A handbook of statistical analyses using R* Second Edition, CRC Press.

[19] Sundberg, R. (2010): *Lecture notes on statistical modelling by exponential families*. Stockholm University.

[20] Dobson, Annette J./Barnett, Adrian G. (2001): *An introduction to generalized linear models.*Third edition, CRC Press.

[21] Agresti Alan (2002): *Categorical data analysis*. Second edition, John Wiley and Sons.