An overview of statistical methods for active pharmacovigilance with applications to diabetes patients

By

Lan Zhuo

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree of

Master of Science

Ottawa-Carleton Institute for mathematics and Statistics

Carleton University

1125 Colonel By Drive, Ottawa, Ontario

Canada K1S 5B6

©Zhuo, Lan 2010

December, 2010

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-79589-7
Our file Notre référence
ISBN: 978-0-494-79589-7

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

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

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux 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.

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

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

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

Canada

Abstract

The primary goal of active pharmacovigilance is to detect the association between certain drugs and particular adverse drug reactions to these drugs through cohort data. Several statistical methods, specifically the logistic regression model, the logistic regression model with James-Stein shrinkage, the Cox model, and the random effects Cox model have been proposed to investigate drug-event association. In this thesis, for each method, we describe the underlying model, the estimation techniques, as well as their properties. We also apply these four models to a diabetes data set, which is extracted from a cohort database, in order to analyze the association between particular drugs of interest (Actos, Avandia, Metformin, Insulin, and Sulfonylurea) and certain adverse drug reactions (heart failure and acute myocardial infarction). We also consider the effects of age, gender, time since first exposure to a drug, and cumulative dose.

Acknowledgements

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I am deeply indebted to my supervisors Dr. Patrick Farrell from Carleton University and Dr. Daniel Krewski from the University of Ottawa, whose help, stimulating suggestions and encouragement helped me throughout my studies.

Also, I would like to thank the School of Mathematics and Statistics at Carleton University for training and teaching me the skills that I needed to complete this thesis.

Finally, I would like to give my special thanks to my parents Dehua Zhuo and Juxiang Yu, and my sister Chaowei Zhuo, whose patient love and encouragement enabled me to complete this thesis.

Table of Contents

Chapter 1
Introduction

Since drug safety is one of the greatest concerns within the pharmaceutical community, pre-marketing safety testing is applied before prescription drugs are approved into the marketplace. However some adverse drug reactions (ADRs) that were not identified in pre-marketing safety testing may become apparent once the drug is introduced into the marketplace. Then post-marketing safety testing becomes an important component of the overall drug safety evaluation process. The World Health Organization (WHO) defines pharmacovigilance as the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem.

Pharmacovigilance aims at achieving the best treatment outcome with medication. No one wants to harm patients, but sometimes a medication can do this for many different factors. Once a medicine has been introduced into the marketplace, we want to determine what risks and adverse reactions were not discovered in pre-marketing testing. This will help each patient to receive optimum therapy at a lower cost. As a result, good pharmacovigilance will identify these risks in the shortest possible time after the medicine has been marketed and will help to establish risk factors.

Pharmacovigilance can be classified into two categories by the form of surveillance: passive pharmacovigilance and active pharmarcovigilance. Passive surveillance means that no active measures are taken to look for adverse effects other than the encouragement of health professionals and others to report safety concerns. Reporting is entirely dependent on the initiative and motivation of the potential reporters. It is referred to as spontaneous reporting. This is the most common form of pharmacovigilance, and this form of reporting is mandatory in some countries. Active surveillance means that active measures are taken to detect adverse events. This is managed by active follow-up after treatment by which events may be detected by asking patients directly or screening patient records. The most comprehensive method is cohort event monitoring (CEM).

A spontaneous reporting system (SRS) is a voluntary, passive pharmacovigilance surveillance system that collects reports of suspected ADRs from both health care professional and consumers. The most popular data mining methods to analyze the SRS datasets are the so called disproportionality-based methods, which include the frequentist approach and the Bayesian approach. The frequentist methods include the Proportional Reporting Ratio (PRR), the Reporting Odds Ratio (ROR) and Relative Reporting Ratios (RRR). Two common Bayesian approaches are the Bayesian Confidence Propogation Neural Network (BCPNN) and the empirical Bayes Screening (EBS). A detailed reviewed of these disproportionality-based methods has been conducted by Gravel (2009).

The disproportionality-based methods listed above for passive pharmacovigilance suffer from two clear limitations. First of all, they are unable to correct for confounding when multiple medications are taken by the same patient. This means, for example, if one drug A truly causes some ADR and another drug B is frequently used together with drug A, then these passive methods will signal that both drugs are the likely cause of the ADR. In this situation it can be said that there is "signal leakage" from drug A to drug B, and that drug B is referred to as an "innocent bystander". The second problem is "masking", which is related to the background reporting rate. When studying a particular drug-ADR combination, the disproportionality-based methods assume that all reports except for the drug of interest constitute the general background reporting of the ADR. However, if there are one or more drugs which occur frequently in these reports, the background reporting becomes substantial, which increases the expected number of reports of the drug-ADR combination of interest. Thus, when the observed number of reports is compared to the expected number, the false conclusion could be drawn that this drug-ADR combination should not be highlighted as a signal. In addition to these two limitations, another concern regarding passive pharmacovigilance is that since the data arise from an SRS, there will be limited information on such characteristics as medication dosage, exposure time, etc.

By contrast, in active pharmacovigilance, once a drug has been introduced into the market, active measures are taken to detect adverse reactions. In this thesis, we will

describe and discuss the statistical methods for active pharmacovigilance. Four different statistical methods will be discussed here; namely logistic regression modeling, logistic regression models with James-Stein shrinkage estimation, Cox models and Cox models with random effects. We will apply these methods to a diabetes dataset in order to investigate the effects of various drugs on heart failure and acute myocardial infarction.

In all the above statistical methods for active pharmacovigilance, the value of some binary dependent variable is described by a set of predictor variables, and the respective degrees of contribution of each of the variables is estimated. Specifically, an indicator variable indicating the presence or absence of a particular ADR on a given patient is treated as the dependent variable. The set of predictor variables includes indicator variables for the usage/non-usage of the group of drugs of interest. This avoids the problem encountered in passive pharmacovigilance of confounding due to a patient taking multiple medications. Furthermore, because all the models considered here include an intercept term which is a function of the background reporting rate, these approaches could in theory be expected to avoid the problems associated with masking. Finally, unlike an SRS dataset, additional variables such as medication dose, exposure time etc. can be simultaneously included in the analysis in active pharmacovigilance.

This thesis is organized as follows. In Chapter 2, we describe and discuss the

statistical methods for active pharmacovigilance. In Chapter 3, we illustrate these methods on a diabetes dataset from Cerner's HealthFacts™ Datawarehouse in order to study the effects of different drugs on the prevalence of heart failure and acute myocardial infarction. Finally, we provide conclusions and discussion in Chapter 4.

Chapter 2
Methods for active pharmacovigilance

The primary task of mining adverse drug reaction databases in active pharmacovigilance is to detect "signals", i.e. to identify drug-event associations worthy of further investigation. Logistic regression methods have become an integral component of the analysis. These approaches are concerned with describing the relationship between a response variable (presence or absence of an ADR in our case) and one or more continuous or discrete explanatory variables. The outcome variable in such a model is discrete, taking on two possible values. The logistic regression model has become, in many fields, one of the standard methods of analysis in this situation. If many similar adverse drug reactions need to be modeled simultaneously, James-Stein type shrinkage estimation can be incorporated with the logistic regression model. Such shrinkage estimation is effective in detecting signals, as it combines information and borrows strength across medically-related adverse drug reactions. Signal detection can also be treated as a survival problem if survival times are available, which is the case with the data we are to consider here. In this situation, it is possible to use the Cox model to explore the relationship between the survival experience of a patient and explanatory variables. Moreover, if patients are grouped by hospital (also the case for the data to be studied), say, it may be useful to study Cox models with

random effects.

In this chapter, we discuss these four methods in turn, showing how signal detection can be formulated into a test of hypothesis for each approach.

In what follows, we assume that the available data contains a record of adverse drug reactions for each of n patients, where each record indicates which drugs the patient has taken, whether or not the patient experienced particular adverse reactions with each of drugs, along with supplementary information such as the age of patient, gender etc. For notational purposes, we shall specify that there are d different drugs and k distinct adverse drug reactions over the entire data set. Finally, although it is already possible for patients to be prescribed more than one drug, we shall restrict our attention here to those who only received a single drug from the group under study.

2.1 Logistic Regression Models

Logistic regression can be used to model each of the k adverse drug reactions separately. Let $Y_i = 1$ denote that the adverse drug reaction of interest is present in the record of the i-th patient and $Y_i = 0$ denote that the adverse drug reaction of interest is not present. In addition, define $\pi_i = \pi(\mathrm{x}_i) = P(Y_i = 1)$, where for the i-th record, $\mathbf{x}_i$ is a vector of values for the explanatory variables augmented by the constant one, so that

$$\mathbf{x}_i = (1, x_{i1}, x_{i2}, \cdots, x_{ip})^T \qquad (1)$$

Provided that we introduce a parameter vector

$$\boldsymbol{\beta}^{\mathrm{T}} = (\beta_0, \beta_1, \cdots, \beta_p)$$

the ordinary logistic regression model is of the form

$$\pi_i = \pi(\mathbf{x_i}) = P\left(y_i = 1 \mid \mathbf{x_i^T}, \boldsymbol{\beta}\right) = \frac{\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x_i}\right)}{1 + \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x_i}\right)} \tag{2}$$

or

$$\mathrm{z_i} = \log\frac{P(y_i = 1)}{1 - P(y_i = 1)} = \mathrm{logit}[\pi(\mathbf{x_i})] = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x_i} \tag{3}$$

To derive maximum likelihood estimates for multiple logistic regression models, we begin by considering the distribution of the data, given by

$$\prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}$$

Since Y_i is a Bernoulli random variable, and all Y_i are assumed to be independent, the log-likelihood of the data is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi_i) \tag{4}$$

The maximum likelihood estimator (MLE) is the $\widehat{\boldsymbol{\beta}}$ that maximizes this log-likelihood. We use the Newton-Raphson algorithm to find the MLE of $\boldsymbol{\beta}$.

The Newton-Raphson method, named after Isaac Newton and Joseph Raphson, is perhaps the best known method for finding successively better approximations to the zeroes of a real-valued function. This method can often converge remarkably quickly, especially if the iteration begins 'sufficiently near' the desired root. It can also be used to find the minimum or maximum of a function.

For the log-likelihood given in (4) above, the maximum likelihood equations are

$$\partial l(\boldsymbol{\beta})/\partial \beta_0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i = 0$$

and

$$\partial l(\boldsymbol{\beta})/\partial \beta_j = \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \pi_i = 0 \qquad j = 1,\ldots,p$$

These equations can be solved to obtain an estimate for $\boldsymbol{\beta}$ by using Newton-Raphson algorithm. We let

$$\mathbf{q}' = \left(\partial l(\boldsymbol{\beta})/\partial \beta_0, \partial l(\boldsymbol{\beta})/\partial \beta_1, \ldots, \partial l(\boldsymbol{\beta})/\partial \beta_p\right)$$

$$= \left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi_i, \sum_{i=1}^{n} x_{i1} y_i - \sum_{i=1}^{n} x_{i1} \pi_i, \cdots, \sum_{i=1}^{n} x_{ip} y_i - \sum_{i=1}^{n} x_{ip} \pi_i\right)$$

and

$$\mathbf{H} = \begin{bmatrix} \partial^2 l(\boldsymbol{\beta})/\partial \beta_0^2 & \partial^2 l(\boldsymbol{\beta})/\partial \beta_0 \partial \beta_1 & \ldots & \partial^2 l(\boldsymbol{\beta})/\partial \beta_0 \partial \beta_p \\ \partial^2 l(\boldsymbol{\beta})/\partial \beta_0 \partial \beta_1 & \partial^2 l(\boldsymbol{\beta})/\partial \beta_1^2 & \ldots & \partial^2 l(\boldsymbol{\beta})/\partial \beta_1 \partial \beta_p \\ \vdots & \vdots & & \vdots \\ \partial^2 l(\boldsymbol{\beta})/\partial \beta_0 \partial \beta_p & \partial^2 l(\boldsymbol{\beta})/\partial \beta_1 \partial \beta_p & \ldots & \partial^2 l(\boldsymbol{\beta})/\partial \beta_p^2 \end{bmatrix}$$

so that

$$\mathbf{H} = \begin{bmatrix} -\sum_{i=1}^{n} \pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & \ldots & -\sum_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) \\ -\sum_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_{i1}^2\pi_i(1-\pi_i) & \ldots & -\sum_{i=1}^{n} x_{i1}x_{ip}\pi_i(1-\pi_i) \\ \vdots & \vdots & & \vdots \\ -\sum_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) & -\sum_{i=1}^{n} x_{i1}x_{ip}\pi_i(1-\pi_i) & \ldots & -\sum_{i=1}^{n} x_{ip}^2\pi_i(1-\pi_i) \end{bmatrix}$$

To estimate $\boldsymbol{\beta}$, we start with an initial guess, say $\hat{\boldsymbol{\beta}}^{(0)}$ and perform an iterative procedure. At the *m*-th step of this iterative process we obtain $\hat{\boldsymbol{\beta}}^{(m+1)}$ using

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} - (\hat{\mathbf{H}}^{(m)})^{-1}\hat{\mathbf{q}}^{(m)} \qquad (5)$$

where $\hat{\mathbf{q}}^{(m)}$ and $\hat{\mathbf{H}}^{(m)}$ are equivalent to $\mathbf{q}$ and $\mathbf{H}$ evaluated at $\hat{\boldsymbol{\beta}}^{(m)}$, the *m*-th guess for $\boldsymbol{\beta}$. The algorithm continues until successive estimates of $\boldsymbol{\beta}$ converge.

If the algorithm converges at iteration M, then $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(M)}$ is the maximum likelihood estimate for $\boldsymbol{\beta}$, and we can use it to determine estimates for $\pi_i = \pi(\mathbf{x_i})$ according to

$$\hat{\pi}_i = \hat{\pi}(\mathbf{x_i}) = \frac{\exp(\mathbf{x_i^T}\widehat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x_i^T}\widehat{\boldsymbol{\beta}})} \tag{6}$$

In addition $-\mathbf{H}^{-1}$ when evaluated at $\widehat{\boldsymbol{\beta}}$ serves as an estimated asymptotic covariance matrix.

The main advantages of the logistic regression model are that it enjoys a body of supporting theory and algorithms, features prominently in commercial statistical software, and its predictive accuracy is often competitive. However, there are some limitations as well. First of all, for the 'short' and 'fat' data set, which means that the number of predictor variables is large and usually exceeds the number of observations and the sparse data set, computing the maximum likelihood fit of a logistic regression model is often impossible since standard software relies on matrix inversion. Even when this barrier is overcome, numerical ill-conditioning can result in a lack of convergence, large estimated coefficient variances, poor predictive accuracy, and/or reduced power for testing hypotheses concerning model assessment (Pike, Hill, and Smith 1980).

2.2 James-Stein Shrinkage in Logistic Regression Models

In many applications, there may be uncertain prior information available about the parameters in the statistical model used to describe the available data. Since the

validity of the prior assumption is not tested, neither the pooled nor unrestricted estimators make use of the available information in an optimal way. The James-Stein type shrinkage estimator incorporates this uncertain prior information, and combines the restricted and unrestricted estimators in a superior manner.

Shrinkage estimation was first proposed by Stein (1956) and James & Stein (1961). Since then, Ahmed & Saleh (1999), Casella & Hwang (1986), An et al. (2006) have conducted considerable research on shrinkage estimation of location parameters, while Ahmed and Krzanowski (2004) considered the estimation of the intercept vector in regression models; An et al. (2010) consider shrinkage estimation in logistic regression models, while An et al. (2009) discuss the more general case of generalized linear models.

In the context of our example, for a given set of drugs, James-Stein shrinkage estimation allows for different ADRs to be studied simultaneously. In order to describe the approach, we assume that the available data contain a total of n adverse drug reactions records, where each report involves one drug and at least one adverse drug reaction. In addition, we assume that there are d different drugs and k distinct adverse drug reactions across the entire data set. Following An et al. (2010), we then consider k logistic regression models, one for each drug, of the form given in equation (3). The l-th model $(l = 1, \cdots, k)$ will be based on a data set of size n_l. The k models can be combined into one single model, with the general form

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7}$$

where $\mathbf{z} = (z_1, z_2, \cdots, z_n)^T$ is a vector of length n with $z_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, $n = \sum_{i=1}^{k} n_l, (l = 1,2,\cdots,k)$; $\mathbf{X}$ is an $n \times qk$ design matrix where $q = p + 1$, and p represents the total number of difference variables, including indicator variables and other explanatory variables such as age and gender in the model; $\boldsymbol{\beta} = \left(\beta_{01}, \beta_{02}, \cdots, \beta_{0k}, \beta_{11}, \beta_{12}, \cdots, \beta_{1k}, \cdots, \beta_{p1}, \beta_{p2}, \cdots, \beta_{pk}\right)^T$ is a vector of length $q \times k$; and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ is the error term. We assume that $\boldsymbol{\varepsilon}$ has mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

The design matrix $\mathbf{X}$ is based on k smaller matrices, one for each adverse drug reaction of interest. The l-th of the matrices, $\mathbf{X}_l$ say, is given by

$$\mathbf{X}_l = \begin{bmatrix} 1 & x_{1,1,l} & \cdots & x_{1,p,l} \\ 1 & x_{2,1,l} & \cdots & x_{2,p,l} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n_l,1,l} & \cdots & x_{n_l,p,l} \end{bmatrix}$$

Appropriately combining these $\mathbf{X}_l$ matrices with zero entries allows us to specify the design matrix $\mathbf{X}$ as:

$\mathbf{X} =$

$$\begin{bmatrix}
1 & 0 & \cdots & 0 & x_{1,1,1} & 0 & \cdots & 0 & x_{1,2,1} & 0 & \cdots & 0 & \cdots\cdots & x_{1,p,1} & 0 & \cdots & 0 \\
\vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 & x_{n_1,1,1} & 0 & \cdots & 0 & x_{n_1,2,1} & 0 & \cdots & 0 & \cdots\cdots & x_{n_1,p,1} & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & 0 & x_{1,1,2} & \cdots & 0 & 0 & x_{1,2,2} & \cdots & 0 & \cdots\cdots & 0 & x_{1,p,2} & \cdots & 0 \\
\vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \vdots & \cdots & 0 \\
0 & 1 & \cdots & 0 & 0 & x_{n_2,1,2} & \cdots & 0 & 0 & x_{n_2,2,2} & \cdots & 0 & \cdots\cdots & 0 & x_{n_2,p,2} & \cdots & 0 \\
 & \vdots & & & & \vdots & & & \vdots & & & & \cdots\cdots & & \vdots & & \\
 & \vdots & & & & \vdots & & & \vdots & & & & & & \vdots & & \\
0 & 0 & \cdots & 1 & 0 & \cdots & 0 & x_{1,1,k} & 0 & \cdots & 0 & x_{1,2,k} & \cdots\cdots & 0 & 0 & \cdots & x_{1,p,k} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots\cdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & 0 & \cdots & 0 & x_{n_2,1,2} & 0 & \cdots & 0 & x_{n_k,2,k} & \cdots\cdots & 0 & 0 & \cdots & x_{n_k,p,k}
\end{bmatrix} \tag{8}$$

The MLE of $\boldsymbol{\beta}$ in equation (7) can be obtained by the Newton-Raphson algorithm described above.

If it is suspected that some of the parameters may be restricted to a particular subspace for $\boldsymbol{\beta}$, we can test the hypothesis

$$H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \qquad \textbf{vs} \qquad H_A: \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d} \tag{9}$$

For example, one hypothesis of particular interest is whether certain drugs may increase the risk of several or all of the adverse drug reactions according to a similar biological mechanism. For this hypothesis, the $\mathbf{C}$ matrix would have $r = (k-1) \times p$ rows and $q \times k$ columns, and $\mathbf{d}$ would be a vector of length r, where $\boldsymbol{d} = (0,0,\cdots,0)^T$. For the case where $p = 2$, the $\mathbf{C}$ matrix is,

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & -1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & -1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & -1 \end{bmatrix}$$

In what follows, we describe the approach for conducting the hypothesis test in (9). First, denote the unrestricted maximum likelihood estimate of $\boldsymbol{\beta}$ in equation (7) by $\widehat{\boldsymbol{\beta}}^{un}$, and the maximum likelihood estimate obtained for $\boldsymbol{\beta}$ under H_0 in equation (9) by $\widetilde{\boldsymbol{\beta}}^{re}$. If H_0 is true, we can combine information that borrows strength across all these related adverse drug reactions. Then, the restricted estimator $\widetilde{\boldsymbol{\beta}}^{re}$ is computed by

$$\widetilde{\boldsymbol{\beta}}^{re} = \widehat{\boldsymbol{\beta}}^{un} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}\left(\mathbf{C}\widehat{\boldsymbol{\beta}}^{un} - \mathbf{d}\right) \tag{10}$$

If the null hypothesis is true, the restricted estimator $\widetilde{\boldsymbol{\beta}}^{re}$ is expected to be better than $\widehat{\boldsymbol{\beta}}^{un}$ since it pools information and borrows strengths across the k data sets. However, if the null hypothesis is not true, the restricted estimator may lead to higher quadratic risk due to the bias. Alternatively, we can use the James-Stein type shrinkage method to combine the restricted and unrestricted estimators in an optimal way to achieve an improved estimator of the model parameters. Note that An, Fung, et al. (2009) showed that the James-Stein type shrinkage estimator is uniformly better than the maximum likelihood estimator in terms of mean squared error in generalized linear models. Before introducing the James-Stein estimator, we need first to consider a test statistic for the hypothesis in equation (9).

In order to test the null hypothesis in (9), the F test statistic is given by

$$F_{(r,m)} = \frac{\left[\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{d}\right]^{\mathrm{T}}\left[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}\right]^{-1}\left[\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{d}\right]}{r s_e^2} \tag{11}$$

where $m = n - qk$ and s_e^2 is the mean square error under the full model. Under H_0, $F_{(r,m)}$ has a central F-distribution with (r, m) degrees of freedom, while under H_A, $F_{(r,m)}$ has a noncentral F-distribution with (r, m) degrees of freedom and noncentrality parameter $\Delta^2/2$ (Saleh, 2004), where

$$\Delta^2 = \frac{\left[\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{d}\right]'\left[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\right]^{-1}\left[\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{d}\right]}{\sigma^2} \tag{12}$$

We can compute s_e^2 by using the Pearson's chi-square statistic (see McCullagh, 1983) as follows,

$$s_e^2 = \frac{1}{n-qk}\sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2/[\hat{\pi}_i(1-\hat{\pi}_i)] \quad (13)$$

Then the James-Stein estimator is given by

$$\hat{\boldsymbol{\beta}}^{JS} = \hat{\boldsymbol{\beta}}^{re} + \left(1 - \frac{c}{F}\right) I(F \geq c)\left(\hat{\boldsymbol{\beta}}^{un} - \tilde{\boldsymbol{\beta}}^{re}\right) \quad (14)$$

where $c = \frac{(r-2)m}{r(m+2)}$, and $I(F \geq c)$ is an indicator function. Hence the James-Stein estimator is a weighted average of the unrestricted and restricted estimates. The weight constant $c = \frac{(r-2)m}{r(m+2)}$ minimizes the mean squared error of the estimator (Saleh, 2006).

The James-Stein estimator can test and incorporate uncertainty about any subset of the parameter space. The advantage of the James-Stein estimator is that it combines information and borrows strength across several data sets. In addition, it has lower quadratic risk than the maximum likelihood estimator. It also provides shorter confidence intervals than the maximum likelihood estimator, while maintaining a desired coverage probability. However, there are some disadvantages for James-Stein estimators. First, the formula to compute s_e^2 in equation (13) is more stable for grouped data. Secondly, when compared with logistic regression models, the asymptotic covariance matrix of the estimators of the model parameters cannot be obtained directly.

2.3 Cox Models

Survival analysis is concerned with longitudinal data from a time origin until the

occurrence of some particular event. The time origin will often correspond to the recruitment of an individual into an experimental study, and the end point is the death of the person or an analogous type of outcome such as the occurrence of an adverse event. Survival analysis focuses on the distribution of survival time. Through a modeling approach to the analysis of survival data, we can examine the relationship between the survival experience of patients and one or more explanatory variables. The basic model for survival data is the proportional hazards model. This model was proposed by Cox (1972) and has also come to be known as the Cox regression model. In what follows, we describe this model in the context of an application to active pharmacovigilance.

Suppose that medical records are available on n patients that have been prescribed a drug for diabetes. We define the time origin as the first time the patient takes the drug, and the end point as the occurrence of a particular adverse drug reaction of interest. The hazard of an adverse drug reaction occurrence at a particular time is assumed to depend on the values $x_1, x_2, \cdots, x_p$ of p explanatory variables, $X_1, X_2, \cdots, X_p$. The values of these variables will be assumed to have been recorded at the time origin of the study. The set of values of the explanatory variables in the proportional hazards model will be represented by the vector $\mathbf{x_i}$

$$\mathbf{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ip})^T \qquad i = 1, \ldots, n \tag{15}$$

In the model, we also include $\beta_1, \beta_2, \cdots, \beta_p$ as the unknown regression coefficients associated with the explanatory variables.

A nearly universal feature of survival data is censoring, the most common form of which is right-censoring: the period of observation expires, or an individual is removed from or leaves the study, before the event occurs. For example, in our dataset, some individuals may not have experienced the particular adverse event of interest at the end of our study, or alternatively, they may drop out of the study for various reasons prior to its termination. On the other hand, an observation is left-censored when the time at which it was initially exposed to risk is unknown.

Let T represent survival time, which is the time for initial exposure to risk to the onset of an adverse reaction. We regard T as a random variable with probability density function $f(t)$ and cumulative distribution function

$$F(t) = \mathrm{P}(\mathrm{T} < t) = \int_0^t f(u)\,\mathrm{d}u$$

which represents the probability that the survival time is less than some value t. The survivor function $S(t)$ is the complement of the cumulative distribution function, which is defined to be the probability that the survival time is greater than or equal to t, and so

$$S(t) = P(T \geq t) = 1 - F(t)$$

The hazard function is used to express the risk or hazard of death at some time t, which assesses the instantaneous risk of demise at time t, conditional on the fact that the individual has survived to time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$
$$= \frac{f(t)}{S(t)}$$

The hazard function for the *i*-th individual can be written as

$$h_i(t) = \exp(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \quad (16)$$

The constant α in the model represents a kind of log-baseline hazard, since $h_i(t) = e^{\alpha}$ when all of the explanatory variables that make up the vector $\mathbf{x_i}$ are zero. We denote the baseline hazard function as $h_0(t)$. Thus, the hazard function for the *i*-th individual is equivalent to

$$h_i(t) = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})\, h_0(t) \quad (17)$$

and this model can be re-expressed as

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (18)$$

The baseline hazard can take any form in this model. Consider, for example, two observations *i* and *j* that differ in their explanatory variable values, with the corresponding predictors

$$\gamma_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

and

$$\gamma_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_p x_{jp}$$

The hazard ratio for these two observations

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)e^{\gamma_i}}{h_0(t)e^{\gamma_j}}$$
$$= \frac{e^{\gamma_i}}{e^{\gamma_j}}$$

is independent of time t. Consequently, the Cox model is a proportional-hazards model. Remarkably, even though the baseline hazard is unspecified, the Cox model can still be estimated by the method of partial likelihood, developed by Cox (1972).

Let $t_{(1)} < t_{(2)} < \cdots < t_{(k)} < \cdots < t_{(K)}$ denote the K distinct, ordered failure times, with d_k indicating the multiplicity of failures occurring at time $t_{(k)}$. That is, d_k is the size of the set D_k of individuals that fail at $t_{(k)}$. Let $\mathbf{x}_{j(k)}$ be the vector of explanatory variables for the j-th individual who fails at the k-th ordered failure time, $t_{(k)}$, which has the same form of equation (15). Let R_k denote the risk set just before the k-th ordered failure time. This suggests that R_k is the group of individuals who are alive and uncensored at a time just prior to $t_{(k)}$. Then the approximate partial likelihood function proposed by Breslow (1974) is of the form

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{e^{\boldsymbol{\beta}' \sum_{j \in D_k} \mathbf{x}_{j(k)}}}{\left[\sum_{l \in R_k} e^{\boldsymbol{\beta}' \mathbf{x}_{l(k)}}\right]^{d_k}} \tag{19}$$

Let Y_i be an event indicator, which is zero if the i-th survival time $t_i, i = 1, 2, \cdots, n$, is right-censored, and one otherwise. Then the partial likelihood function in equation (19) can be expressed in the form

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{e^{\boldsymbol{\beta}' \sum_{j \in D_k} \mathbf{x}_{j(k)}}}{\left[\sum_{l \in R_k} e^{\boldsymbol{\beta}' \mathbf{x}_{l(k)}}\right]^{d_k}} \right\}^{y_i} \tag{20}$$

The corresponding log-likelihood function is given by

$$l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \left\{ \boldsymbol{\beta}^T \sum_{j \in Di} \mathbf{x}_{(j)} - d_i \log \sum_{l \in R_i} e^{\boldsymbol{\beta}^T \mathbf{x}_{(l)}} \right\} \tag{21}$$

We can apply the Newton-Raphson procedure to maximize the partial log-likelihood

function (21) to obtain the maximum likelihood estimates (MLE) of $\boldsymbol{\beta}$.

To do so, we initially determine the vector $\boldsymbol{u}(\boldsymbol{\beta})$, which is the set of first derivatives of the log-likelihood function in equation (21) with respect to the $\boldsymbol{\beta}$-parameters. This quantity is known as the vector of efficient scores.

$$\boldsymbol{u}(\boldsymbol{\beta}) = \left(\partial l(\boldsymbol{\beta})/\partial\, \beta_1, \partial l(\boldsymbol{\beta})/\partial\, \beta_2, \ldots, \partial l(\boldsymbol{\beta})/\partial\, \beta_p\right)$$

We also determine the matrix $\mathbf{H}(\boldsymbol{\beta})$ as the $p \times p$ matrix of second partial derivatives of the log-likelihood function, $l(\boldsymbol{\beta})$. The (j,k)th element of $\mathbf{H}(\boldsymbol{\beta})$ is then

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}$$

for $j = 1, 2, \cdots, p; k = 1, 2, \cdots, p$, and $\mathbf{H}(\boldsymbol{\beta})$ is called Hessian matrix. As we know, the variance-covariance matrix of the p maximum likelihood estimates can be approximated by $-\mathbf{H}(\boldsymbol{\beta})^{-1}$.

According to the Newton-Raphson procedure, an estimate of the vector of the $\boldsymbol{\beta}$ parameters at the m-th cycle of the iterative procedure, $\widehat{\boldsymbol{\beta}}^{(m)}$, is

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} - \left(\widehat{\mathbf{H}}\left(\boldsymbol{\beta}^{(m)}\right)\right)^{-1} \widehat{\boldsymbol{u}}\left(\boldsymbol{\beta}^{(m)}\right) \qquad (22)$$

The process can be started by an initial guess, say $\widehat{\boldsymbol{\beta}}^{(0)}$ and an iterative procedure based on equation (22) as performed. The process is terminated when the change in the log-likelihood function is sufficiently small, or when the largest of the relative changes in the values of the parameter estimates is sufficiently small.

For the Cox model, although the resulting estimates are not as efficient as maximum likelihood estimates for a correctly specified parametric hazard regression model, not having to make arbitrary, and possibly incorrect assumptions about the form of the baseline hazard is a compensating virtue of Cox's specification. However, there are additional limitations associated with the Cox model. When compared with the logistic regression model, we are faced with the challenge of specifying survival time. This can be difficult for some patients when the time that they were first exposed to risk is not known. Moreover, conditional on the values of any covariates and an individual's survival to a particular time, right-censoring must be independent of the future value of the hazard for the individual. If this condition is not met, then the estimates of the survival distribution can be seriously biased.

2.4 Random Effects Cox Models

It is sometimes the case in a survival experiment that individuals within a certain subset of the population share a common, unobservable, random frailty. For example, such a frailty could be an unobservable genetic or environmental effect. If the frailty is known, the Cox proportional hazards model for the observable covariates is valid; however, we include a random effect for the frailty as a multiplicative factor on the hazard rate.

Survival models with a single level of random effects have been proposed by Sastry

(1997), Sargent (1998) and Yau (2001). Nested frailty survival models were also considered by Sastry (1997) and Yau (2001); these studies specified gamma and log-normal distributions, respectively, for the random effects. Sargent (1998) discussed the Bayesian approaches to nested random effects Cox models. Ma (2003) proposed a Poisson modeling approach for nested random effects Cox proportional hazards models which can be used to fit flexible frailty models.

In this thesis, we consider the random effects Cox models by using the Poisson modeling approach. This method characterizes the random effects Cox model as an auxiliary random effects Poisson regression model. The orthodox best linear unbiased predictors approach is used to obtain the best linear unbiased predictors. We describe the methodology in what follows.

Consider a Cox model with a single level of random effects. Suppose that the cohort of interest consisting of n individuals is grouped in m independent clusters, and that these clusters serve as the random effects in the model. We assume that the cluster-level random effects $U_1, U_2, \cdots U_m$ are independent and identically distributed with

$$\mathrm{E}(U_j) = 1, \qquad \mathrm{var}(U_j) = \sigma^2 \tag{23}$$

Assumption (23) covers a wide range of common distributions used to describe random effects, including the gamma, inverse Gaussian and log-normal distributions.

Suppose that there are n_j individuals within the j-th cluster. Let individual (i,j) denote the i-th individual in the j-th cluster, where $i = 1,2,\cdots,n_j$. In addition, suppose that the cohort is stratified on the basis of one or more relevant covariates and that these strata are indexed by $s = 1,2,\cdots,a$. Let the hazard function for the i-th individual in the j-th cluster from stratum s at time t be denoted by $h_{ij}^{(s)}(t)$. Again, suppose that the hazard of an adverse drug reaction occurrence at a particular time depends on the values $x_1, x_2, \cdots, x_p$ of p explanatory variables, $X_1, X_2, \cdots, X_p$. Then the vector of explanatory variables for the i-th individual in the j-th cluster from stratum s is denoted as $\boldsymbol{x}_{ij}^{(s)} = \left(x_{ij1}^{(s)}, x_{ij2}^{(s)}, \cdots, x_{ijp}^{(s)}\right)^T$ with associated parameter vector $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \cdots, \beta_p)$. Given the random effects, the hazard functions for different individuals are conditionally independent, with

$$\begin{aligned} h_{ij}^{(s)}(t) &= h_0^{(s)}(t) u_i \exp\left(\beta_1 x_{ij1}^{(s)} + \beta_2 x_{ij2}^{(s)} + \cdots + \beta_p x_{ijp}^{(s)}\right) \\ &= h_0^{(s)}(t) u_i \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right) \end{aligned} \tag{24}$$

It is assumed that the distribution of random effects does not depend on the regression parameter vector $\boldsymbol{\beta}$.

Let $\tau_{s1} < \tau_{s2} \cdots < \tau_{sq_s}$ denote the distinct failure times in the s-th stratum, where q_s is the number of failures observed in stratum s. In addition, let d_{sh} represent the multiplicity of failures occurring at time τ_{sh}, where $s = 1, \cdots, a$ and $h = 1, \cdots, q_s$. The risk set at time τ_{sh} is $R(\tau_{sh}) = \left\{(i,j): t_{ij}^{(s)} \geq \tau_{sh}\right\}$, where $t_{ij}^{(s)}$ is the observed survival time for individual (i,j) from the s-th stratum. In addition, let $y_{ij,h}^{(s)} = 1$ if an adverse event occurs for individual (i,j) from the s-th stratum at time τ_{sh}, and 0

otherwise.

Suppose that we let $\mathbf{Y}$ and $\mathbf{U}$ be vectors containing the responses $Y_{ij,h}^{(s)}$ and the random effects U_j. Given the random effects $\boldsymbol{U} = \boldsymbol{u}$, the conditional partial likelihood function proposed by Cox and Oakes (1984) is

$$L_p(\boldsymbol{\beta}; \mathbf{Y}|\boldsymbol{u}) = \prod_{s=1}^{a}\prod_{h=1}^{q_s} \frac{\prod_{(i,j)\in R(\tau_{sh})} u_j^{y_{ij,h}^{(s)}} \left\{\exp\left(\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_{ij}^{(s)}\right)\right\}^{y_{ij,h}^{(s)}} (d_{sh}!)}{\left\{\sum_{(i,j)\in R(\tau_{sh})} u_j \exp\left(\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_{ij}^{(s)}\right)\right\}^{d_{sh}}} \qquad (25)$$

We use the Poisson modeling approach for this random effect Cox model. Specifically, we assume that, given that the random effects $\boldsymbol{U} = \boldsymbol{u}$, $\mathbf{Y}$ are conditionally and independently distributed with

$$Y_{ij,h}^{(s)} \sim \text{Poisson}\left\{u_j exp\left(\alpha_{sh} + \boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_{ij}^{(s)}\right)\right\} \qquad (i,j) \in R(\tau_{sh}) \qquad (26)$$

We know that for a Poisson random variable X with parameter λ, the probability density function for X is $f(x,\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$. Thus, the conditional distribution for $Y_{ij,h}^{(s)}$ given $\boldsymbol{u}$ is

$$\begin{aligned} f\left(Y_{ij,h}^{(s)}\middle|\boldsymbol{u}\right) &= \frac{\left(u_j exp\left(\alpha_{sh} + \boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_{ij}^{(s)}\right)\right)^{y_{ij,h}^{(s)}} e^{-u_j exp\left(\alpha_{sh}+\boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_{ij}^{(s)}\right)}}{y_{ij,h}^{(s)}!} \\ &= \frac{u_j^{y_{ij,h}^{(s)}} \left\{exp\left(\alpha_{sh} + \boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_{ij}^{(s)}\right)\right\}^{y_{ij,h}^{(s)}}}{\exp\left\{u_j exp\left(\alpha_{sh} + \boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{x}_{ij}^{(s)}\right)\right\}} \end{aligned} \qquad (27)$$

where $(i,j) \in R(\tau_{sh})$.

Finally, the conditional likelihood for the random effects Poisson model is

$$L(\alpha, \boldsymbol{\beta}; \mathbf{Y}|\boldsymbol{u}) = \prod_{s=1}^{a} \prod_{h=1}^{q_s} \frac{\prod_{(i,j)\in R(\tau_{sh})} u_j^{y_{ij,h}^{(s)}} \left\{exp\left(\alpha_{sh} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right)\right\}^{y_{ij,h}^{(s)}}}{\exp\left\{\sum_{(i,j)\in R(\tau_{sh})} u_j exp\left(\alpha_{sh} + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right)\right\}} \quad (28)$$

so that the log-likelihood is

$$l(\alpha, \widehat{\boldsymbol{\beta}}; \mathbf{Y}|\boldsymbol{u}) = \sum_{\mathrm{s}} \sum_{\mathrm{h}} \sum_{(i,j)\in R(\tau_{sh})} \left\{y_{ij,h}^{(s)} \log u_j + y_{ij,h}^{(s)}\left(\alpha_{sh} + \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right) - u_j exp\left(\alpha_{sh} + \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right)\right\}$$

We obtain the maximum likelihood conditional Poisson estimates $(\hat{\alpha}, \widehat{\boldsymbol{\beta}})$ for $(\alpha, \boldsymbol{\beta})$ by satisfying the equation

$$\frac{\partial l}{\partial \alpha_{sh}} = \sum_{(i,j)\in R(\tau_{sh})} \left\{y_{ij,h}^{(s)} - u_j exp\left(\alpha_{sh} + \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right)\right\} = 0$$

that is

$$\sum_{(i,j)\in R(\tau_{sh})} y_{ij,h}^{(s)} = \sum_{(i,j)\in R(\tau_{sh})} u_j exp\left(\alpha_{sh} + \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right) \quad (29)$$

We know that $\sum_{(i,j)\in R(\tau_{sh})} y_{ij,h}^{(s)}$ can be explained as the multiplicity of failures occurring at time τ_{sh}; thus, according to our notation above

$$\sum_{(i,j)\in R(\tau_{sh})} y_{ij,h}^{(s)} = d_{sh} \quad (30)$$

so that equation (29) can be written as

$$d_{sh} = \sum_{(i,j)\in R(\tau_{sh})} u_i exp\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right) \exp\left(\alpha_{sh}\right)$$

That is, the maximum likelihood conditional Poisson estimates for α_{sh} satisfy the equation

$$\exp(\hat{\alpha}_{sh}) = \frac{d_{sh}}{\sum_{(i,j)\in R(\tau_{sh})} u_j exp\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right)} \quad (31)$$

The estimate of the cumulative baseline hazard function, $\widehat{\boldsymbol{\Lambda}}_0^{(s)}(t)$, for the s-th stratum proposed by Whitehead (1980) is given by

$$\widehat{\boldsymbol{\Lambda}}_0^{(s)}(t) = \sum_{\tau_{sh} \le t} \exp(\hat{\alpha}_{sh}) \quad (32)$$

This hazard function $\widehat{\Lambda}_0^{(s)}(t)$ provides a useful interpretation of the auxiliary Poisson model.

Substituting equation (31) into the conditional likelihood for the random effects Poisson model given in equation (28), and using some simple algebra, we obtain

$$L(\hat{\alpha}, \widehat{\boldsymbol{\beta}}; \mathbf{Y}|\boldsymbol{u})$$

$$= \prod_{s=1}^{a} \prod_{h=1}^{q_s} \frac{{d_{sh}}^{d_{sh}} \exp(-d_{sh})}{d_{sh}!} \frac{\prod_{(i,j)\in R(\tau_{sh})} {u_i}^{y_{ij,h}^{(s)}} \left\{\exp\left(\boldsymbol{\beta}^{\mathbf{T}} \boldsymbol{x}_{ij}^{(s)}\right)\right\}^{y_{ij,h}^{(s)}} (d_{sh}!)}{\left\{\sum_{(i,j)\in R(\tau_{sh})} u_i \exp\left(\boldsymbol{\beta}^{\mathbf{T}} \boldsymbol{x}_{ij}^{(s)}\right)\right\}^{d_{sh}}}$$

$$= \frac{{d_{sh}}^{d_{sh}} \exp(-d_{sh})}{d_{sh}!} L_p(\widehat{\boldsymbol{\beta}}; \mathbf{Y}|\boldsymbol{u}) \qquad (33)$$

where the term $\frac{{d_{sh}}^{d_{sh}} \exp(-d_{sh})}{d_{sh}!}$ does not depend on the parameters of interest. This indicates that the maximum likelihood estimator for $\boldsymbol{\beta}$ from the Poisson likelihood in equation (28) is the same as the maximum likelihood estimate for $\boldsymbol{\beta}$ from the Cox likelihood in equation (25). We can therefore make inference about the random effects Cox model by fitting the auxiliary random effects Poisson model.

Since the random effects are unknown, they must be estimated. Algorithms for fitting random effects models usually iterate between updating the random effects and updating the regression parameters until convergence is achieved. From equation (33), we can conclude that, given the predicted random effects, the estimate of $\boldsymbol{\beta}$ for the auxiliary random effects Poisson regression model is also the estimate for the corresponding random effects Cox models. Therefore, in our example, we use the random effects predictors for the auxiliary random effects Poisson regression model to

approximate the random effects in the Cox model. Details surrounding the estimation of the random effects and regression parameters in the Cox model can be found in Krewski et al. (2009); see, in particular, Appendix B (Algorithmic description of the Cox Poisson program) and Appendix C (Computer program for random effects Cox model using the Cox-Poisson program).

The random effects Cox model is an improvement on the Cox model, since we consider unobservable genetic or environmental effects in real-life situations, treat them as random variables and incorporate random effects into the standard Cox model. An important feature of using the Poisson modeling approach to estimate random effects Cox proportional hazards models is that the principal results depend only on the first and second moments of the unobserved random effects. The orthodox best linear unbiased predictor approach is applied to the random effects Poisson model to obtain estimates of the prediction random effects, which enable us to determine optimal and consistent parameter estimates. Brockwell and Davis (1991) proved that the orthodox best linear unbiased predictors are truly the best linear unbiased predictors in the literal sense. In addition, Ma (2003) justified the asymptotic properties for cases where the number of clusters increases with the sample size. The disadvantage of the random effects Cox model is that the computational procedure is quite complicated. In addition to estimating the regression parameters and random effects, the dispersion parameter σ^2 in equation (23) is typically unknown and must also be estimated. Thus, the algorithm for fitting the model needs to iterate between

updating the regression parameter estimates via the Newton scoring algorithm, updating the random effect estimates via the orthodox best linear unbiased predictor, and updating the estimate for the dispersion parameter via the adjusted Pearson estimates.

Chapter 3
Examples

To illustrate the methods for active pharmacovigilance that were discussed in the previous chapter, we make use of the HealthFacts^{TM} data. This database is a cohort event monitoring study consisting of all records for all patients having one or more encounters with a diagnosis of diabetes and a corresponding discharge date between January 1, 2000 and June 30, 2009. In total, the database contains 614,401 patients, some of which have more than one record. Specifically, for all patients above, the database includes information on all encounters occurring between January 1, 2000 and June 30, 2009.

Cardiovascular disease is an important cause of morbidity among persons with diabetes. Records involving drugs frequently used to treat diabetes, namely Metformin (MET), Actos (AC), Avandia (AV), Insulin (INS), Chlorpropamide, Gliclazide, Glimepiride, Glipizide, Glyburide, and Tolbutamide were extracted from the above diabetes database. We decided to combine Chlorpropamide, Gliclazide, Glimepiride, Glipizide, Glyburide, and Tolbutamide, since all of these drugs contain similar active ingredients. We shall henceforth refer to this set of drugs as a

Sulfonylurea group (SUL). Although AC, AV, INS and SUL have good clinical outcomes in treating diabetes, use of these drugs was suspected to bear adverse cardiac effects. Some clinical outcomes suggest that these four drugs increase the risk of heart failure (HF) and acute myocardial infarction (AMI). In what follows, we compare the risks of HF and AMI between persons treated with AC, AV, INS and SUL to patients treated with MET, after adjusting for age and gender. In our analysis, we restrict our investigation to patients who have taken only one of the above types of drugs through the entire study period.

Thus, the selected population consisted of all patients with diabetes in the database who were treated with one and only one of the five drugs between January 1, 2000 and June 30, 2009. The cohort entry date was defined as the date of the first prescription of the drug. Each patient was followed until they either experienced an event (HF or AMI), or until June 30, 2009 if they did not experience an event. The date of admission for each outcome was taken as the index date. Based on the above, the data set under consideration consisted of 137,047 patient records. Table 1 presents the percentages of males and females taking each of the different drugs. The percentages of females prescribed each drug are similar, and slightly higher than the corresponding percentages for males. The distributions of patient ages for each of the five drugs considered given in Table 2 are also very similar.

Table 1: Distribution of patients by gender

Drug	No. of Patients	Percentage Female	Percentage Male
MET	15284	53.44%	46.56%
AC	3397	50.69%	49.31%
AV	2075	51.47%	48.53%
INS	97424	52.48%	47.52%
SUL	18867	50.04%	49.96%
Total	137047	52.20%	47.80%

Table 2: Distribution of patients by age

Drug	No. of Patients	Min	First quartile	Median	Mean	Third quartile	Max
MET	15284	9.00	54.00	65.00	63.47	74.00	90.00
AC	3397	11.00	56.00	66.00	65.03	75.00	90.00
AV	2075	16.00	55.00	65.00	64.28	75.00	90.00
INS	97424	0.00	54.00	66.00	63.80	76.00	90.00
SUL	18867	10.00	59.00	69.00	67.57	77.00	90.00
Total	137047	0.00	55.00	66.00	64.32	76.00	90.00

The number of patients with HF, AMI and both HF and AMI are shown in Table 3. Although most patients who experienced adverse events experienced either HF or AMI, a small number (just over 2%) experienced both HF and AMI. Because the joint occurrence of HF and AMI in the same patient occurs rarely, we analyzed the data for HF and AMI separately.

The last four drugs listed in Table 3 will be referred to as the test drugs, whereas

Metformin is treated as the reference drug in all analysis conducted in this thesis. The choice of the reference group warrants some discussion. From a public health perspective, it could be argued that the rate of adverse events (HF and AMI) within the general population should be used to evaluate the adverse event rate observed among diabetics taking one of the five drugs considered here. However, because diabetics represent a subpopulation whose health status is somewhat different from that of the general population, comparisons of adverse event rates in pharmacovigilance are generally made with respect to one of the drugs used to treat individuals in such subpopulations. In this case, we have chosen Metformin as the reference drug, since it is associated with the lowest adverse event rate among the diabetic drugs considered here.

Table 3: Data summary and rate ratios for use of test drugs vs Metformin

Drug	No. of Patients	HF outcomes		AMI outcomes		HF and AMI outcomes	
		No.	Percentage	No.	percentage	No.	Percentage
MET	15284	870	5.69%	150	0.98%	49	0.32%
AC	3397	315	9.27%	62	1.83%	35	1.03%
AV	2075	242	11.66%	54	2.60%	30	1.45%
INS	97424	16510	16.95%	3676	3.77%	2339	2.40%
SUL	18867	2786	14.77%	601	3.19%	327	1.73%
Total	137047	20723	15.12%	4543	3.31%	2780	2.03%

From Table 3, we can see that for heart failure (HF) cases, compared with the patients who consumed Metformin, the rate of experiencing HF for patients who consumed Actos is 63% higher (9.27% to 5.69%); the rate of experiencing HF for patients who

consumed Avandia is 105% higher; the rate of experiencing HF for patients who consumed Insulin is 198% higher; and the rate of experiencing HF for patients who consumed Sulfonylurea is 160% higher.

We find a similar result in Table 3 for the acute myocardial infarction (AMI) cases. Compared with the patients who consumed Metformin, the rate of experiencing AMI for patients who consumed Actos is 87% higher (1.83% to 0.98%); the rate of experiencing AMI for patients who consumed Avandia is 165% higher; the rate of experiencing AMI for patients who consumed Insulin is 285% higher; and the rate of experiencing AMI for patients who consumed Sulfonylurea is 226% higher.

As an illustration of the statistical methods for active pharmacovigilance described in Chapter 2, in what follows, we shall compare the odds of HF and AMI associated with the use of test drugs to that of the reference drug. Specifically, we shall fit the standard logistic regression model, a logistic regression model using James-Stein shrinkage, a Cox model, and a Cox model with random effects to the data that are summarized in Table 3. Each model will be used to analyze this diabetes dataset independently. Note that SAS is used to fit the standard logistic and Cox models, while logistic regression with James-Stein estimation and the random effects Cox model are fitted using the R software package. Examples of the code and output for each of the four models are presented in Appendix A.

3.1 Application in Logistic Regression Models

As stated above, the data set contains records of 137,047 diabetes patients. Each patient record indicates the type of drug the patient had been taking and whether the patient suffered HF or AMI. The record also contains supplementary information such as age, gender etc. The logistic regression model is used to study HF and AMI separately. For each adverse reaction, we let Y_i be one if it is present and zero otherwise. Taking heart failure as the example, if the i-th individual experienced heart failure during the study period (between January 1, 2000 and June 30, 2009), we set $Y_i = 1$. If heart failure was not experienced during the study period, we set $Y_i = 0$.

Since we have five different drugs in the entire data set, we define four indicator variables, X_1, X_2, X_3, X_4 for the model. Table 4 presents the definition of these indicator variables.

Table 4

Level of drugs	X_1	X_2	X_3	X_4
MET	0	0	0	0
AC	1	0	0	0
AV	0	1	0	0
INS	0	0	1	0
SUL	0	0	0	1

For example, if the i-th individual consumed Metformin, then $(x_{i1}, x_{i2}, x_{i3}, x_{i4}) = (0,0,0,0)$, and if the k-th individual consumed Actos, then $(x_{k1}, x_{k2}, x_{k3}, x_{k4}) = (1,0,0,0)$, etc. Note that the model can also be extended to include additional

explanatory variables. Here we will initially include the patient's age and gender along with the four indicator variables above. Then for the *i*-th individual, if $\mathbf{x_i}$ is a vector of values for the explanatory variables augmented by the constant one to allow for a intercept term in the model, we have

$$\mathbf{x_i} = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})^T$$

where x_{i5} and x_{i6} represent the age and gender for the *i*-th individual respectively. Introducing a parameter vector $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_0, \beta_1, \cdots, \beta_6)$, the logistic regression model is of the form

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}+\beta_6 x_{i6}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}+\beta_6 x_{i6}}}$$

The maximum likelihood estimate, $\widehat{\boldsymbol{\beta}}$, for the multiple logistic regression model for heart failure cases is presented in Table 5, and the results for acute myocardial infarction cases is presented in Table 6.

Table 5: Maximum Likelihood Estimates for HF cases with age and gender

Parameter	DF	Estimate	Standard error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-2.8585	0.0470	3694.9305	<.0001
AC	1	0.5237	0.0687	58.1462	<.0001
AV	1	0.7805	0.0768	103.3343	<.0001
INS	1	1.2167	0.0359	1147.0770	<.0001
SUL	1	1.0490	0.0405	669.8200	<.0001
Age	1	0.0009	0.0005	3.7020	0.0543
Gender	1	-0.0136	0.0152	0.7963	0.3722

Table 6: Maximum Likelihood Estimates for AMI cases with age and gender

Parameter	DF	Estimate	Standard error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-4.6418	0.1031	2028.5399	<.0001
AC	1	0.6191	0.1521	16.5782	<.0001
AV	1	0.9831	0.1603	37.5985	<.0001
INS	1	1.3688	0.0835	268.7527	<.0001
SUL	1	1.1870	0.0918	167.2653	<.0001
Age	1	0.0010	0.0010	1.1268	0.2885
Gender	1	-0.0586	0.0303	3.7557	0.0526

Consider the heart failure cases first. According to the results presented in Table 5, we get the maximum likelihood estimates for $\boldsymbol{\beta}$. Thus, π_i can be estimated using

$$\hat{\pi}_i = \frac{e^{-2.8585+0.5237x_{i1}+0.7805x_{i2}+1.2167x_{i3}+1.0490x_{i4}+0.0009x_{i5}-0.0136x_{i6}}}{1+e^{-2.8585+0.5237x_{i1}+0.7805x_{i2}+1.2167x_{i3}+1.0490x_{i4}+0.0009x_{i5}-0.0136x_{i6}}}$$

To assess the overall fit of the model, we would test the hypothesis

$H_0: \beta_1 = \beta_2 = \cdots = \beta_6 = 0$ vs H_a: At least one β in H_0 is not zero (34)

It is possible to conduct this test using a likelihood ratio chi-square statistic. For a likelihood ratio test here, we maximize the likelihood without all the variables in the model and also with all the variables included. We then let Λ denote the ratio of these maximized likelihoods and consider $G^2 = -2\log\Lambda$ as the test statistic. Since the statistic follows a chi-square distribution with degrees of freedom p, we can calculate a P-value, and use it to conclude whether our model fits the data or not.

According to the SAS output attached in Appendix A, for heart failure cases we get the likelihood ratio chi-square statistic $G^2 = 1710.1735$. Since this statistic follows a chi-square distribution with 6 degrees of freedom, the associated P-value is less than

0.0001. Alternatively, the chi-square statistic for performing a Wald test is 1299.9874 with 6 degrees of freedom, and the P-value is also less than 0.0001. The small P-value implies that there is strong evidence that the model fits the data of heart failure cases. Similarly, the likelihood ratio chi-square statistic for acute myocardial infarction cases is $G^2 = 453.654$, which also follows a chi-square distribution with 6 degrees of freedom. The P-value for testing model fit is again less than 0.0001, which indicates that the model appropriately describes the acute myocardial infarction cases.

To assess the significance of each of the variables in the model, we would test the hypothesis

$$H_0: \beta_j = 0 \qquad vs \qquad H_a: \beta_j \neq 0 \qquad \text{where } j = 1,2,\cdots,6 \qquad (35)$$

It is possible to conduct this test using a Wald statistic test (1943). Under this test, the maximum likelihood estimate $\hat{\beta}_j$ is normally distributed with mean 0 and standard error $se(\beta_j)$. Then for each of the β_js, the test statistic is $\frac{\beta_j}{se(\beta_j)}$, which follows a standard normal distribution. Alternatively, for each of the β_j, the square of this test statistic, $\frac{{\beta_j}^2}{(se(\beta_j))^2}$, is distributed as chi-square with one degree of freedom. In what follows, we use the Wald test statistic $W = \frac{{\beta_j}^2}{(se(\beta_j))^2}$ to test the significance of x_j in the model. We set the significance level as $\alpha = 0.05$ in the analysis.

For both heart failure cases and acute myocardial infarction cases, we find that the P-values for the indicator variables of AC, AV, INS and SUL are less than 0.0001, which implies strong evidence of a difference in the occurrence of the corresponding

ADRs of each drug relative to Metformin. Note also that, for heart failure cases, the P-value for age is 0.0543 and the P-value for gender is 0.3722; thus both are greater than $\alpha = 0.05$. Similarly, for acute myocardial infarction cases, the P-value for age is 0.2885 and P-value for gender is 0.0526, both of which are also larger than $\alpha = 0.05$. Thus, it would appear that the variables of age and gender are not strongly significant in both models. However, it is best to test for the effect of age and gender simultaneously. To do so, we test

$$H_0: \beta_5 = \beta_6 = 0 \qquad vs \qquad H_a: at\ least\ one\ of\ \beta_5\ and\ \beta_6 \neq 0 \qquad (36)$$

In order to conduct this test, we need to compare the model with the indicator variables for the drugs, age and gender to a model that only contains the indicator variables for the drugs (a model without age and gender). For HF cases, the likelihood ratio chi-square statistic for the model containing age and gender is $G_F^2 = 1710.1735$, which follows a chi-square distribution with 6 degrees of freedom. Similarly, the likelihood ratio chi-square statistic for the model without age and gender is $G_R^2 = 1705.8382$ which follows a chi-square distribution with 4 degrees of freedom. Note that the estimates for this model are given in Table 7. Then $\Delta G^2 = G_F^2 - G_R^2 = 1710.1735 - 1705.8382 = 4.3353$ follows a chi-square distribution with degrees of freedom $\Delta df = 6 - 4 = 2$. Thus we can calculate the P-value as $P(\chi_2^2 \geq 4.3353) = 0.1144$. The P-value is close to 0.10, suggesting that the effect of age and gender simultaneously are not very strong. We will examine this further in what follows.

Table 7: Maximum Likelihood Estimates for HF cases without age and gender

Parameter	DF	Estimate	Standard error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-2.8063	0.0349	6468.5890	<.0001
AC	1	0.5256	0.0687	58.5607	<.0001
AV	1	0.7815	0.0768	103.6097	<.0001
INS	1	1.2171	0.0359	1147.9631	<.0001
SUL	1	1.0533	0.0405	667.9631	<.0001

From Table 5 and Table 7, we find that the maximum likelihood estimates of the coefficients associated with the indicator variables for AC, AV, INS and SUL from the logistic model without age and gender are very close to estimates from the model including age and gender, which indicates that the effect of age and gender simultaneously for HF cases does not appear to be strong. We therefore conclude that the predictor variables of age and gender can be excluded from the model.

To test the effect of age and gender simultaneously on AMI cases, the likelihood ratio chi-square statistic for the model with age and gender is $G_F^2 = 453.6541$ with degrees of freedom $df = 6$. Similarly, the likelihood ratio chi-square statistic for the model without age and gender is $G_R^2 = 448.9788$ with degrees of freedom $df = 4$. To test the null hypothesis of (36), the test statistic is $\Delta G^2 = G_F^2 - G_R^2 = 453.6541 - 448.9788 = 4.6753$, which follows a chi-square distribution with degrees of freedom $\Delta df = 6 - 4 = 2$. The P-value is calculated as $P(\chi_2^2 \geq 4.6753) = 0.0966$. The P-value is also close to 0.10, which suggests that the effect of age and gender simultaneously on AMI cases is also not very strong. The results for fitting the model without age and gender are listed in Table 8.

Table 8: Maximum Likelihood Estimates for AMI cases without age and gender

Parameter	DF	Estimate	Standard error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-4.6074	0.0818	3173.8337	<.0001
Actos	1	0.6223	0.1520	16.7524	<.0001
Avandia	1	0.9850	0.1603	37.7524	<.0001
Insulin	1	1.3697	0.0835	269.1201	<.0001
Sulfonylurea	1	1.1932	0.0917	169.3482	<.0001

For AMI cases, comparing the results in Table 6 and Table 8, we also find that the maximum likelihood estimates for the coefficients associated with the indicator variables for the drug in the model with age and gender are very close to counterparts in the model without age and gender, which suggests that the simultaneous effect of age and gender on AMI cases is not very strong. As with the HF cases, we conclude that both of the predictor variables can be excluded from the model.

Since there are no multiple drugs consumed by the individuals in our data set and we have concluded that the effects of age and gender are not very strong, we can consider for HF the estimates in Table 7. Thus, the explanatory variables for users of Actos should be the form of $(x_1, x_2, x_3, x_4)_{ac} = (1,0,0,0)$, and the explanatory variables for Metformin users have the form of $(x_1, x_2, x_3, x_4)_{met} = (0,0,0,0)$. Thus, from the logistic regression model without age and gender, we can show that the odds that a patient who consumed Actos will experience HF is

$$\frac{\pi(\boldsymbol{x_{ac}})}{1-\pi(\boldsymbol{x_{ac}})} = \frac{\dfrac{e^{\beta_0+\beta_1(1)+\beta_2(0)+\beta_3(0)+\beta_4(0)}}{1+e^{\beta_0+\beta_1(1)+\beta_2(0)+\beta_3(0)+\beta_4(0)}}}{\dfrac{1}{1+e^{\beta_0+\beta_1(1)+\beta_2(0)+\beta_3(0)+\beta_4(0)}}} = e^{\beta_0+\beta_1}$$

while the odds that a patient who consumed Metformin will experience HF is

$$\frac{\pi(\boldsymbol{x}_{met})}{1-\pi(\boldsymbol{x}_{met})} = \frac{\frac{e^{\beta_0+\beta_1(0)+\beta_2(0)+\beta_3(0)+\beta_4(0)}}{1+e^{\beta_0+\beta_1(1)+\beta_2(0)+\beta_3(0)+\beta_4(0)}}}{\frac{1}{1+e^{\beta_0+\beta_1(0)+\beta_2(0)+\beta_3(0)+\beta_4(0)}}} = e^{\beta_0}$$

Therefore, the ratio of the odds that a patient who takes Actos will experience HF to the odds that a patient who takes Metformin will experience HF is

$$\frac{\frac{\pi(\boldsymbol{x}_{ac})}{1-\pi(\boldsymbol{x}_{ac})}}{\frac{\pi(\boldsymbol{x}_{met})}{1-\pi(\boldsymbol{x}_{met})}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Thus e^{β_1} represents the ratio of the odds of the occurrence of HF with Actos to the odds of the occurrence of HF with Metformin. Similarly, e^{β_2}, e^{β_3}, e^{β_4} represent the ratio of the odds of the occurrence of HF with Avandia, Insulin and Sulfonylurea respectively to the odds of the occurrence of HF with Metformin. Therefore, in this case, each β_j, specifically $\beta_1, \beta_2, \beta_3, \beta_4$, could be interpreted as a log odds ratio.

We get the estimate for β_1 from Table 7 as $\hat{\beta}_1 = 0.5256$; thus $e^{\hat{\beta}_1} = e^{0.5256} = 1.691$. Therefore, the odds of experiencing heart failure for patients who consumed Actos is 1.691 times the odds of experiencing heart failure for patients who consumed Metformin. In addition, from Table 7, a 95% confidence interval for β_1 is $0.5256 \pm 1.96(0.0687)$ or $(0.3909, 0.6603)$. Thus, a 95% confidence interval for the odds ratio e^{β_1} is $(e^{0.3909}, e^{0.6603})$ or $(1.478, 1.935)$. Therefore, we estimate with 95% confidence that the odds of experiencing heart failure for patients who consumed Actos is anywhere from 1.478 to 1.935 times the odds of experiencing heart failure for patients who consumed Metformin.

As a result, compared with Metformin users, the users of Actos were at a significantly increased risk of heart failure (odds ratio, 1.691; 95% confidence interval 1.478-1.935; P<.0001). Similarly, the odds of experiencing heart failure for patients who consumed Avandia is 2.183 times the odds of experiencing heart failure for patients who consumed Metformin (odds ratio, 2.185; 95% confidence interval 1.880-2.540; P<.0001). The odds of experiencing heart failure for patients who consumed Insulin is 3.377 times the odds of experiencing heart failure for patients who consumed Metformin (odds ratio, 3.377; 95% confidence interval 3.148-3.624; P<.0001). Moreover, the odds of experiencing heart failure for patients who consumed Sulfonylurea is 2.867 times the odds of experiencing heart failure for patients who consumed Metformin (odds ratio, 2.867; 95% confidence interval 2.648-3.104; P<.0001). Of note is the fact that patients who took Metformin are at significantly lower risk for experiencing heart failure than patients who were prescribed the other drugs. Patients who took Insulin seemed to be at the greatest risk, followed by those who consumed Sulfonylurea. Patients who took Avandia and Actos were at lower risk, but still at significantly higher risk than those taking Metformin. The odds ratio point and interval estimates summarized above for HF cases are given in Table 9.

Table 9: Odds ratio estimates for HF cases without age and gender

Effect	Point Estimate	95% Wald Confidence limits	
AC	1.691	1.478	1.935
AV	2.185	1.880	2.540
INS	3.377	3.148	3.624
SUL	2.867	2.648	3.104

Similarly, we can determine analogous estimates for acute myocardial infarction cases. The results are presented in Table 10.

Table 10: Odds ratio estimates for AMI cases without age and gender

Effect	Point Estimate	95% Wald Confidence limits	
AC	1.863	1.383	2.510
AV	2.678	1.956	3.667
INS	3.934	3.340	4.434
SUL	3.298	2.755	3.947

From Table 10, the users of Actos were at 86.3% increased risk of AMI versus users of Metformin (odds ratio, 1.863; 95% confidence interval, 1.383-2.510; P<.0001), while the users of Avandia were at 167.8% increased risk of AMI versus users of Metformin (odds ratio, 2.678; 95% confidence interval, 1.956-3.667; P<.0001). Moreover, the users of Insulin were at a 293.4% increased risk of AMI compared with users of Metformin (odds ratio, 3.934; 95% confidence interval, 3.340-4.434; P<.0001), while the users of Sulfonylurea were at a 229.8% increased risk of AMI compared with users of Metformin (odds ratio, 3.298; 95% confidence interval, 2.755-3.947; P<.0001). Comparing the results in Tables 9 and 10, we find that the relative sizes of the risks for different drug users experiencing AMI are the same as that for the in HF cases.

3.2 Application in James-Stein Shrinkage Estimation

In the previous section, logistic regression was used to study the occurrence of HF and AMI for patients taking the various drugs of interest. The results indicated that for drugs where the odds of HF was relatively low, the odds of AMI was also relatively low, and vice versa. As we know, heart failure and myocardial infarction are both adverse cardiovascular events. It is reasonable therefore to suspect that a drug that increases the risk of one of these events will also increase the risk associated with the other. Thus, we apply a James-Stein type shrinkage estimation approach in a logistic regression context to simultaneously model the HF and AMI data, thereby consolidating information and borrowing strength across medically related adverse drug reactions.

Specifically, we are interested in the simultaneous association between the two adverse events HF and AMI and the drugs Actos, Avandia, Insulin, Sulfonylurea and Metformin. To date, we have fit two independent logistic regression models based on the two data sets for HF and AMI respectively. We now combine these two models into a single one, using James-Stein estimation for fitting. The design matrix $\mathbf{X}$ will have the form of equation (8).

From the results of the logistic regression models for both HF cases and AMI cases in Table 5 through Table 9, we discovered that age and gender do not have strong association with both HF and AMI when the effect of different drugs is taken into

account. Therefore, we shall exclude the predictor variables of age and gender in the design matrix **X**. Suppose the logistic regression parameters for HF cases are denoted as $\boldsymbol{\beta}_{\mathbf{HF}}{}^{\mathrm{T}} = (\beta_{01}, \beta_{11}, \cdots, \beta_{41})$, and the logistic regression parameters for AMI cases are denoted as $\boldsymbol{\beta}_{\mathbf{MI}}{}^{\mathrm{T}} = (\beta_{02}, \beta_{12}, \cdots, \beta_{42})$; then, in the full model, the unrestricted parameter vector corresponding to the **X** matrix assumes the form

$$\widehat{\boldsymbol{\beta}}^{un} = (\beta_{01}, \beta_{02}, \beta_{11}, \beta_{12}, \cdots, \beta_{41}, \beta_{42})^{T}$$

In order to test whether, relative to Metfomin, the four test drugs simultaneously increase the risk of HF and AMI, we can test the hypothesis H_0: $\mathbf{C\beta} = \mathbf{d}$. Here $\mathbf{C}$ is a matrix with 4 rows and 10 columns of the form

$$\mathrm{C} = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

and $\mathbf{d} = (0, 0, 0, 0)^{T}$. The restricted estimator $\widetilde{\boldsymbol{\beta}}^{re}$, computed under the null hypothesis, is obtained using equation (10).

In order to test the null hypothesis, the test statistic is calculated by equation (11). The quantity s_e^2 in equation (11) is the mean square error under the full model; it is calculated according to equation (13) by using the Pearson's chi-square statistic. Equation (11) yields the test statistic $F_{(r,m)} = 40.43$ with degrees of freedom $(r, m) = (4, 274100)$. It is obvious that the P-value is close to zero, which indicates that the null hypothesis should be rejected.

The James-Stein estimates are obtained using equation (14); these estimates, denoted as $\widehat{\boldsymbol{\beta}}^{JS}$, are presented in Table 11.

Table 11: The result for James-Stein shrinkage estimation

Covariate	Event	$\widehat{\boldsymbol{\beta}}^{un}$	$\exp(\widehat{\boldsymbol{\beta}}^{un})$	$\widetilde{\boldsymbol{\beta}}^{re}$	$\exp(\widetilde{\boldsymbol{\beta}}^{re})$	$\widehat{\boldsymbol{\beta}}^{JS}$	$\exp(\widehat{\boldsymbol{\beta}}^{JS})$
Intercept	HF	-2.8063	0.060	-2.8729	0.057	-2.8071	0.060
	AMI	-4.6074	0.010	-4.5408	0.011	-4.6065	0.010
AC	HF	0.5256	1.691	0.5739	1.775	0.5262	1.692
	AMI	0.6223	1.863	0.5739	1.775	0.6217	1.862
AV	HF	0.7815	2.185	0.8833	2.419	0.7828	2.188
	AMI	0.9850	2.678	0.8833	2.419	0.9838	2.675
INS	HF	1.2171	3.377	1.2934	3.645	1.2180	3.381
	AMI	1.3697	3.934	1.2934	3.645	1.3687	3.930
SUL	HF	1.0533	2.867	1.1232	3.075	1.0541	2.870
	AMI	1.1932	3.398	1.1232	3.075	1.1924	3.295

By using James-Stein shrinkage estimation, we pool the data on HF and AMI cases to 'borrow strength' to obtain improved estimates of the logistic regression parameters. From the results in Table 11, we find that compared with patients consuming Metformin, the patients who consumed Actos have 69.2% increased risk of experiencing HF (since the odds ratio $e^{\beta_{11}} = e^{0.5262} = 1.692$) and 86.2% increased risk of experiencing AMI; the patients who consumed Avandia have 118.8% increased risk of experiencing HF and 167.5% increased risk of experiencing AMI; the patients who consumed Insulin have 238.1% increased risk of experiencing HF and 293.0% increased risk of experiencing AMI; and the patients who consumed Sulfonylurea have 187.0% increased risk of experiencing HF and 229.5% increased risk of experiencing AMI. Comparing the odds ratio estimates associated with the James-Stein shrinkage estimators to those from the two logistic regression models, we

find that they are very similar.

3.3 Application in Cox Models

For the cohort study in the diabetes database, a patient's record contains information on the date that a drug was initially prescribed after January 1, 2000, and on the dates of diagnosis of an adverse event. In order to incorporate this information into the study of likelihood of particular drug-ADR combinations, we can make use of a Cox model.

Using the above data set of 137,047 diabetes patients who have taken only one of the five drugs (Actos, Avandia, Insulin, Sulfonylurea and Metformin) during the study period, the date for first drug usage was recognized as the time origin. The date of the onset of an adverse drug reaction (HF and AMI) was treated as the endpoint. Some of these patients did not experience the ADRs by the time that the study was completed, so these individuals were treated as right-censored. The endpoint for these patients was recorded as the last day of the study period (June 30, 2009). The survival time is referred to as the time period between the time origin and the endpoint. The values of explanatory variables such as the age of the patients in years, their gender etc. were recorded at the time origin.

To illustrate the fitting of a Cox model, we begin by considering HF as the ADR of

interest. If HF was experienced by the i-th patient over the course of the study period, the dependent variable Y_i was coded as one, and zero if the individual had not experienced HF by the end of the study (Note that these data would be censored). Since we want to explore whether the consumption of the drugs of interest Actos, Avandia, Insulin and Sulfonylurea will increase the risk of HF versus the arbitrary chosen reference drug Metformin, we include four indicator variables, X_1, X_2, X_3, X_4, with the values presented in Table 4 in our model. In addition, we also include predictor variables X_5 and X_6 for age and gender, respectively. Thus, the vector of explanatory variables for the i-th individual is the form of $\mathbf{x_i} = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})^T$, while the parameter vector is $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \cdots, \beta_6)$. Then if t represents the time period that passes from when a patient was initially prescribed a drug to the day on which an ADR is experienced, or the end of the study, the hazard function for the i-th individual is

$$h_i(t) = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_6 x_{i6})\, h_0(t)$$

where $h_0(t)$ represents the baseline hazard function.

The Cox model employed here involves the assumption of proportional hazards, with the effect of each covariate being to proportionally increase or decrease the hazard at each point in time. We used SAS to evaluate the proportional hazards assumption for each covariate included in the Cox model (Harrell, 1986). This was done by finding the correlation between the Schoenfeld residuals (Schoenfeld, 1982) for a particular covariate and the ordered individual failure times. If the proportional hazards

assumption is met, the correlation between the residuals and the ordered failure times for any given covariate should be near zero. The P-values for testing a correlation of zero for each covariate are given in Table 12 for both the HF and AMI cases. Since the P-values for the HF cases are all greater than 0.05, the proportional hazards assumption does not appear unreasonable. A similar conclusion can be drawn for the AMI cases.

Table 12: P-values for the testing of the proportional hazard assumption

Cases \ Variables	AC	AV	INS	SUL	Age	Gender
HF	0.6465	0.4965	0.0582	0.0631	0.0742	0.5217
AMI	0.7819	0.0528	0.0858	0.1035	0.8808	0.0615

The maximum likelihood estimates for the Cox model fit to HF cases are presented in Table 13.

Table 13: Analysis of MLE for HF cases using Cox model with age and gender

Parameter	DF	Parameter estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio CL	
AC	1	0.48636	0.0657	54.73	<.0001	1.626	1.430	1.850
AV	1	0.61563	0.0727	71.77	<.0001	1.851	1.605	2.134
INS	1	1.14158	0.0348	1079.33	<.0001	3.132	2.926	3.352
SUL	1	0.92150	0.0389	562.39	<.0001	2.513	2.329	2.712
Age	1	0.00094	0.0004	4.53	0.0333	1.001	1.000	1.002
Gender	1	-0.0100	0.0139	0.52	0.4714	0.990	0.963	1.017

Table 14: Analysis of MLE for AMI cases using Cox model with age and gender

Parameter	DF	Parameter estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio CL	
AC	1	0.58887	0.1509	15.24	<.0001	1.802	1.341	2.422
AV	1	0.81778	0.1586	26.59	<.0001	2.265	1.660	3.091
INS	1	1.33905	0.0830	260.07	<.0001	3.815	3.242	4.490
SUL	1	1.07332	0.0912	138.667	<.0001	2.925	2.447	3.497
Age	1	0.00110	0.0009	1.34	0.2466	1.001	0.999	1.003
Gender	1	-0.05440	0.0297	3.35	0.0671	0.947	0.893	1.004

As before, the likelihood ratio test statistic can be used to assess the overall fit of the model. From the SAS output, we obtain $G^2 = 1731.3002$, which follows a chi-square distribution with 6 degrees of freedom; thus the corresponding P-value is less than 0.0001, suggesting that the model fits the data well.

A similar analysis was performed using AMI as the ADR of interest. The maximum likelihood estimates for the Cox model are given in Table 14. This model also appears to fit the data well. The likelihood ratio test statistic, which follows a chi-square distribution with 6 degrees of freedom, is $G^2 = 457.6473$. The associated P-value is again less than 0.0001.

Before we investigate the effect of the different drugs on HF and AMI using these two Cox models, it is worthwhile to test the simultaneous effects of age and gender in both

models independently. The null hypothesis for both HF and AMI cases is of the form (37). For HF, the likelihood ratio test statistic for the model with age and gender is $G_F^2 = 1731.3002$, which follows a chi-square distribution with 6 degrees of freedom. Similarly, the likelihood ratio test statistic for the model without age and gender is $G_R^2 = 1726.3863$, which follows a chi-square distribution with 4 degrees of freedom. Then $\Delta G^2 = G_F^2 - G_R^2 = 1731.3002 - 1726.3863 = 4.9139$ follows a chi-square distribution with degrees of freedom $\Delta df = 6 - 4 = 2$. Thus, the P-value for testing the simultaneous contribution of age and gender in the Cox model for HF is P-value $= P(\chi_2^2 \geq 4.9139) = 0.0857$. Similarly, for AMI cases, we obtain $G_F^2 = 457.6473$ and $G_R^2 = 453.1663$, so that $\Delta G^2 = G_F^2 - G_R^2 = 4.481$, which follows a chi-square distribution with $\Delta df = 6 - 4 = 2$ degrees of freedom. Here, the P-value is $P(\chi_2^2 \geq 4.481) = 0.1064$. Since the P-values for both HF and AMI cases are greater than 0.05, it is suggested that the effect of age and gender simultaneously in HF and AMI cases are not very strong. In addition, Table 15 and 16 present the maximum likelihood estimates for the coefficients associated with the indicator variables for drugs in the Cox models for HF and AMI that do not include age and gender. A comparison of the estimates in Tables 15 and 16 with counterparts in Tables 13 and 14 suggest that inclusion of age and gender in the two Cox models has little effect on the drug/ADR relationship for both HF and AMI.

Table 15: Analysis of MLE for HF case using Cox model without age and gender

Parameter	DF	Parameter estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio CL	
AC	1	0.48802	0.0657	55.11	<.0001	1.629	1.432	1.853
AV	1	0.61662	0.0727	72.00	<.0001	1.853	1.607	2.136
INS	1	1.14205	0.0348	1080.27	<.0001	3.133	2.927	3.354
SUL	1	0.92571	0.0388	568.89	<.0001	2.524	2.339	2.723

Table 16: Analysis of MLE for AMI case using Cox model without age and gender

Parameter	DF	Parameter estimate	Standard Error	Chi-Square	Pr>ChiSq	Hazard Ratio	95% Hazard Ratio CL	
AC	1	0.59178	0.1508	15.39	<.0001	1.807	1.345	2.429
AV	1	0.81977	0.1586	26.72	<.0001	2.270	1.664	3.097
INS	1	1.33983	0.0830	260.38	<.0001	3.818	3.245	4.493
SUL	1	1.07962	0.0911	140.58	<.0001	2.944	2.462	3.519

Thus, we can draw appropriate conclusions using estimates in Tables 15 and 16 instead of Tables 13 and 14. Hence, compared with patients consuming Metformin, the patients who consumed Actos were at 62.9% greater risk of experiencing HF (hazard ratio, 1.629; 95% confidence interval 1.432-1.853; P<.0001) and 80.7% greater risk of experiencing AMI (hazard ratio, 1.807; 95% confidence interval, 1.345-2.429; P<.0001); the patients who consumed Avandia were at 85.3% greater risk of HF (hazard ratio, 1.853; 95% confidence interval, 1.607-2.136; P<.0001) and 127.0% greater risk of AMI (hazard ratio, 2.270; 95% confidence interval,

1.664-3.097; P<.0001); the patients who consumed Insulin were at 213.3% (hazard ratio, 3.133; 95% confidence interval, 2.927-3.354; P<.0001) and 281.8% greater risk of HF and AMI respectively (hazard ratio, 3.818; 95% confidence interval, 3.245-4.493; P<.0001); and the patients who consumed Sulfonylurea were at 152.4% (hazard ratio, 2.524; 95% confidence interval, 2.339-2.723; P<.0001) and 194.4% (hazard ratio, 2.944; 95% confidence interval, 2.462-3.519; P<.0001) greater risk of HF and AMI respectively.

3.4 Application in Cox Models with Random Effects

In practice, it may be the case that the time elapsing between initial drug prescription and the occurrence of an ADR for individuals in some subgroups of the population are associated since members in these subgroups share a common unobserved trait. Since the cohort of interest in our example is grouped by 84 different hospitals, we now wish to acknowledge the effect of hospital on a particular drug/ADR relationship. To do so, we consider two independent Cox models with the random effects, one for each of the ADRs of interest, HF and AMI. We also consider age and gender of patients through the stratification of the data set according to these two variables. We describe this stratification below.

Consider first the random effects Cox model to study the drug/reaction relationship when the ADR is HF. Again, the data set contains 137,047 records, which are

clustered by hospitals. We incorporate these as random effects into the Cox proportional hazards model. Since there are 84 different hospitals, we define the hospital random effects as $U_1, \cdots, U_{84}$ and assume that they are independently and identically distributed positive random effects with $\mathrm{E}(U_j) = 1$, $\mathrm{var}(U_j) = \sigma^2, (j = 1, \cdots, 84)$. Within each hospital j, there are n_j individuals. The cohort was stratified by five-year age categories (ages are between 0 and 90 in the data set) for each gender, yielding 36 strata. We index these by $s = 1, 2, \cdots, 36$. Let $\tau_{s1} < \tau_{s2} \cdots < \tau_{sq_s}$ denote the distinct failure times in the s-th age/gender stratum, where q_s represents the total number of patients in stratum s that experienced an ADR over all 84 hospitals. We also let $y_{ij,h}^{(s)}$ be an indicator variable for the i-th patient at the j-th hospital falling into stratum s, and set $y_{ij,h}^{(s)} = 1$ if the ADR of interest, HF, occurs for this individual by time τ_{sh}, and $y_{ij,h}^{(s)} = 0$ otherwise. In the Cox random effects model for describing the occurrence of HF, we include 4 variables X_1, X_2, X_3, X_4, with the values presented in Table 4, which indicate which of the drugs of interest (Actos, Avandia, Insulin, Sulfonylurea and Metformin) was consumed by a given individual. Thus, the vector of explanatory variables for individual (i, j) in stratum s is denoted as $\boldsymbol{x}_{ij}^{(s)} = \left(x_{ij1}^{(s)}, x_{ij2}^{(s)}, x_{ij3}^{(s)}, x_{ij4}^{(s)}\right)^T$, while the associated parameter vectors is $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \cdots, \beta_4)$. Given the hospital random effects, the hazard functions for individuals are conditionally independent, with

$$h_{ij}^{(s)}(t) = h_0^{(s)}(t) u_i \exp\left(\beta_1 x_{ij1}^{(s)} + \beta_2 x_{ij2}^{(s)} + \beta_3 x_{ij3}^{(s)} + \beta_4 x_{ij4}^{(s)}\right)$$

$$= h_0^{(s)}(t) u_i \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{ij}^{(s)}\right)$$

where $i = 1, \cdots, n_j; j = 1, \cdots, 84;\ s = 1, \cdots, 36$.

By characterizing the random effects Cox model as an auxiliary random effects Poisson regression model as illustrated in Chapter 2, we obtain estimates of the model parameter by using the orthodox best linear unbiased predictor approach. The results are listed in Table 17. Similarly, repeating the analysis for AMI produces the results in Table 18.

Table 17: Analysis of parameter estimates using Random effects Cox model for HF cases

Parameter	Coefficient	Std. Error	t	exp(Coef)	Lower 95%	Upper 95%
AC	0.4868	0.0660	7.379	1.627	1.430	1.852
AV	0.6135	0.0730	8.406	1.847	1.601	2.131
INS	1.1463	0.0350	32.762	3.147	2.938	3.370
SUL	0.9237	0.0389	23.721	2.519	2.334	2.718

The random effects dispersion parameter is $\hat{\sigma}_{\mathrm{HF}}{}^2 = 0.0006$. Estimates for hospital random effects are given in Appendix A.

Table 18: Analysis of parameter estimates using Random effects Cox model for AMI cases

Parameter	Coefficient	Std. Error	t	exp(Coef)	Lower 95%	Upper 95%
AC	0.5893	0.1509	3.906	1.803	1.341	2.423
AV	0.8175	0.1586	5.154	2.265	1.660	3.091
INS	1.3375	0.0831	16.091	3.810	3.237	4.484
SUL	1.0711	0.0912	11.747	2.919	2.441	3.490

The random effects dispersion parameter is $\hat{\sigma}_{\mathrm{AMI}}{}^2 = 6.167\mathrm{e}-06$. Estimates for hospital random effects are given in Appendix A.

From the R output in Appendix A, we find that the estimates of the hospital random effects are all very close to one for both the HF and AMI models. In addition, the estimates of the dispersion parameters are very small $\left(\hat{\sigma}_{HF}{}^2 = 0.0006 \text{ and } \hat{\sigma}_{AMI}{}^2 = 6.167e-06\right)$ for both HF cases and AMI cases. This suggests that there is not a significant difference in the occurrence of HF and AMI for patients at different hospitals. It also explains why the parameter estimates in the standard Cox model (See Tables 15 and 16) are close to the parameter estimates in the Cox models with random effects in Tables 17 and 18; these latter estimates may be interpreted as follows.

Compared with patients who consumed Metformin, the patients who consumed Actos were at 62.7% greater risk of experiencing HF (hazard ratio, 1.627; 95% confidence interval, 1.430-1.852) and 80.3% greater risk of AMI (hazard ratio, 1.803; 95% confidence interval, 1.341-2.423); the patients who consumed Avandia were at 84.7% greater risk of HF (hazard ratio, 1.847; 95% confidence interval, 1.601-2.131) and 126.5% greater risk of AMI (hazard ratio, 2.265; 95% confidence interval, 1.660-3.091); the patients who consumed Insulin were at 214.7% (hazard ratio, 3.147; 95% confidence interval, 2.938-3.370) and 281.0% (hazard ratio, 3.810; 95% confidence interval, 3.237-4.484) greater risk of HF and AMI respectively; and the patients who consumed Sulfonylurea were at 151.9% (hazard ratio, 2.519; 95% confidence interval, 2.334-2.718) and 191.9% (hazard ratio, 2.919; 95% confidence interval, 2.441-3.490) greater risk in HF and AMI cases respectively.

3.5 Comparison of the Results in Section 3.1 to 3.4

We have used four different models to fit the diabetes data set to explore the association of the five drugs (Actos, Avandia, Metformin, Insulin and Sulfonylurea) to the adverse drug reactions of HF and AMI in Section 3.1 through 3.4. In this section, we compare the results for HF and AMI separately. For HF, the odds ratio estimates obtained under the four models for the different drugs relative to Metformin are given in Table 19. Table 20 presents the analogous estimates for AMI.

Table 19: Rate ratio estimate for HF cases using four different models

Parameter	LR	JS	Cox	RE Cox
AC	1.691	1.692	1.629	1.627
AV	2.185	2.188	1.859	1.847
INS	3.377	3.381	3.139	3.147
SUL	2.867	2.870	2.514	2.519

Table 20: Rate ratio estimate for AMI cases using four different models

Parameter	LR	JS	Cox	RECox
AC	1.863	1.862	1.807	1.803
AV	2.678	2.675	2.270	2.265
INS	3.934	3.930	3.818	3.810
SUL	3.289	3.295	2.944	2.919

Abbreviation: LR - logistic regression model; JS – James-Stein shrinkage estimation; Cox - Cox model; RE Cox – random effect Cox model.

We find that, for both the HF and AMI cases, there is little difference for the odds

ratio estimates obtained using a simple logistic regression model, a logistic model with James-Stein shrinkage, a Cox model, and a random effects Cox model. Thus, for the particular application under consideration, standard logistic regression seems sufficient to investigate the drug/ADR relationship for both HF and AMI.

3.6 Effect of Time Since Exposure using Logistic Model

For the dataset consisting of patients with diabetes, we have investigated the relationship between particular drugs (AC, AV, MET, INS, SUL) and the ADRs of HF and AMI. Another question of interest is whether the time since first exposure (TSE) to a drug affects the occurrence of HF and AMI and if so, whether the relationship between a particular ADR and TSE is different for different drugs. We define TSE for a patient here as the difference between the first time to take a given drug and the occurrence of the ADR. If the ADR is not observed, TSE is defined as the difference between the time that the drug was initially prescribed and the last day of the study period (June 30, 2009). Since the results in Section 3.5 suggest that the more complicated models considered here do not add much over logistic regression, in so far as the relationship between particular drug/ADR combinations of interest are concerned, we shall restrict our analysis in this section to the logistic regression model only.

As stated earlier, the dataset of diabetes patients of interest consisted of 137,047

individuals who were prescribed one and only one of the five drugs under study over the period from January 1, 2000 to June 30, 2009. However, for patients that were prescribed a drug shortly after the start date of the study, there was some concern as to whether or not this represented the individual's initial exposure to the medication. We therefore decided to eliminate from the dataset those patients who were prescribed one of the five drugs of interest during the first three months of the study. That is, we only considered those patients who were initially prescribed medication from April 1, 2000 onwards. This results in the elimination of 921 patients records, reducing the overall dataset to 136,134 individuals.

Thus, the time period under consideration is from April 1, 2000 to June 30, 2009, and encompasses 3378 days in total. Tables 21 and Table 22 summarize the mean time since exposure to Actos, Avandia, Insulin, Metformin and Sulfonylurea for HF and AMI cases respectively.

Table 21: Summary of time since exposure for HF cases for different drugs

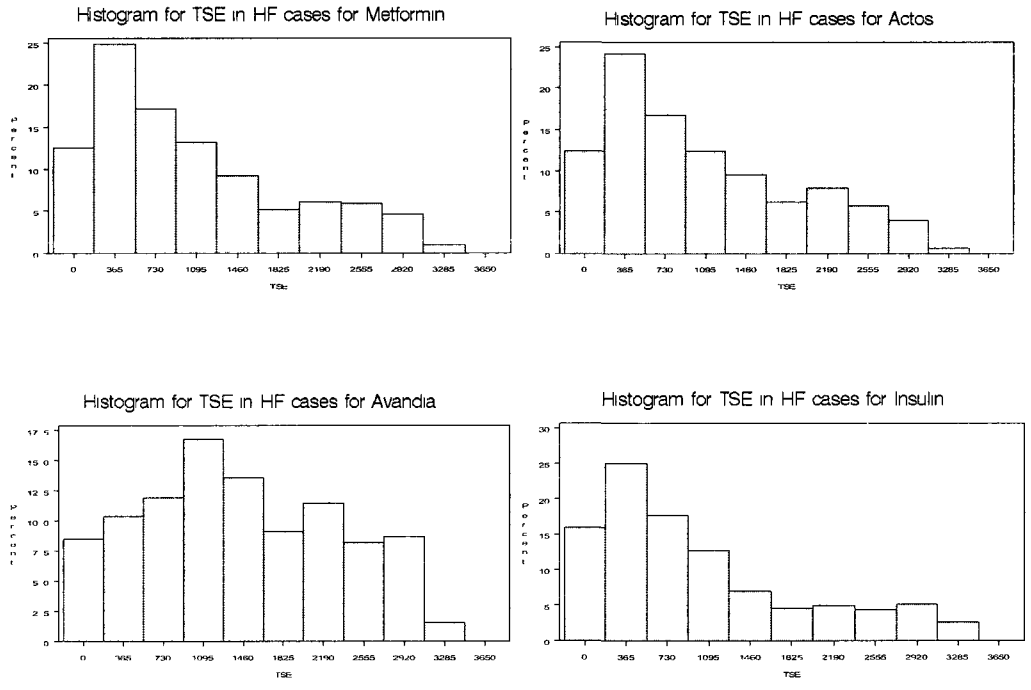
Drug level	N obs	Mean(days)
MET	15272	1050
AC	3397	1064
AV	2075	1437
INS	96564	1009
SUL	18826	1305

Table 22: Summary of time since exposure for AMI cases for different drugs

Drug level	N obs	Mean(days)
MET	15272	1095
AC	3397	1054
AV	2075	1556
INS	96564	1123
SUL	18826	1442

Histograms of TSE for each of the five drugs are presented for HF cases in Figure 1, while Figure 2 shows the analogous histograms for AMI cases. The largest number of the 136,134 patients under investigation was prescribed Insulin, while individuals taking Avandia were the least common. The tables also suggest that despite the drastic difference in the number of patients taking the various drugs, the mean TSE was roughly the same for each drug for both the HF and AMI cases.

Figure 1: Histograms of time since exposure in HF cases



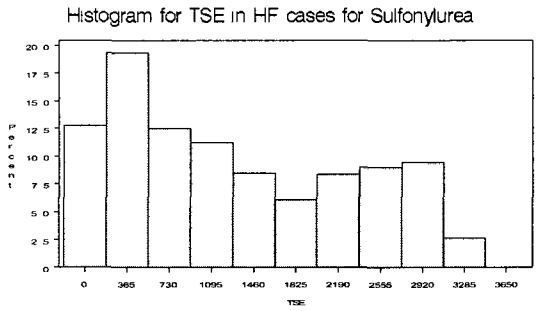
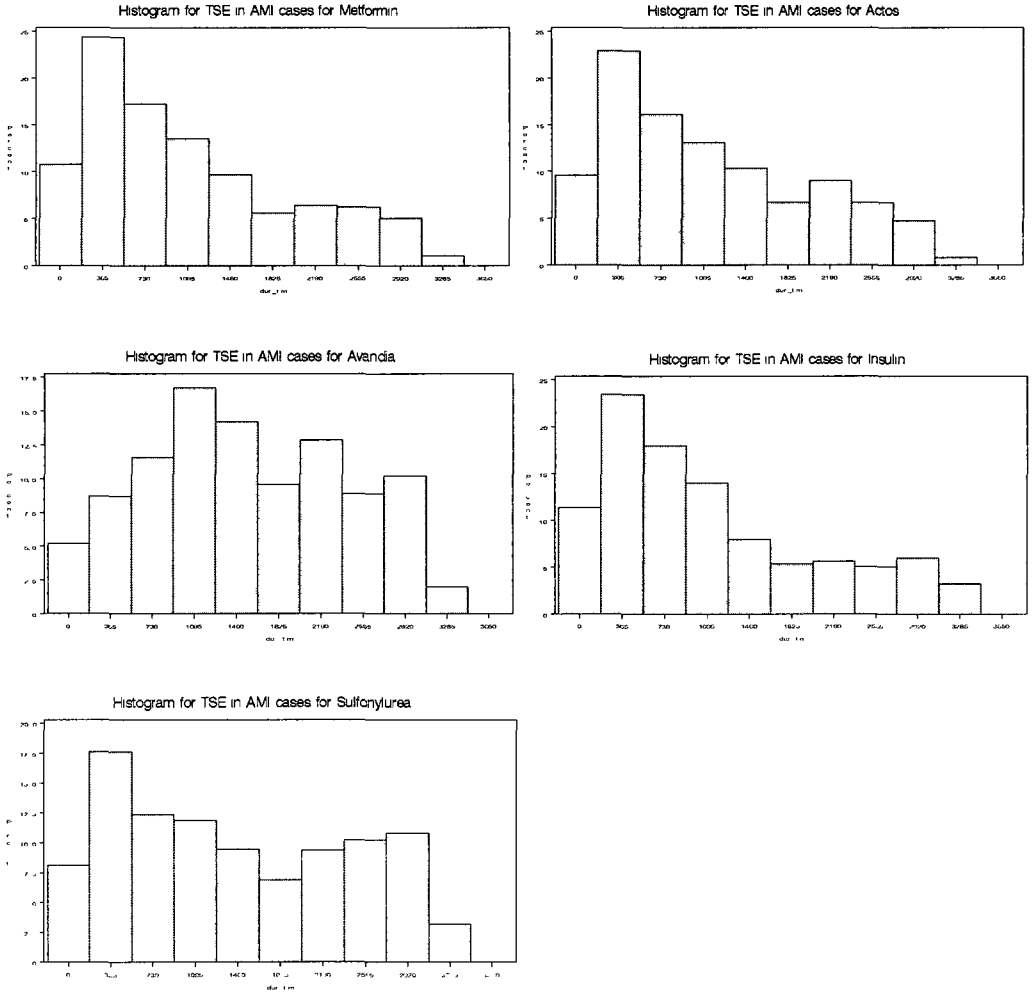


Figure 2: Histograms of time since exposure in AMI cases



In addition, by comparing the five histograms of TSE in Figure 1 for HF cases, we can see that the shapes of TSE distributions for Metformin, Actos and Insulin are

similar. For example, they are all right skewed. The histograms of TSE in Figure 2 for AMI cases are similar to those in Figure 1.

Before assessing the effect of TSE and a drug/TSE interaction on the occurrence of HF or AMI, we first decided to check if there was any impact on the relationship of the occurrence of these ADRs and the five drugs when patients that were prescribed medication in the first three months of the study were removed from the data set. For HF, again Y_i is a dependent variable that equals one if the ADR was observed on patient i during the study period and zero otherwise, while X_1, X_2, X_3, X_4, defined by Table 4, are used as indicator variables to identify the drug prescribed. We use the logistic regression model in equation (2) to find the maximum likelihood estimate of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^T$, which is presented in Table 23.

Table 23: MLE for HF cases only with indicator variables after removing the first 3 months of data

Parameter	DF	Estimate	Standard error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-2.8066	0.0349	6462.7334	<.0001
AC	1	0.5259	0.0687	58.6140	<.0001
AV	1	0.7819	0.0768	103.6692	<.0001
INS	1	1.2139	0.0360	1140.0854	<.0001
SUL	1	1.0549	0.0405	678.3130	<.0001

Table 24: Odds ratio estimates for HF cases only with indicator variables after removing the first 3 months of data

Effect	Point Estimate	95% Wald Confidence limits	
AC	1.692	1.479	1.936
AV	2.186	1.880	2.541
INS	3.367	3.138	3.613
SUL	2.872	2.652	3.109

Comparing the results in Table 24 with the results in Table 9, we can see that the results obtained after eliminating the 921 patients whose first drug records are in the first three months are very close to the results obtained using the original data. Using these parameter estimates, we determined estimates and 95% confidence intervals for the ratios of the odds of HF under each drug relative to Metformin. Those results, given in Table 24, are also similar to those obtained when the complete data set from January 1, 2000 to June 30, 2009 was used. Similar conclusions are arrived at when AMI is considered as the ADR. Tables 25 and 26 summarize the analogous parameter and odds ratio estimates respectively when patients initially prescribed a drug during the first three months of the study were removed from the analysis. These results are very similar to those in Tables 8 and 10 respectively.

Table 25: MLE for AMI cases with only indicator variables after removing the first 3 months of data

Parameter	DF	Estimate	Standard error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-4.6133	0.0821	3161.1020	<.0001
AC	1	0.6282	0.1522	17.0399	<.0001
AV	1	0.9910	0.1605	38.1418	<.0001
INS	1	1.3646	0.0838	265.2554	<.0001
SUL	1	1.1980	0.0920	169.7001	<.0001

Table 26: Odds ratio estimates for AMI cases only with indicator variables after removing the first 3 months of data

Effect	Point Estimate	95% Wald Confidence limits	
AC	1.874	1.391	2.526
AV	2.694	1.967	3.689
INS	3.914	3.321	4.613
SUL	3.313	2.767	3.968

Thus, regardless of the ADR under consideration, the removal of the first three months of data has a negligible effect on the ADR/drug relationship.

Thus, we now consider the logistic regression model to explore the relationship between the occurrence of the two ADRs of interest (HF and AMI) and TSE after eliminating patients with an initial prescription date within the first three months of the study. Again Y_i is the dependent variable, taking one when the i-th individual experiences the ADR and zero otherwise. In exploring the relationship between the occurrence of ADR and TSE using logistic regression, we include the indicator variables X_1, X_2, X_3, X_4 defined in Table 4 to identify which drug was prescribed for a particular patient. We also include a variable X_5 to represent TSE. In order to acknowledge the possibility that the relationship between the occurrence of a particular ADR and TSE is different for different drugs, we include interaction terms in the model as well. Specifically, we define X_6, X_7, X_8, X_9 as the interactions of indicator variables reflecting the five drugs of interest of X_1, X_2, X_3, X_4 with X_5. For example, X_6 is the interaction of X_1 and X_5, given by $X_6 = X_1 \times X_5$. Note that X_6 is equivalent to the TSE for a patient who took Actos, and zero when the patient took another drug. The other interaction terms, $X_7 = X_2 \times X_5$, $X_8 = X_3 \times X_5$ and $X_9 = X_4 \times X_5$ can be interpreted similarly. Finally, we also decided to include variables in the model for age and gender, reflected by X_{10} and X_{11} respectively. As a result, the vector of explanatory variables augmented by the constant one for the i-th patient in our model is $\mathbf{x_i} = (1, x_{i1}, x_{i2}, \cdots, x_{i11})^T$. We introduce a parameter vector

$\boldsymbol{\beta}^{\mathrm{T}} = (\beta_0, \beta_1, \cdots, \beta_{11})$. By using the logistic regression model of equation (2), we can get the maximum likelihood estimates of $\widehat{\boldsymbol{\beta}}$ for HF and AMI cases independently. Tables 27 and 28 summarize the results.

Table 27: MLE for HF cases with TSE in different drugs with age and gender

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-1.6463	0.0630	683.8380	<0.001
AC	1	0.5497	0.1050	27.3823	<0.001
AV	1	1.8771	0.1335	197.7828	<0.001
INS	1	1.0593	0.0544	378.6847	<0.001
SUL	1	1.3711	0.0619	489.9997	<0.001
TSE	1	-0.00183	0.000091	406.6089	<0.001
AC_TSE	1	0.000062	0.000170	0.1315	0.7169
AV_TSE	1	-0.00080	0.000193	17.0512	<0.001
INS_TSE	1	0.000331	0.000093	12.7737	0.0004
SUL_TSE	1	-0.00006	0.000102	0.3714	0.5422
AGE	1	0.000569	0.000517	1.2103	0.2713
GENDER	1	0.0229	0.0161	2.0143	0.1558

Table 28: MLE for AMI cases with TSE in different drugs with age and gender

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-3.2023	0.1408	517.2949	<0.001
AC	1	0.6095	0.2305	6.9879	0.0082
AV	1	2.1102	0.2452	74.0775	<0.001
INS	1	1.2281	0.1267	93.8847	<0.001
SUL	1	1.3966	0.1388	101.2516	<0.001
TSE	1	-0.00254	0.000270	88.5948	<0.001
AC_TSE	1	0.000284	0.000459	0.3842	0.5354
AV_TSE	1	-0.00081	0.000488	2.7368	0.0981
INS_TSE	1	0.000428	0.000274	2.4445	0.1179
SUL_TSE	1	0.000361	0.000290	1.5533	0.2126
AGE	1	0.000868	0.00100	0.7492	0.3867
GENDER	1	0.0696	0.0311	4.9912	0.0255

Abbreviation: AC_TSE is the interaction of Actos and TSE; AV_TSE is the interaction of Avandia and TSE; INS_TSE is the interaction of Insulin and TSE; SUL_TSE is the

interaction of Sulfonylurea and TSE.

Consider the model fit to the HF cases first. To assess its fit, we can use the likelihood ratio test statistic, $G^2 = 17032.3074$, which follows a chi-square distribution with 11 degrees of freedom. The corresponding P-value is less than 0.0001 which indicates that the model fits the data well. Similarly the likelihood ratio test statistic for the model fit to the AMI cases is $G^2 = 5620.8795$; the P-value is also less than 0.0001, suggesting that this model also fit very well.

Now we want to test the effect of age and gender simultaneously in both of these models. We will therefore consider analogous logistic regression models without age and gender. For HF cases, the likelihood ratio test statistic for the model with age and gender is $G_F{}^2 = 17032.3074$, which follows a chi-square distribution with 11 degrees of freedom. Fitting a model without age and gender (See the output in Table 29) yields a likelihood ratio test statistic $G_R{}^2 = 17029.2431$ which follows a chi-square distribution with 9 degrees of freedom. Then $\Delta G^2 = G_F{}^2 - G_R{}^2 = 17032.3074 - 17029.2431 = 3.0643$ follows a chi-square distribution with degrees of freedom $\Delta df = 11 - 9 = 2$. Thus, the associated P-value $= P(\chi_2^2 \geq 3.0643) = 0.2160$. The large P-value suggests that the simultaneous effects of age and gender are not significant in the model. In the same manner, we can calculate the P-value associated with testing the significance of age and gender for AMI cases. Output is given in Table 30. Here we obtain P-value $= P(\chi_2^2 \geq 5.543) = 0.0625$.

Since the P-value is also larger than 0.05, we conclude that the simultaneous effects of age and gender in AMI cases are not very strong. We will examine this further in what follows.

Table 29: MLE in HF case with TSE in different drugs without age and gender

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-1.5995	0.0528	917.2819	<0.001
AC	1	0.5518	0.1050	27.5948	<0.001
AV	1	1.8779	0.1335	197.9605	<0.001
INS	1	1.0598	0.0544	379.0628	<0.001
SUL	1	1.3741	0.0619	492.8050	<0.001
TSE	1	-0.00183	0.000091	406.5806	<0.001
AC_TSE	1	0.000061	0.000170	0.1294	0.7169
AV_TSE	1	-0.00080	0.000193	17.0314	<0.001
INS_TSE	1	0.000331	0.000093	12.7693	0.0004
SUL_TSE	1	-0.00006	0.000102	0.3710	0.5424

Table 30: MLE for AMI cases with TSE in different drugs without age and gender

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-3.1147	0.1241	630.0491	<0.001
AC	1	0.6141	0.2306	7.0940	0.0077
AV	1	2.1081	0.2451	73.9741	<0.001
INS	1	1.22961	0.1268	94.0940	<0.001
SUL	1	1.4021	0.1387	102.1207	<0.001
TSE	1	-0.00254	0.000270	88.5721	<0.001
AC_TSE	1	0.000284	0.000459	0.3821	0.5365
AV_TSE	1	-0.00080	0.000487	2.6971	0.1005
INS_TSE	1	0.000429	0.000274	2.4445	0.1179
SUL_TSE	1	0.000362	0.000290	1.5594	0.2117

Specifically, comparing the results in Table 29 with those in Table 27, and those in Table 30 with counterparts in Table 28, we find that the results in the two sets of tables are very close to each other, further confirming that the simultaneous effects of

age and gender for both HF and AMI cases are not very strong. We therefore decided to exclude the predictor variables of age and gender in the further analysis of the ADR/TSE relationship below.

By checking the P-values for the drug/TSE interaction terms in Table 29 for HF cases, we find that those for the parameters AC_TSE (β_6) and SUL_TSE (β_9) are 0.7169 and 0.5424 respectively, which are both much larger than 0.05. However, we are more interested in assessing the significance of the drug/TSE interaction simultaneously across all drugs. That is, we want to test the hypothesis

$$H_0: \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0 \text{ vs } H_a: \text{At least one of the terms not equal to zero} \quad (37)$$

In order to conduct the test, we consider the model fit for the HF cases with the indicator variables for the five different drugs, the variable TSE, and the drug/TSE interaction terms, and compare it with the model containing only the indicator variables and TSE (the interaction terms are excluded). The likelihood ratio test statistic for the model with the interaction terms is $G_F^2 = 17029.2431$, and it has 9 degrees of freedom. For the model without the interaction terms, the likelihood ratio test statistic is $G_R^2 = 16904.9645$, and it has 5 degrees of freedom. Thus, $\Delta G^2 = G_F^2 - G_R^2 = 17029.2431 - 16904.9645 = 124.2786$ follows a chi-square distribution with 4 degrees of freedom ($\Delta df = 9 - 5 = 4$). The associated P-value is $P(\chi_4^2 \geq 124.2786)$, which is almost equal to zero. It suggests that the simultaneous effects of the drug/TSE interaction terms are very strong for HF cases. Similarly, we

can test the significance of the simultaneous effects of the interaction terms for AMI cases. The associated P-value is calculated as $P(\chi_4^2 \geq 5615.3365 - 5601.3534) = P(\chi_4^2 \geq 13.9831) = 0.0073$, which indicates that the set of drug/TSE interaction terms in the model for AMI cases are also very significant.

Thus, the drug/TSE interactions are significant in each of the models for HF and AMI. Therefore, under these models, the ratio of the odds of the occurrence of a particular ADR for a given drug relative to Metformin is actually a function of TSE. We will illustrate this with the following example. Suppose we wish to estimate the odds of HF for a patient taking Actos over 500 days to the odds of HF for a patient taking Metformin for 500 days. The odds of HF for a patient taking Actos for 500 days is

$$e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}+\beta_5 x_{i5}+\beta_6 x_{i6}+\beta_7 x_{i7}+\beta_8 x_{i8}+\beta_9 x_{i9}}$$

$$= e^{\beta_0+\beta_1(1)+\beta_2(0)+\beta_3(0)+\beta_4(0)+\beta_5(500)+\beta_6(500)+\beta_7(0)+\beta_8(0)+\beta_9(0)}$$

$$= e^{\beta_0+\beta_1+500\beta_5+500\beta_6}.$$

Similarly, the odds of HF for a patient taking Metformin for 500 days is

$$e^{\beta_0+\beta_1 x_{j1}+\beta_2 x_{j2}+\beta_3 x_{j3}+\beta_4 x_{j4}+\beta_5 x_{j5}+\beta_6 x_{j6}+\beta_7 x_{j7}+\beta_8 x_{j8}+\beta_9 x_{j9}}$$

$$= e^{\beta_0+\beta_1(0)+\beta_2(0)+\beta_3(0)+\beta_4(0)+\beta_5(500)+\beta_6(0)+\beta_7(0)+\beta_8(0)+\beta_9(0)}$$

$$= e^{\beta_0+500\beta_5}.$$

Thus the ratio of these two odds is equal to

$$\frac{e^{\beta_0+\beta_1+500\beta_5+500\beta_6}}{e^{\beta_0+500\beta_5}} = e^{\beta_1+500\beta_6}.$$

Note that the result includes the value 500, which reflects the TSE. Thus, if we wish

to conduct a similar comparison for patients with a different TSE, the odds ratio would change. Returning to the example, we can compute a point estimate for the odds ratio $e^{\beta_1+500\beta_6}$ using the output in Table 29 to calculate $e^{\hat{\beta}_1+500\hat{\beta}_6} = e^{0.5518+500(0.000061)} = 1.790$. Thus, when the TSE is 500 days, the odds of experiencing HF for patients who took Actos is estimated to be 1.790 times the odds of experiencing HF for patients who took Metformin. If we wish to determine a confidence interval (say 95%) for this odds ratio, we would require a measure of the variability in $e^{\hat{\beta}_1+500\hat{\beta}_6}$. While this is not available from the logistic regression model fit, we do have an asymptotic estimated covariance matrix for $\hat{\boldsymbol{\beta}}$. Thus, we can find $\widehat{\text{var}}\{\hat{\beta}_1 + 500\hat{\beta}_6\} = \text{var}(\hat{\beta}_1) + (500)^2\text{var}(\hat{\beta}_6) + 2(500)\text{cov}(\hat{\beta}_1, \hat{\beta}_6)$ to determine a confidence interval for $\beta_1 + 500\beta_6$, and transform it to an interval for $e^{\beta_1+500\beta_6}$. For our example, to compute a 95% confidence interval for the odds ratio represented by $e^{\beta_1+500\beta_6}$, we therefore proceed as follows:

$\hat{\beta}_1 + 500\hat{\beta}_6 = 0.5518 + 500(0.000061) = 0.5823$

$\widehat{\text{var}}\{\hat{\beta}_1 + 500\hat{\beta}_6\} = 0.011033 + (500)^2(2.894e^{-8}) + 2(500)(-0.00001)$

$= 0.008268$

95% C.I. for $\beta_1 + 500\beta_6$: $0.5823 \pm 1.96(0.008268)$ or $(0.5661, 0.5985)$.

95% C.I. for $e^{\beta_1+500\beta_6}$: $(e^{0.5661}, e^{0.5985})$ or $(1.7914, 1.8194)$

Thus, the 95% confidence interval for the odds of experiencing heart failure for patients who consumed Actos is anywhere from 1.7614 to 1.8194 times the odds of experiencing heart failure for patients who consumed Metformin.

Note that we can perform the above estimation of odds ratios for comparison of Metformin to other drugs, as well as for different values of TSE. For the occurrence of HF, Table 31(a) through Table 31(d) present odds ratio estimates and 95% confidence intervals for each drug relative to Metformin. For each table, we chose values for TSE ranging from one to nine years (we assume one year is simply equal to 365 days, and nine years equals 3285 days, for example) to coincide with the time period of the data set. Table 32(a) to 32(d) present similar results for AMI cases.

Table 31 (a): Actos-Metformin comparison on TSE in HF cases

Years	TSE(days)	OddsEst	95% Lower CI	95% Upper CI
1	365	1.775469686	1.496793468	2.106030441
2	730	1.81544388	1.466564742	2.247317414
3	1095	1.856318082	1.371647737	2.512246204
4	1460	1.898112554	1.26108008	2.856940909
5	1825	1.940848018	1.151604746	3.270992968
6	2190	1.984545657	1.048293407	3.756983914
7	2555	2.029227137	0.952614941	4.322588905
8	2920	2.074914608	0.864785002	4.978428883
9	3285	2.121630719	0.784538656	5.737533613

Table 31 (b): Avandia-Metformin comparison on TSE in HF cases

Years	TSE(days)	OddsEst	95% Lower CI	95% Upper CI
1	365	4.883684717	4.090191365	5.831115048
2	730	3.646982087	3.044220207	4.369092061
3	1095	2.723451474	2.082409848	3.561829068
4	1460	2.03378787	1.381422818	2.99422671
5	1825	1.51876879	0.907335399	2.54223371
6	2190	1.134168746	0.59353548	2.167248274
7	2555	0.846961534	0.387487096	1.851271555
8	2920	0.632484225	0.252685869	1.583136785
9	3285	0.472319318	0.164666646	1.354770646

Table 31 (c): Insulin-Metformin comparison on TSE in HF cases

Years	TSE(days)	OddsEst	95% Lower CI	95% Upper CI
1	365	3.256376259	3.026172226	3.504092149
2	730	3.674547515	3.35713577	4.021970026
3	1095	4.146418709	3.602902883	4.771926601
4	1460	4.678885778	3.829643173	5.716452195
5	1825	5.279730211	4.05787163	6.869500477
6	2190	5.957732764	4.293957025	8.266170221
7	2555	6.722801785	4.540686149	9.953575817
8	2920	7.586118014	4.799715021	11.99012572
9	3285	8.560297977	5.07227968	14.44689688

Table 31 (d): Sulfonylurea-Metformin comparison on TSE in HF cases

Years	TSE(days)	OddsEst	95% Lower CI	95% Upper CI
1	365	3.865921209	3.552708995	4.206746687
2	730	3.782177871	3.419002852	4.183930247
3	1095	3.700248576	3.170445318	4.318585608
4	1460	3.620094028	2.906503804	4.508881341
5	1825	3.541675782	2.654418175	4.725505374
6	2190	3.464956228	2.420267879	4.960575549
7	2555	3.389898568	2.204951125	5.211640371
8	2920	3.316466802	2.007837923	5.478007923
9	3285	3.24462571	1.827804259	5.759695519

Relative to Metformin, it would appear that the odds of experiencing HF for a patient taking Actos increase slightly the longer that the individual has been taken the drug. For a patient taking Actos for 365 days, the odds of experiencing HF is estimated to be 1.775 times the odds for a patient taking Metformin for the same length of time. These odds increase slightly to 2.122 times if a patient has been taking medication for nine years.

By contrast, relative to Metformin, the odds of HF decrease slightly for a patient taking Sulfonylurea as TSE increases. For a patient taking this drug for one year, the odds of HF is approximately 3.866 times the odds for an individual taking Metfomin. This drops slightly to 3.245 for a TSE of nine years. Of note, however, is that despite the drug, patient taking Sulfonylurea would appear to be at far greater risk still after nine years of taking the drug.

The same pattern observed when comparing Actos to Metformin can be seen when contrasting the odds of HF when taking Insulin to the odds when taking Metformin. However, the odds under Insulin relative to Metformin are significantly higher than counterpart odds when Actos was compared to Metformin. In addition, the increase in odds for the Insulin/Metformin comparison is also substantially more noticeable as TSE increases, jumping from 3.256 from one year since first exposure to 8.5603 after nine years.

The odds that a patient taking Avandia for one year will experience HF is estimated to be 4.884 times the odds that a patient taking Metformin for the same length of time will experience HF. Interestingly, these odds decrease the longer the patients are taking the drugs to the point where, after 7 years, the point estimate for the odds ratio is less than one.

Table 32 (a): Actos-Metformin comparison on TSE in AMI cases

Years	TSE(days)	OddsEst	95% Lower CI	95% Upper CI
1	365	2.049836431	1.524846671	2.755575018
2	730	2.273726238	1.476791713	3.500717778
3	1095	2.522070018	1.24329454	5.116114459
4	1460	2.797538714	1.014086956	7.717506684
5	1825	3.103094998	0.818970402	11.75768812
6	2190	3.442025134	0.658611196	17.9886663
7	2555	3.817974323	0.528507641	27.58130025
8	2920	4.234985906	0.423575269	42.34219258
9	3285	4.697544851	0.339209838	65.05391407

Table 32 (b): Avandia-Metformin comparison on TSE in AMI cases

Years	TSE(days)	OddsEst	95% Lower CI	95% Upper CI
1	365	6.147835078	4.481494444	8.433766151
2	730	4.591009801	2.89631284	7.277311587
3	1095	3.428421667	1.614013999	7.282511266
4	1460	2.560237429	0.869588218	7.537838666
5	1825	1.911904757	0.463566991	7.885332371
6	2190	1.427750316	0.24600775	8.286206297
7	2555	1.066199013	0.130250263	8.72766248
8	2920	0.796203876	0.068869372	9.204971614
9	3285	0.594580003	0.036383727	9.716579538

Table 32 (c): Insulin-Metformin comparison on TSE in AMI cases

Years	TSE(days)	OddsEst	95% Lower CI	95% Upper CI
1	365	3.999562579	3.523757939	4.539613987
2	730	4.677529098	3.763691893	5.813249087
3	1095	5.470417834	3.688467091	8.113254244
4	1460	6.397709272	3.569950384	11.46533692
5	1825	7.482186035	3.443513568	16.25755402
6	2190	8.750492636	3.316985713	23.08454965
7	2555	10.2337901	3.192906812	32.80097602
8	2920	11.96852156	3.072261765	46.62542431
9	3285	13.99730764	2.955450893	66.29263297

Table 32 (d): Sulfonylurea-Metformin comparison on TSE in AMI cases

Years	TSE(days)	OddsEst	95% Lower CI	95% Upper CI
1	365	3.999562579	3.523757939	4.539613987
2	730	4.677529098	3.763691893	5.813249087
3	1095	5.470417834	3.688467091	8.113254244
4	1460	6.397709272	3.569950384	11.46533692
5	1825	7.482186035	3.443513568	16.25755402
6	2190	8.750492636	3.316985713	23.08454965
7	2555	10.2337901	3.192906812	32.80097602
8	2920	11.96852156	3.072261765	46.62542431
9	3285	13.99730764	2.955450893	66.29263297

For AMI cases, the Avandia-Metformin comparison is very similar to HF cases. In addition, the comparisons of Actos-Metformin and Insulin-Metformin are quite similar to HF cases, except that the odds of experiencing AMI increase faster as TSE increases. The biggest difference is in Sulfonylurea-Metformin comparison. In HF cases, the odds of experiencing HF decrease slightly for a patient taking Sulfonylurea as TSE increases. However, in AMI cases, the odds of experiencing AMI increase significantly for a patient taking Sulfonylurea as the TSE increases, jumping from 4.00 for one year since first exposure to 14.00 after nine years.

3.7 Effect of Cumulative Dosage using Logistic Model

In Section 3.6, we investigated whether the time since first exposure (TSE) to a drug affects the occurrence of HF and AMI. We concluded that this was in fact the case, and that for the five drugs under consideration in our study, the relationship between a particular ADR and TSE was different for different drugs. In this section, we conduct

a similar investigation using standard logistic regression, focusing on cumulative dosage of a drug on each patient over the entire study rather than TSE. Note that we were not able to consider the effect of intensity of exposure since the time interval covered by a given dose is unknown.

As before when we studied the effect of TSE to investigate the effect of cumulative dose, we again remove patients who were initially prescribed a drug in the first three months of the study, leaving 136,134 individuals out of the original 137,047. In addition, there were 24,902 patient records where no information on dosage was available, further reducing the applicable data set to a total of 111,232 individuals. The cumulative dosage for each patient is determined by adding all doses of the particular drug prescribed for the patient over the period from April 1, 2000 to June 30, 2009.

For HF and AMI cases respectively, Tables 33 and 34 summarize the mean cumulative doses for patients taking one of the five drugs under study for the period from April 1, 2000 to June 30, 2009. As can be seen from the tables, unlike the mean TSE values in Tables 21 and 22, the mean cumulative doses are extremely different for the various drugs.

Table 33: Summary of cumulative dosage for HF cases for different drugs

Drug level	N obs	Mean(mg)
MET	15236	3354.21
AC	3392	144.71
AV	2069	27.32
INS	71987	73107.14
SUL	18548	41.61

Table 34: Summary of cumulative dosage for AMI cases for different drugs

Drug level	N obs	Mean(mg)
MET	15236	3354.21
AC	3392	147.11
AV	2069	28.92
INS	71987	73159.37
SUL	18548	42.68

Similar to the case where the effect of TSE was investigated, before considering the effect of cumulative doses, we first check the impact of removing the first three months of data and the patients for which there is no dosage information. Fitting logistic regression models on the remaining data in which only the four indicator variables for the drugs of interest are included, we obtained the maximum likelihood estimates in Table 35 and 36 when the response was the occurrence of HF and AMI, respectively.

Table 35: MLE for HF cases with indicator variables only after removing useless data

Parameter	DF	Estimate	Standard error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-2.8078	0.0350	6446.2001	<.0001
AC	1	0.5181	0.0689	56.4794	<.0001
AV	1	0.7863	0.0768	104.7444	<.0001
INS	1	1.2936	0.0363	1270.7074	<.0001
SUL	1	1.0612	0.0406	682.7695	<.0001

Table 36: MLE for AMI cases with indicator variables only after removing useless data

Parameter	DF	Estimate	Standard error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-4.6109	0.0821	3157.7678	<.0001
AC	1	0.6273	0.1522	16.9915	<.0001
AV	1	0.9915	0.1605	38.1846	<.0001
INS	1	1.4315	0.0842	288.8923	<.0001
SUL	1	1.1985	0.0921	169.4444	<.0001

Comparing the values in Table 35 with those in Table 7, it can be seen that for HF cases, the results obtained after removal of some of the patient records as described above does not dramatically affect the parameter estimates. A similar conclusion can be drawn by comparing the results for AMI cases in Table 36 with those in Table 8. The same is true of the estimates and 95% confidence intervals for the odds ratios of HF and AMI under each drug relative to Metformin. This is verified by comparing the results in Tables 37 and 38 below to those in Tables 9 and 10 respectively.

Table 37: Odds ratio estimates for HF cases with only indicator variables after removing useless data

Effect	Point Estimate	95% Wald Confidence limits	
AC	1.679	1.467	1.922
AV	2.195	1.888	2.552
INS	3.646	3.396	3.915
SUL	2.890	2.669	3.129

Table 38: Odds ratio estimates for AMI cases with only indicator variables after removing useless data

Effect	Point Estimate	95% Wald Confidence limits	
AC	1.873	1.390	2.523
AV	2.695	1.968	3.692
INS	4.185	3.548	4.936
SUL	3.315	2.768	3.971

Now we consider the standard logistic regression model to explore the relationship between the occurrence of the two ADRs of interest (HF and AMI) and cumulative dose. We use the data set consisting of 111,232 patients described earlier in this section. Again, Y_i is the dependent variable, taking on one when the i-th individual experiences the ADR and zero otherwise. In exploring the relationship between the occurrence of the ADR and cumulative dose using the logistic regression, we include the indicator variables defined in Table 4 to identify which drug was prescribed for a particular patient. We also include a variable X_5 to represent the cumulative dosage. Similar to the model we considered in Section 3.6, we also include drug/cumulative dose interaction terms to acknowledge the possibility that the relationship between the occurrence of a particular ADR and cumulative dose is different for different drugs. The variable X_6 is the interaction of X_1 and X_5, given by $X_6 = X_1 \times X_5$. Note that X_6 is equivalent to the cumulative dose for a patient who took Actos and zero when the patient took other drugs. The other interaction terms, $X_7 = X_2 \times X_5$, $X_8 = X_3 \times X_5$ and $X_9 = X_4 \times X_5$ can be interpreted similarly. Given the fact that the effects of age and gender in the models considered to date have not been very strong (including

those models that only contain indicator variables for different drugs), we decided not to include these two variables in the model. As a result, the vector of explanatory variables for the i-th patient is the form of $\mathbf{x_i} = (1, x_{i1}, x_{i2}, \cdots, x_{i9})^T$ with associated parameter vector $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \cdots, \beta_9)$. Fitting two separate models to the HF and AMI cases yield the results in Tables 39 and 40 respectively.

Table 39: MLE for HF cases with cumulative dosage in different drugs

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-2.8189	0.0357	6248.3006	<0.001
AC	1	0.5180	0.0698	55.0959	<0.001
AV	1	0.7882	0.0782	101.5320	<0.001
INS	1	1.3044	0.0370	1245.8398	<0.001
SUL	1	1.0731	0.0413	675.7324	<0.001
CUMDOSE	1	2.938E-6	1.635E-6	3.2301	0.0723
AC_CUMDOSE	1	0.000061	0.000040	2.3526	0.1251
AV_CUMDOSE	1	0.000305	0.000392	0.6048	0.4367
INS_CUMDOSE	1	-2.93E-6	1.635E-6	3.2207	0.0727
SUL_CUMDOSE	1	-0.00002	0.00061	0.1282	0.7203

Table 40: MLE for AMI cases with cumulative dosage in different drugs

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-4.6159	0.0834	3061.7089	<0.001
AC	1	0.7085	0.1687	17.6454	<0.001
AV	1	1.0471	0.1677	39.0007	<0.001
INS	1	1.4364	0.0856	281.8788	<0.001
SUL	1	1.2052	0.0935	166.2482	<0.001
CUMDOSE	1	1.368E-6	3.851E-6	0.1262	0.7224
AC_CUMDOSE	1	-0.00106	0.00128	0.6837	0.4083
AV_CUMDOSE	1	-0.00339	0.00428	0.5929	0.4413
INS_CUMDOSE	1	-1.37E-6	3.87E-6	0.1261	0.7225
SUL_CUMDOSE	1	-0.00004	0.00015	0.0814	0.7754

Abbreviation: CUMDOSE means cumulative dose; AC_CUMDOSE is the interaction between Actos and cumulative dose; AC_CUMDOSE is the interaction between

Avandia and cumulative dose; INS_CUMDOSE is the interaction between Insulin and cumulative dose; SUL_CUMDOSE is the interaction between Sulfonylurea and cumulative dose.

Consider the model fit for HF cases first. To assess its fit, we can use $G^2 = 1917.7164$, which follows a chi-square distribution with 9 degrees of freedom. The corresponding P-value is less than 0.0001. Thus, we conclude that our model fits the data very well. Similarly, the likelihood ratio test statistic for the model fit to the AMI cases is $G^2 = 488.3315$; the P-value is also less than 0.0001, suggesting that the model also fits very well.

The P-value associated with the drug/cumulative dose interaction terms for both the HF and AMI model fits summarized in Table 39 and 40 are quite large. Of interest is to test whether these interaction terms are significant in these two models. Regarding HF cases, the likelihood ratio test statistic for the model with interaction terms is $G_F{}^2 = 1917.7164$; it has 9 degrees of freedom. We compare it to a model fit with only the indicator variables for drugs and the variable cumulative dosage. The test statistic for this model without interaction terms is $G_R{}^2 = 1912.1178$; it has 5 degrees of freedom. Thus $\Delta G^2 = G_F{}^2 - G_R{}^2 = 5.5986$; it follows a chi-square distribution with 4 degrees of freedom $(\Delta df = 9 - 5 = 4)$. As a result, the P-value is $P(\chi_4^2 \geq 5.5986) = 0.2311$, which is larger than 0.05. Thus the drug/cumulative dose interaction terms in the HF model are not significant. Performing an identical analysis

on the model for AMI cases leads to a P-value $= P(\chi_4^2 \geq 488.3315 - 484.3506) = P(\chi_4^2 \geq 3.9809) = 0.4085$ for testing the significance of the drug/cumulative dose interaction. Thus, similar to the situation above for HF cases, the drug/cumulative dose interaction is not significant in the model for AMI cases.

Table41: MLE for HF cases with cumulative dosage excluding the interaction terms

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-2.8078	0.0350	6446.2656	<0.001
AC	1	0.5181	0.0698	56.4823	<0.001
AV	1	0.7863	0.0768	104.7482	<0.001
INS	1	1.2933	0.0363	1270.0709	<0.001
SUL	1	1.0612	0.0406	682.7877	<0.001
CUMDOSE	1	4.279E-9	2.478E-9	2.9821	0.0842

Table 42: MLE for AMI cases with cumulative dosage excluding the interaction terms

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	1	-4.6109	0.0821	3157.7718	<0.001
AC	1	0.6273	0.1522	16.9916	<0.001
AV	1	0.9915	0.1605	38.1848	<0.001
INS	1	1.4314	0.0842	288.8533	<0.001
SUL	1	1.1985	0.0921	169.4452	<0.001
CUMDOSE	1	8.63E-10	1.329E-9	0.4216	0.5161

Note that the models fit for the HF and AMI cases without the drug/cumulative dose interaction terms, and only the drug and cumulative dose main effects, are presented in Tables 41 and 42, respectively. Of further note from these tables is the fact that even the main effect for cumulative dose is not very strong in either model. Thus, the variable cumulative dose does not appear to significantly improve either the HF or AMI models provided that the drug main effects is accounted for.

Chapter 4
Conclusion

Data mining from a cohort study is the most comprehensive method in active pharmocovigilance. In the regard, the objective is to use effective statistical methods to detect signals of adverse drug reactions that warrant our attention.

In this thesis, we described four statistical models, specifically the logistic regression model, the logistic regression model with James-Stein shrinkage, the Cox model, and the Cox model with random effects, to explore the association between prescription drug usage and adverse health outcomes. The logistic regression model has been the standard method of analysis for this kind of investigation. Building on this model, the James-Stein shrinkage estimation method incorporates prior notions into logistic models; it is effective in detecting signals as it combines information and borrows strength across medically related adverse drug reactions. Cox models can be used to explore how the survival experience of adverse drug reaction patients depends on the values of one or more explanatory variables. In addition, random effects can be incorporated into the Cox models if the population shares a common, unobservable, random frailty. All four of these methods can avoid the confounding and masking

problems encountered in the disproportionality-based methods in passive pharmacovigilance.

In applying these methods to a diabetes data set in order to explore the effect of a particular set of drugs (Actos, Avandia, Metformin, Insulin and Sulfonylurea) on the occurrence of heart failure (HF) and acute myocardial infarction (AMI), we discovered that all four produced similar results. In particular we found that ignoring the time since first exposure, the Actos, Avandia, Insulin and Sulfonylurea treatments were associated with a significant increase in the risk of HF and AMI among patients with diabetes compared with those treated with Metformin. Compared with Metformin users, the patients who consumed Insulin appeared to have the highest increased risk of HF and AMI, the Sulfonylurea consumers experienced the second highest increased risk, while the patients who consumed Actos and Avandia have similar increased risk of experiencing HF and AMI, but both were lower than the Sulfonylurea users. It was also discovered that age and gender had little effect on these associations.

Since the four methodologies produced similar results when investigating the above drug/ADR relationships, we decided to extend the simplest of the four, the logistic regression model to explore the relationship between the ADRs (HF and AMI) and the time since first exposure (TSE) to different drugs. We also considered the relationship between ADRs and the cumulative dose of different drugs independently. We

discovered that TSE to different drugs had a significant effect for both HF and AMI cases, and that there is a significant drug/TSE interaction for both of these ADRs; however, there is not a significant relationship between the cumulative dose of different drugs and the ADRs of interest (HF and AMI).

The applications presented in this thesis only include a small number of covariates. A number of additional important prognostic factors exist that were not taken into account, such as exposure to other diabetic drugs, cardiovascular history, diabetes duration, etc. Future research could involve the incorporation of these covariates into the statistical models considered here.

References

[1] Ahmed, S.E. and Krzanowski, W.J. (2004), Biased estimation in a simple multivariate regression model: *Computational Statistics and Data Analysis*, v. 45, p. 689-696.

[2] Ahmed, S.E. and Saleh, AKMdE. (1999), Improved nonparametric estimation of location vectors in multivariate regression models: *Journal of Nonparametric Statistics*, v. 11, p. 52-78.

[3] An, L., Ahmed, S.E. and Ali, A. (2006), Tumor growth rate approximation-assisted estimation: *Cancer Informatics*, v. 2, p. 214-221.

[4] An, L., Fung, K. and Krewski, D. (2010), Mining pharmacovigilance data using Bayesian logistic regression with James-Stein type shrinkage estimation: *Journal of Biopharmaceutical Statistics*, v. 20, no. 5, p. 998-1012.

[5] An, L., Nkurunziza, S., Fung, K. and Krewski, D. (2009), Shrinkage estimation in general linear models: *Computational Statistics & Data Analysis*, v. 53, no. 7, p. 2537-2549.

[6] Breslow N.E. (1974), Covariance analysis of censored survival data: *Biometrics,* v. 30, p. 89-99.

[7] Brockwell, P.J. and Davis, R.A. (1991), *Time Series: Theory and Methods*: New York, Springer-Verlag.

[8] Casella, G. and Hwang, J.T. (1986), Confidence sets and the Stein-effect: *Communications in Statistics: theory and methods*, v. 15, p. 2043-2063.

[9] Casella, O., (2007), Mining the WHO Drug Safety Database Using Lasso Logistic Regression: Master Thesis, Uppsala University.

[10] Collett, D. (2003), *Modelling Survival Data in Medical Research*: New York, Chapman & Hall.

[11] Cox, D.R. (1972), Regression models and life tables (with discussion): *J. R. Statist. Soc.* B v. 55, p. 187-220.

[12] Cox, D.R. and Oakes, D. (1984), *Analysis of Survival Data*: New York, Chapman and Hall.

[13] Fung, K. and Krewski, D. (1999), Evaluation of regression calibration and SIMEX method in logistic regression when one of the predictor is subject to additive measurement error: *Journal of Epidemiology and Biostatistics*, v. 4, no. 2, p. 65-74.

[14] Genkin, G., Lewis, D. and Madigan, D. (2007), Large-scale Bayesian logistic regression for text categorization: *Technometrics*, v. 49, no. 3, p. 291-304.

[15] Gravel, C. (2009), Statistical Methods for Signal Detection in Pharmacovigilance: Master Thesis, School of Mathematics and Statistics, Carleton University, Ottawa, Canada.

[16] Harrell, F.E. and Lee, K.L. (1986), Proceeding of the eleventh annual SAS users group international, p. 823-828.

[17] Hauben, M. and Bete, A. (2009), Decision support methods for the detection of advrse events in post-marketing data: *Drug Discovery Today*, v. 14, no. 7/8, p.

343-357.

[18] Hosmer, D.W. and Lemeshow, S. (1989), *Applied Logistic Regression*: John Wiley & Sons, Inc.

[19] Kalbfleisch, J.D. and Prentice, R.L. (1980), *Statistical Analysis of Failure Time Data*: New York, Wiley.

[20] Krewski, D., Jerrett, M., Burnett, R.T., Ma, R., Hughes, E., Shi, Y., Turner, M.C., Pope, C.A., Thurston, G., Calle, E.E., and Thun, M.J. (2009), Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality: Boston, Massachusetts, Health Effects Institute, 140, p. 1-140.

[21] Lipscombe, L.L., Gomes, T., Levesque, L.E., Hux, J.E., Juurlink, D.N. and Alter, D.A. (2007), Thiazolidinediones and cardiovascular outcomes in older patients with diabetes: *Journal of the American Medical Association*, v. 298, no. 22, p. 2634-2643.

[22] Ma, R., Krewski, D. and Burnett, R.T. (2003), Random effects Cox models: A Poisson modelling approach: *Biometrika*, v. 90, no. 1, p. 157-169.

[23] McCullagh, P. and Nelder, J.A. (1984), *Generalized Linear Models*, New York: Chapman and Hall.

[24] Pike, M.C., Hill, A.P. and Smith, P.G. (1980), Bias and efficiency in logistic analysis of stratified case-control studies: *American Journal of Epidemiology*, v. 9, p.89-95.

[25] Saleh, AKMdE. (2006), *Theory of Preliminary Test and Stein-Type Estimation with Applications*: New Jersey, John Wiley & Sons, Inc..

[26] Sargent, D.J. (1998), A general framework for random effects survival analysis in Cox proportional hazards setting: *Biometrics*, v. 54, p. 1486-1497.

[27] Sastry, N. (1997), A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil: *J. Am. Statist. Assoc.*, v. 92, p. 426-435.

[28] Schoenfeld, D. (1982), Partial residuals for the proportional hazards regression model: *Biometrika*, v. 69, p. 239-241.

[29] Stein, C. (1956), Inadmissibility of the usual estimator of the mean of a multivariate normal distribution: *In Proceeding of the Third Bekeley Symposium on Mathematical Statistics and Probability*, University Of California Press: Berkeley, CA.

[30] Whitehead, J. (1980), Fitting Cox's regression model to survival data using GLIM: *Appl. Statist.*, v. 29, p. 268-275.

[31] World Health Organization (2007), *A practical handbook on the pharmacovigilance of antimalarial medicines*.

[32] Yau, K.K.W. (2001), Multi-level models for survival analysis with random effects: *Biometrics*, v. 57, p. 96-102.

Appendix A: Computer Programs

1) SAS Code and output for logistic regression models

```
/******* Logistic model for HF case ***********/
data hf_comb_time_re;
set diab.hf_comb_time_re;

proc logistic descending;
model HF = Actos Avand Ins Sul AGE gender / covb;
title 'Logistic model for Heart Failure Patients';
run;

/****** Logistic regression model for AMI case *******/
data mi_comb_time_re;
set diab.mi_comb_time_re;

proc logistic descending;
model MI = Actos Avand Ins Sul AGE gender / covb;
title 'Logistic model for Acute Myocardial Infaction Patients';
run;
```

SAS output for heart failure case using logistic regression models

Logistic model for Heart Failure Patients

The LOGISTIC Procedure

Model Information

Data Set	WORK.HF_COMB_TIME_RE
Response Variable	HF
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	137047
Number of Observations Used	137047

Response Profile

Ordered Value	HF	Total Frequency
1	1	20728
2	0	116327

Probability modeled is HF=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	116458.04	114759.87
SC	116467.87	114828.66
-2 Log L	116456.04	114745.87

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1710.1735	6	<.0001
Score	1427.0121	6	<.0001
Wald	1299.9874	6	<.0001

Logistic model for Heart Failure Patients

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.8585	0.0470	3694.9305	<.0001
Actos	1	0.5237	0.0687	58.1462	<.0001
Avand	1	0.7805	0.0768	103.3343	<.0001
Ins	1	1.2167	0.0359	1147.0770	<.0001
Sul	1	1.0490	0.0405	669.8200	<.0001
Age	1	0.000935	0.000486	3.7020	0.0543
Gender	1	-0.0136	0.0152	0.7963	0.3722

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Actos	1.688	1.476	1.932
Avand	2.183	1.878	2.537
Ins	3.376	3.146	3.622
Sul	2.855	2.637	3.091
Age	1.001	1.000	1.002
Gender	0.987	0.958	1.016

Association of Predicted Probabilities and Observed Responses

Percent Concordant	48.5	Somers' D	0.119
Percent Discordant	36.7	Gamma	0.139
Percent Tied	14.8	Tau-a	0.030
Pairs	2411226056	c	0.559

Estimated Covariance Matrix

Parameter	Intercept	Actos	Avand	Ins	Sul	AGE	gender
Intercept	0.002211	-0.0012	-0.00121	-0.00121	-0.00116	-0.00001	-0.0001
Actos	-0.0012	0.004717	0.001218	0.001218	0.001219	-3.72E-7	6.928E-6
Avand	-0.00121	0.001218	0.005896	0.001218	0.001218	-1.94E-7	4.813E-6
Ins	-0.00121	0.001218	0.001218	0.00129	0.001218	-8.41E-8	2.174E-6
Sul	-0.00116	0.001219	0.001218	0.001218	0.001643	-9.71E-7	9.327E-6
Age	-0.00001	-3.72E-7	-1.94E-7	-8.41E-8	-9.71E-7	2.361E-7	-4.01E-7
Gender	-0.0001	6.928E-6	4.813E-6	2.174E-6	9.327E-6	-4.01E-7	0.000231

SAS output for acute myocardial infarction case using logistic regression model

Logistic model for Acute Myocardial Infaction Patients

The LOGISTIC Procedure

Model Information

Data Set	WORK.MI_COMB_TIME_RE
Response Variable	MI
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	137047
Number of Observations Used	137047

Response Profile

Ordered Value	MI	Total Frequency
1	1	4548
2	0	132507

Probability modeled is MI=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	39923.582	39481.928
SC	39933.410	39550.725
-2 Log L	39921.582	39467.928

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	453.6541	6	<.0001
Score	355.2896	6	<.0001
Wald	313.3010	6	<.0001

Logistic model for Acute Myocardial Infarction Patients

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.6418	0.1031	2028.5399	<.0001
Actos	1	0.6191	0.1521	16.5782	<.0001
Avand	1	0.9831	0.1603	37.5985	<.0001
Ins	1	1.3688	0.0835	268.7527	<.0001
Sul	1	1.1870	0.0918	167.2653	<.0001
Age	1	0.00103	0.000968	1.1268	0.2885
Gender	1	-0.0586	0.0303	3.7557	0.0526

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Actos	1.857	1.379	2.502
Avand	2.673	1.952	3.659
Ins	3.931	3.337	4.629
Sul	3.277	2.738	3.923
Age	1.001	0.999	1.003
Gender	0.943	0.889	1.001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	42.8	Somers' D	0.125
Percent Discordant	30.2	Gamma	0.172
Percent Tied	27.0	Tau-a	0.008
Pairs	602641836	c	0.563

Estimated Covariance Matrix

Parameter	Intercept	Actos	Avand	Ins	Sul	AGE	gender
Intercept	0.010621	-0.00661	-0.00665	-0.00667	-0.00646	-0.00006	-0.00037
Actos	-0.00661	0.023121	0.00669	0.006689	0.006695	-1.48E-6	0.000027
Avand	-0.00665	0.00669	0.025703	0.006689	0.006692	-7.82E-7	0.000019
Ins	-0.00667	0.006689	0.006689	0.006971	0.00669	-3.74E-7	8.56E-6
Sul	-0.00646	0.006695	0.006692	0.00669	0.008424	-3.85E-6	0.000037
Age	-0.00006	-1.48E-6	-7.82E-7	-3.74E-7	-3.85E-6	9.361E-7	-1.61E-6
Gender	-0.00037	0.000027	0.000019	8.56E-6	0.000037	-1.61E-6	0.000915

2) R code and output for James-Stein shrinkage estimation approach

```
#Read the data into R
HF.data<-read.csv(file="I:/ JS data/HF-MET.csv")
MI.data<-read.csv(file="I:/ JS data/MI-MET.csv")
C<-read.csv(file="I:/ JS data/C-matrix.csv",header=FALSE)
C<-as.matrix(C,4,10)
Age<-HF.data$X6
Gender<-HF.data$X5
Actos<-HF.data$X1
Avand<-HF.data$X2
Ins<-HF.data$X3
Sul<-HF.data$X4
HF<-HF.data$HF
MI<-MI.data$MI

k<-2        # number of adverse events
nX<-4       # drugs without age and gender
p<-(nX+1)*k       # dimension of beta
n<-length(Gender)
nn<-n*k             # numbers of rows for big X matrix
m<-k*n-p
X<-rep(1,n)         # a column of 1's
X0<-rep(0,n)        # a column of 0's
Xplus<-array(0,dim=c(n,(nX+1)*k,k))

# Construct the designed matrix used in full model and reduced model
for (i in 1:k){
Xplus[,,i]<-matrix(c(rep(X0,i-1),X,rep(X0,k-1),Actos,
rep(X0,k-1),Avand,rep(X0,k-1),Ins,
rep(X0,k-1),Sul,rep(X0,k-i)),n,(nX+1)*k)
}
XplusV<-matrix(0,p,n*k)
for (i in 1:p){
XplusV[i,]<-as.vector(Xplus[,i,])
}
XXplus<-t(X1plusV)   # XXplus is the design matrix for the full model

Y<-c(HF,MI)
len<-length(Y)

# using logistic regression model to get the estimate of beta (without age and gender) in full model
beta.glm<-glm(Y~XXplus[,1]+XXplus[,2]+XXplus[,3]+XXplus[,4]
+XXplus[,5]+XXplus[,6]+XXplus[,7]+XXplus[,8]
```

```
+XXplus[,9]+XXplus[,10]-1,family=binomial)
summary(beta.glm)

Call:
glm(formula = Y ~ XXplus[, 1] + XXplus[, 2] + XXplus[, 3] +
    XXplus[, 4] + XXplus[, 5] + XXplus[, 6] + XXplus[, 7] +
    XXplus[, 8] + XXplus[, 9] + XXplus[, 10] - 1, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6095  -0.6095  -0.2775  -0.2775   3.0389

# the estimate of beta in full model using logistic regression model
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
XX1plus[, 1]  -2.80631    0.03489 -80.428  < 2e-16 ***
XX1plus[, 2]  -4.60742    0.08178 -56.337  < 2e-16 ***
XX1plus[, 3]   0.52555    0.06868   7.653  1.97e-14 ***
XX1plus[, 4]   0.62233    0.15204   4.093  4.26e-05 ***
XX1plus[, 5]   0.78154    0.07678  10.179  < 2e-16 ***
XX1plus[, 6]   0.98506    0.16032   6.144  8.03e-10 ***
XX1plus[, 7]   1.21710    0.03592  33.882  < 2e-16 ***
XX1plus[, 8]   1.36969    0.08349  16.405  < 2e-16 ***
XX1plus[, 9]   1.05328    0.04048  26.020  < 2e-16 ***
XX1plus[, 10]  1.19322    0.09169  13.014  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 379997  on 274110  degrees of freedom
Residual deviance: 154223  on 274100  degrees of freedom
AIC: 154243
Number of Fisher Scoring iterations: 7

bbUL<-c(beta.glm$coefficients)
bbUL            # bbUL is estimate of beta in full model

r<-(k-1)*nX
d<-rep(0,r)
SS<-solve(t(XXplus)%*%XXplus)

# the estimate of beta (without age and gender) in the reduced model
(betaRe<-bbUL-SS%*%t(C)%*%solve(C%*%SS%*%t(C))%*%(C%*%bbUL-d))
       BetaRe
 [1,] -2.8729230
```

```
 [2,] -4.5408096
 [3,] 0.5739388
 [4,] 0.5739388
 [5,] 0.8832987
 [6,] 0.8832987
 [7,] 1.2933938
 [8,] 1.2933938
 [9,] 1.1232496
[10,] 1.1232496

###################### Using Pearson's chi-squared statistic to estimate sigma-square
rhatL<-XXplus%*%bbUL
PhatL<-exp(rhatL)/(1+exp(rhatL))
chi<-sum(((Y-PhatL)^2)/(PhatL*(1-PhatL)))
(se2L1<-chi2/m1)  #  se^2  is the residual mean squared under the full model
[1] 1.000036

###################### James-Stein shrinkage estimators
```

Note: It is reasonable to suspect that certain drugs may increase the risk of all of these adverse cardiovascular events (HF and MI here) according to similar biological mechanisms.

The null hypothesis is $H_0: C\beta = d$

Where $C = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$

And $\beta = (\beta_{h0}, \beta_{m0}, \beta_{h1}, \beta_{m1}, \beta_{h2}, \beta_{m2},, \beta_{h3}, \beta_{m3}, \beta_{h4}, \beta_{m4}, \beta_{h5}, \beta_{m5})^T$

$d = (0,0,0,0)^T$

```
########################

(c<-(r-2)*m/(r*(m+2)))
[1] 0.4999964
LnL<-(t(C%*%bbUL-d))%*%solve(C%*%SS%*%t(C))%*%(C%*%bbUL-d)/(r*se2L1)
(LnL<-LnL[1,1])
[1] 40.42989          # This is the test statistic F-value
ifelse(LnL>c, Ind<-1, Ind<-0)
[1] 1
(bbJSL<-betaRe+(1-c/LnL)*Ind*(bbUL-betaRe))  # JS estimator
#James-Stein shrinkage estimate (without age and gender)
      bbJSL
 [1,] -2.8071370
 [2,] -4.6065955
 [3,] 0.5261503
 [4,] 0.6217273
 [5,] 0.7828002
```

```
 [6,] 0.9837972
 [7,] 1.2180414
 [8,] 1.3687462
 [9,] 1.0541470
[10,] 1.1923523
```

Odds Ratio Estimates for JS estimate (without age and gender)

	Point estimate	Odds ratio	
[1,]	-2.8071370	0.060377605	--Intercept-HF
[2,]	-4.6065955	0.009985757	--Intercept-MI
[3,]	0.5261503	1.692404473	--Actos-HF
[4,]	0.6217273	1.862141810	--Actos-MI
[5,]	0.7828002	2.187589308	--Avandia-HF
[6,]	0.9837972	2.674592979	--Avandia-MI
[7,]	1.2180414	3.380560008	--Insulin-HF
[8,]	1.3687462	3.930419563	--Insulin-MI
[9,]	1.0541470	2.869526336	--Sulfonylurea-HF
[10,]	1.1923523	3.294822392	--Sulfonylurea-MI

Note: There is no big difference between the estimate by using logistic regression model and the estimate by using James-Stein shrinkage estimation.

3) SAS code and output for Cox models

```
/****** Cox model for HF case *******/
data hf_comb_time_re;
set diab.hf_comb_time_re;

proc phreg nosummary;
model failtime*hf(0) = Actos Avand Ins Sul age gender/ rl;
title 'Cox model for Heart Failure Patients ';
run;

/****** Cox model for AMI case *******/
data mi_comb_time_re;
set diab.mi_comb_time_re;

proc phreg nosummary;
model failtime*mi(0) = Actos Avand Ins Sul age gender/ rl;
title 'Cox model for Acute Myocardial Infaction Patients';
run;
```

SAS output heart failure case using Cox model

Cox model for Heart Failure Patient

The PHREG Procedure

Model Information

Data Set	WORK.HF_COMB_TIME_RE
Dependent Variable	failtime
Censoring Variable	HF
Censoring Value(s)	0
Ties Handling	BRESLOW

Number of Observations Read	137047
Number of Observations Used	137047

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	479557.53	477826.23
AIC	479557.53	477838.23
SBC	479557.53	477885.87

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1731.3002	6	<.0001
Score	1411.5263	6	<.0001
Wald	1293.8913	6	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Actos	1	0.48636	0.06574	54.7288	<.0001	1.626	1.430	1.850
Avand	1	0.61563	0.07267	71.7712	<.0001	1.851	1.605	2.134
Ins	1	1.14158	0.03475	1079.3339	<.0001	3.132	2.926	3.352
Sul	1	0.92150	0.03886	562.3908	<.0001	2.513	2.329	2.712
Age	1	0.0009454	0.0004443	4.5280	0.0333	1.001	1.000	1.002
Gender	1	-0.01003	0.01392	0.5186	0.4714	0.990	0.963	1.017

SAS output for acute myocardial infarction case using Cox model

Cox model for Acute Myocardial Infarction Patients

The PHREG Procedure

Model Information

Data Set	WORK.MI_COMB_TIME_RE
Dependent Variable	failtime
Censoring Variable	MI
Censoring Value(s)	0
Ties Handling	BRESLOW

Number of Observations Read	137047
Number of Observations Used	137047

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	105299.27	104841.62
AIC	105299.27	104853.62
SBC	105299.27	104892.15

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	457.6473	6	<.0001
Score	363.3716	6	<.0001
Wald	324.2439	6	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Actos	1	0.58887	0.15085	15.2397	<.0001	1.802	1.341	2.422
Avand	1	0.81778	0.15859	26.5909	<.0001	2.265	1.660	3.091
Ins	1	1.33905	0.08303	260.0652	<.0001	3.815	3.242	4.490
Sul	1	1.07332	0.09115	138.6672	<.0001	2.925	2.447	3.497
Age	1	0.00110	0.0009484	1.3424	0.2466	1.001	0.999	1.003
Gender	1	-0.05440	0.02971	3.3520	0.0671	0.947	0.893	1.004

4) R code and output for random effects Cox model

R output for heart failure case using random effects Cox model

---

Cox-Poisson program version 10.04
Primary records used 136970
Random Effects Model at One-level -- Hospitals; Heart Fail with 4 drugs followup 2000-2009
CoxPoiss estimation with: One-Level, Clusters Independent
Total event count: 20710

Model: survProps(endtime = failtime, event = hf) ~ actos + avand + ins + sul
Call:
CoxPoiss(model = MyModel, primary = datSrc, RandomEffects = reffs,
strata = strata5, logFile = LogFileName, outheading = OutHead,
maxiterations = 500, tolerance = 1e-08)

Sample Size : 110251670
Final Log-Likelihood: -194458.288

	Coefficient	Std. Error	t	exp(Coef)	Lower 95%	Upper 95%
actos	0.486817	0.0659716	7.37919	1.62713	1.42977	1.85173
avand	0.613514	0.0729884	8.40564	1.84691	1.60073	2.13095
ins	1.146330	0.0349894	32.76221	3.14662	2.93807	3.36998
sul	0.923710	0.0389404	23.72110	2.51862	2.33354	2.71837

Wald statistic for H0: { all coeffs = 0 } = 1279.4393 on 4 d.o.f.

Random effects dispersion parameters:
SigmaSq
0.000601144

Random effects:

Index	hospitals	U	Variance
1	12	0.997335	0.000601144
2	13	1.011779	0.000601144
3	14	0.996628	0.000601144
4	15	0.997631	0.000601144
5	16	1.009098	0.000601144
6	17	1.000492	0.000601144
7	18	0.977437	0.000601144
8	19	1.000331	0.000601144
9	20	1.000988	0.000601144
10	21	1.015405	0.000601144

11	22	1.003605	0.000601144
12	23	0.993985	0.000601144
13	24	0.998710	0.000601144
14	25	1.003987	0.000601144
15	26	0.999439	0.000601144
16	27	0.989948	0.000601144
17	28	1.002557	0.000601144
18	29	0.996844	0.000601144
19	30	1.002868	0.000601144
20	32	1.003930	0.000601144
21	33	1.000640	0.000601144
22	34	1.012426	0.000601144
23	35	0.980177	0.000601144
24	36	0.989996	0.000601144
25	40	1.011858	0.000601144
26	41	0.989352	0.000601144
27	42	0.988724	0.000601144
28	43	1.001985	0.000601144
29	44	0.995871	0.000601144
30	45	0.998653	0.000601144
31	46	0.999654	0.000601144
32	47	1.019324	0.000601144
33	48	0.984770	0.000601144
34	49	1.015322	0.000601144
35	50	0.992414	0.000601144
36	53	0.978907	0.000601144
37	54	1.000047	0.000601144
38	65	0.980669	0.000601144
39	66	1.005639	0.000601144
40	67	0.997394	0.000601144
41	68	0.998677	0.000601144
42	70	1.001775	0.000601144
43	71	0.996788	0.000601144
44	72	1.005938	0.000601144
45	74	1.006656	0.000601144
46	78	1.000075	0.000601144
47	79	0.999932	0.000601144
48	80	0.999197	0.000601144
49	81	1.000862	0.000601144
50	84	0.997264	0.000601144
51	85	1.000072	0.000601144
52	86	1.001269	0.000601144
53	87	1.000247	0.000601144
54	88	0.997395	0.000601144

55	90	0.999967	0.000601144
56	91	0.999099	0.000601144
57	92	1.020509	0.000601144
58	93	1.013294	0.000601144
59	94	0.997444	0.000601144
60	95	0.998103	0.000601144
61	96	1.001918	0.000601144
62	97	1.001393	0.000601144
63	98	0.999219	0.000601144
64	99	1.000648	0.000601144
65	107	0.996702	0.000601144
66	109	0.999900	0.000601144
67	110	1.000037	0.000601144
68	111	1.000851	0.000601144
69	113	1.000264	0.000601144
70	114	0.999162	0.000601144
71	116	0.999978	0.000601144
72	117	0.999608	0.000601144
73	118	1.001906	0.000601144
74	119	1.002318	0.000601144
75	122	1.000329	0.000601144
76	123	0.999885	0.000601144
77	125	1.002640	0.000601144
78	126	0.999742	0.000601144
79	127	0.999291	0.000601144
80	128	0.999119	0.000601144
81	129	0.999728	0.000601144
82	131	0.998510	0.000601144
83	132	1.016019	0.000601144
84	134	0.999453	0.000601144

R output for acute myocardial infarction case using random effects Cox model

---

Cox-Poisson program version 10.04
Primary records used 136970
Random Effects Model at One-level -- Hospitals; MI with 4 drugs followup 2000-2009
CoxPoiss estimation with: One-Level, Clusters Independent
Total event count: 4547

Model: survProps(endtime = failtime, event = mi) ~ actos + avand + ins + sul
Call:
CoxPoiss(model = MyModel, primary = datSrc, RandomEffects = reffs,
strata = strata5, logFile = LogFileName, outheading = OutHead,
maxiterations = 500, tolerance = 1e-08)

Sample Size : 24591755
Final Log-Likelihood: -42645.613

	Coefficient	Std. Error	t	exp(Coef)	Lower 95%	Upper 95%
actos	0.589292	0.1508591	3.90624	1.80271	1.34127	2.42291
avand	0.817513	0.1586027	5.15447	2.26486	1.65973	3.09061
ins	1.337521	0.0831216	16.09113	3.80959	3.23687	4.48364
sul	1.071103	0.0911791	11.74725	2.91860	2.44097	3.48968

Wald statistic for H0: { all coeffs = 0 } = 318.76132 on 4 d.o.f.

Random effects dispersion parameters:

SigmaSq
6.16701e-06

Random effects:

Index	hospitals	U	Variance
1	12	0.999976	6.16701e-06
2	13	0.999972	6.16701e-06
3	14	1.000011	6.16701e-06
4	15	1.000007	6.16701e-06
5	16	0.999956	6.16701e-06
6	17	1.000028	6.16701e-06
7	18	1.000038	6.16701e-06
8	19	1.000000	6.16701e-06
9	20	1.000011	6.16701e-06
10	21	1.000075	6.16701e-06
11	22	0.999988	6.16701e-06

12	23	0.999956	6.16701e-06
13	24	1.000010	6.16701e-06
14	25	0.999985	6.16701e-06
15	26	0.999979	6.16701e-06
16	27	0.999979	6.16701e-06
17	28	0.999973	6.16701e-06
18	29	0.999998	6.16701e-06
19	30	1.000047	6.16701e-06
20	32	1.000033	6.16701e-06
21	33	0.999982	6.16701e-06
22	34	0.999881	6.16701e-06
23	35	1.000046	6.16701e-06
24	36	1.000129	6.16701e-06
25	40	1.000007	6.16701e-06
26	41	0.999985	6.16701e-06
27	42	0.999924	6.16701e-06
28	43	1.000043	6.16701e-06
29	44	0.999952	6.16701e-06
30	45	1.000014	6.16701e-06
31	46	1.000026	6.16701e-06
32	47	1.000066	6.16701e-06
33	48	0.999937	6.16701e-06
34	49	1.000037	6.16701e-06
35	50	0.999976	6.16701e-06
36	53	0.999947	6.16701e-06
37	54	1.000021	6.16701e-06
38	65	1.000014	6.16701e-06
39	66	1.000036	6.16701e-06
40	67	1.000011	6.16701e-06
41	68	0.999996	6.16701e-06
42	70	0.999949	6.16701e-06
43	71	1.000000	6.16701e-06
44	72	1.000011	6.16701e-06
45	74	1.000062	6.16701e-06
46	78	0.999999	6.16701e-06
47	79	1.000000	6.16701e-06
48	80	0.999998	6.16701e-06
49	81	0.999998	6.16701e-06
50	84	1.000125	6.16701e-06
51	85	0.999997	6.16701e-06
52	86	0.999981	6.16701e-06
53	87	0.999987	6.16701e-06
54	88	0.999995	6.16701e-06
55	90	1.000017	6.16701e-06

56	91	0.999996	6.16701e-06
57	92	1.000043	6.16701e-06
58	93	0.999928	6.16701e-06
59	94	0.999991	6.16701e-06
60	95	0.999976	6.16701e-06
61	96	0.999990	6.16701e-06
62	97	0.999997	6.16701e-06
63	98	0.999998	6.16701e-06
64	99	0.999995	6.16701e-06
65	107	1.000002	6.16701e-06
66	109	1.000002	6.16701e-06
67	110	1.000001	6.16701e-06
68	111	0.999999	6.16701e-06
69	113	0.999999	6.16701e-06
70	114	0.999998	6.16701e-06
71	116	0.999999	6.16701e-06
72	117	0.999999	6.16701e-06
73	118	0.999998	6.16701e-06
74	119	1.000005	6.16701e-06
75	122	0.999999	6.16701e-06
76	123	1.000000	6.16701e-06
77	125	1.000001	6.16701e-06
78	126	1.000000	6.16701e-06
79	127	0.999999	6.16701e-06
80	128	0.999997	6.16701e-06
81	129	1.000019	6.16701e-06
82	131	0.999995	6.16701e-06
83	132	0.999893	6.16701e-06
84	134	1.000004	6.16701e-06