FACULTY OF SCIENCES
MASTER OF STATISTICS: BIOSTATISTICS
ACADEMIC YEAR 2014-2015

MASTER THESIS

Development of a propensity score to handle confounding of the association between H1N1 (2009) pandemic influenza vaccination and the risk of stroke in elderly

Student:
Kurnia WAHYUDI
1233676

Supervisors:
Germano FERREIRA, Pharm.D., M.P.H., Ph.D.
Prof. dr. Marc AERTS

January, 2015

Abstract

Several retrospective studies have suggested a reduction of stroke risk by influenza vaccination by preventing infections with influenza viruses and concomitant bacterial infections. Some studies controlled the confounding by performing matching based on some covariates in design phase and including other covariates in the model. However, these approach are only able to ensure the exposed and unexposed groups are comparable in a few important ways and do not explicitly balance covariates distributions within levels of the exposure effect estimate. Hence, there may be residual confounding and associated bias in the estimation of the exposure effect. Propensity score, defined as the conditional probability of being exposed given the observed covariates, can be incorporated into the analysis to balance the distributions of the covariates in the exposed and unexposed cohorts, and therefore reduce this bias. We assess the feasibility of developing a valid propensity score method to be used in matching cohorts in a potential hypothetical study to assess the association between H1N1 (2009) pandemic influenza vaccination and the occurrence of first ever stroke in subjects 65 years old and above. Logit model is used to estimate the propensity score with three different covariates selections criteria. With some covariates have missing values, generalized propensity score is more appropriate since it should condition both observed covariates and patterns of missing data. Hence, missingness pattern method is performed, where separate logit models are fitted using the subset of covariates fully observed for each pattern of missing data. After performing nearest neighbor (greedy), without replacement matching, on the logit of generalized propensity score using calipers of width 0.2 of the standard deviation of the logit of generalized propensity score, balanced distributions of the observed covariates and patterns of missing data in the exposed and unexposed cohorts are obtained. Therefore, it is feasibile to use a propensity score matching method to minimize the bias on estimating measure of association between H1N1 (2009) pandemic influenza vaccination and the risk of stroke in elderly.

Keywords: confounding; generalized propensity score; logit model; matching; missing data; missingness pattern; observational study; pandemic influenza vaccination; stroke

Acknowledgments

In the name of Allah, the Beneficent, the Merciful.

All praise and gratitude due to Allah, Lord of all creations. For everything that He has been given to me, especially for His guidance to the straight path of those whom His blessings are upon.

I would like to thank to my internal and external supervisors, for their supports, encouragements, and advices that were proved to be invaluable for the completion of this project.

I would like to thank to the Indonesian Directorate General of Higher Education (DIKTI) of Indonesian Ministry of Education and Culture, for granting me scholarship for enrolling to Master of Statistics: Biostatistics programme, in Hasselt University.

Last but not least, I am very grateful for the love and support from my family as well as staff of Department of Epidemiology and Biostatistics, Faculty of Medicine, Padjadjaran University, Indonesia, my friends, and others that cannot be mentioned here for their kindness and supports during my study period.

”For indeed, with hardship (will be) ease. Indeed, with hardship (will be) ease.”
(Quran chapter 94 verse 5-6)

To Melani and Audrey

Contents

1 Introduction **1**

1.1 Clinical Background . . . 1

1.2 Confounding Control in Observational Studies . . . 1

2 Rationale and Background **2**

2.1 Design of the Potential Hypothetical Study . . . 3

2.2 Population, Period, and Subjects of the Potential Hypothetical Study . . . 3

2.3 Covariates . . . 3

2.4 Data Source . . . 4

3 Objectives of the Project **4**

4 Methodology **4**

4.1 Propensity Score . . . 4

4.1.1 Propensity Score Models . . . 5

4.1.2 Forming Propensity Score Matched Sets . . . 7

4.1.3 Assessing Balance in Baseline Characteristics . . . 7

4.1.4 Propensity Score with Missing Data . . . 8

4.2 Analyses . . . 8

5 Results **9**

5.1 Descriptive Analysis . . . 9

5.2 Estimating Propensity Score . . . 11

5.3 Propensity Score Matching and Balance of Covariates and Patterns of Missng Data Assessment . . . 12

6 Discussion **17**

7 Conclusions **19**

A List of Extracted Covariates and Code List **25**

B Figures **26**

B.1 Distributions of Continuous Covariates under Complete Case Method . . . 26

B.2 Distributions of Continuous Covariates under Missingness Pattern Method . . . 28

C Tables **30**

C.1 Overlapping Range under Complete Case Method . . . 30

C.2 Overlapping Range under Missingness Pattern Method . . . 33

D SAS Code **36**

D.1 Fitting the Propensity Score Model under Complete Case Method . . . 36

D.2 Fitting the Propensity Score Model under Missingness Pattern Method . . . 36

D.3 Forming Propensity Score Matched Sample for Complete Case Method . . . 39

D.4 Absolute Standardized Differences for Matched Sample . . . 47

List of Figures

1 CPRD Data. Boxplots and quantile-quantile plots of ordinary propensity score of exposed and unexposed cohorts before matching based on all models under complete case method. . . . 12

2 CPRD Data. Boxplots and quantile-quantile plots of generalized propensity score of exposed and unexposed cohorts before matching based on all models under missingness pattern method. . . . 13

3 CPRD Data. Boxplots and quantile-quantile plots of ordinary propensity score of exposed and unexposed cohorts after matching based on all models under complete case method. . . . 16

4 CPRD Data. Boxplots and quantile-quantile plots of generalized propensity score of exposed and unexposed cohorts after matching based on all models under missingness pattern method. . . . 16

5 CPRD Data. Boxplots of age (years) of exposed and unexposed cohorts before and after matching based on all models under complete case method. . . . 26

6 CPRD Data. Quantile-quantile plots of age (years) of exposed and unexposed cohorts before and after matching based on all models under complete case method. . . . 26

7 CPRD Data. Boxplots of number of consultations of exposed and unexposed cohorts before and after matching based on all models under complete case method. . . . 27

8 CPRD Data. Quantile-quantile plots of number of consultations of exposed and unexposed cohorts before and after matching based on all models under complete case method. 27

9 CPRD Data. Boxplots of age (years) of exposed and unexposed cohorts before and after matching based on all models under missngness pattern method. . . . 28

10 CPRD Data. Quantile-quantile plots of age (years) of exposed and unexposed cohorts before and after matching based on all models under missingness pattern method. . . . 28

11 CPRD Data. Boxplots of number of consultations of exposed and unexposed cohorts before and after matching based on all models under missingness pattern method. . . . 29

12 CPRD Data. Quantile-quantile plots of number of consultations of exposed and unexposed cohorts before and after matching based on all models under missingness pattern method. . . . 29

List of Tables

1 CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts in the original sample, including their absolute standardized differences. 10

2 CPRD Data. Overview of patterns of missing data and their frequencies. ”O” indicates observed and ”M” indicates missing. 11

3 CPRD Data. Number available cases before matching and match sets based on all models, under complete case and missingness pattern methods. Standard deviation of the logit of ordinary and generalized propensity scores before matching, variance ratio of ordinary and generalized propensity scores before and after matching, and absolute standardized differences of ordinary and generalized propensity scores before and after matching, are also included. 14

4 CPRD Data. Absolute standardized differences of covariates of exposed and unexposed cohorts in the original and propensity score-matched samples based on complete case and missingness pattern methods. 14

5 CPRD Data. Proportion of patterns of missing data after matching based on all models under missingness pattern method. 15

6 CPRD Data. List of extracted covariates and code list used. 25

7 CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of ordinary propensity score using model 1 (all covariates) under complete case method. 30

8 CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of ordinary propensity score using model 2 (potential confounders) under complete case method. 31

9 CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of ordinary propensity score using model 3 (true confounders) under complete case method. 32

10 CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of generalized propensity score using model 1 (all covariates) under missingness pattern method. 33

11 CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of generalized propensity score using model 2 (potential confounders) under missingness pattern method. 34

12 CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of generalized propensity score using model 3 (true confounders) under missingness pattern method. 35

1 Introduction

1.1 Clinical Background

Acute and chronic infections may contribute to stroke risk (Lindsberg and Grau, 2003). These associations were found in young, middle-aged, and elderly subjects, irrespective of ischemic stroke subtype. The infections involved were mainly acute and chronic respiratory tract infections (Syrjänen et al., 1988; Bova et al., 1996; Macko et al., 1996; Grau et al., 1997, 1998). Infections may play a role in promoting the development of atherosclerotic plaques or in triggering their complications (Ross, 1999). Inflammation contributes to stroke risk via various interrelated mechanisms. Infectious diseases, traditional risk factors, and genetic susceptibility may cooperate in stimulating inflammatory pathways (Lindsberg and Grau, 2003). An increased leukocyte counts and mainly neutrophil counts (mostly found in bacterial infection) may indicate a short period of increased risk of stroke (Grau et al., 2004).

Subjects who are vaccinated against influenza are less prone to viral (influenza) infections and subsequent bacterial infection complications (Cox and Subbarao, 1999). Several retrospective studies suggest a reduction of stroke risk by influenza vaccination by preventing infections with influenza viruses and concomitant bacterial infections. A case-control study matched for age, sex, and district of residency, found that influenza vaccination was associated with reduced odds of stroke after adjustment for potential confounders (age, traditional risk factors, and antibiotics use) (odds ratio [OR], 0.50; 95% CI, 0.26 to 0.94) (Lavallée et al., 2002). In a larger case-control study matched for age, sex, and area of residence, after adjustment for vascular risk factors, education, health-related behavior and other factors, influenza vaccination were associated with reduced odds of stroke/transient ischemic attack (odds ratio [OR], 0.46; 95% CI, 0.28 to 0.77), with a trend toward protection from hemorrhagic stroke (Grau et al., 2005). In a cohort study with $>$140 000 elderly subjects, influenza vaccination was associated with a significant reduction of hospitalizations for cerebrovascular disease (16% 1998 to 1999; 23% 1999 to 2000) during influenza seasons (Nichol et al., 2003). In another cohort study based on the United Kingdom General Practice Research Database, within-person comparisons were undertaken and using the self-controlled case-series method, it was found that there was no increase in the risk of stroke in the period after seasonal influenza vaccination (Smeeth et al., 2004).

1.2 Confounding Control in Observational Studies

One of the primary challenges of observational studies is confounding, defined as systematic differences in prognosis between subjects exposed to an intervention of interest and the selected comparator group. Any observed difference in outcome risk between both groups cannot be attributed completely to a causal effect of the exposure on the outcome in the presence of uncontrolled confounding (Brookhart et al., 2013). Confounding in medicine can occur from a variety of sociomedical process (Brookhart et al., 2010). The most familiar form occurs from good medical practice, that is when physicians prescribing medications and performing procedures on patients who are most likely to benefit from them. This is called confounding by indication, which may cause the medical interventions appeared to cause the outcomes that they prevent (Walker, 1996). Patients with near the end of life considered by the physicians, may be less likely to receive preventive medications, leading to confounding by frailty or comorbidity (Glynn et al., 2001). Additional sources of confounding can result from patients' health-related behaviours that is after initiating a preventive medication, they may be more likely than others in engaging other healthy, prevention-oriented behaviours. This confounding is known as healthy user or adherer effect (Simpson et al., 2006).

An observational study is biased if the treated and control groups differ prior to the treatment in ways that matter for the outcomes under study. An overt bias is one that can be seen in the data at hand while a hidden bias can not be seen because the required information was not observed or recorded. Overt biases are controlled using adjustment, such as matching or stratification (Rosenbaum, 2002a,b). However, both approaches are practical only when covariates have low dimension and discrete measurement. As the number of covariates increases, it becomes difficult to find matched pairs with the same or similar values of covariates. Suppose there are hundreds or thousands of subjects and more than 15 covariates, it is likely that many subjects will have unique value of covariates (Rosenbaum, 2002a,b). Although in some observational studies mentioned earlier matching was performed in the design phase to control the confounding (Rubin, 1973a), this approach was only able to ensure the two groups were comparable in a few important ways. Beyond this, there was little to ensure comparability in all relevant ways (Rosenbaum, 2002a,b). Most often, confounding was controlled by including it in the model (i.e. multivariable regression model of the outcome) (Rubin, 1973b; Agresti, 2013; Hosmer et al., 2013; Brookhart et al., 2013). The strength of this approach is that by selecting covariates that are associated in a meaningful clinical or statistical manner with the outcome, and are believed to (or actually do) express statistical evidence of confounding of the treatment effect, adjusted estimates of the treatment effect that are assumed to be free of confounding may be obtained. However, it does not explicitly balance covariates distributions within levels of the treatment effect estimate. Hence, there may be residual confounding and associated bias in the effect estimate (Hosmer et al., 2013). To obtain unbiased estimate using this approach, the investigator must correctly models the effect of the treatment and the covariates on the outcome. However, correct specification of such model can be demanding, especially in a study involving rare outcome, many confounders, or strong treatment effect heterogeneity that must be correctly modeled (Cepeda et al., 2003). Therefore, drawing causal inferences from observational studies is challenging.

A method that is increasingly being used in observational studies, that directly addresses the potentially confounding effects of covariate imbalance, is to incorporate the propensity score, for treatment or exposure, into the analysis (Rosenbaum and Rubin, 1983; Rubin, 1997; Joffe and Rosenbaum, 1999; Rosenbaum, 2002a,b). It has been employed in a variety of disciplines including medicine (Weitzen et al., 2004; Luo et al., 2010), pharmacoepidemiology (Seeger et al., 2005; Glynn et al., 2006), psychology (Harder et al., 2010), social sciences (Thoemmes and Kim, 2011), and law (Rubin, 2001). The purpose of this method is to explicitly balance the distribution of covariates related to the choice of the treatment in order to reconstruct a situation similar to random assignment (Rosenbaum and Rubin, 1983; Rubin, 1997; Joffe and Rosenbaum, 1999; Rosenbaum, 2002a,b), as opposed to handling imbalance via statistical adjustment (Agresti, 2013; Hosmer et al., 2013; Guo and Fraser, 2015). The basic idea of propensity score is to focus on the prediction of treatments rather than on outcomes and to replace all the confounding covariates that play a role in the choice of a given treatment with a function of these covariates (Rosenbaum and Rubin, 1983; Joffe and Rosenbaum, 1999; Rosenbaum, 2002a,b). In other words, the goal is to reduce multidimensional covariates to a one-dimensional score (Guo and Fraser, 2015). The most frequently perfomed analytical method in epidemiology studies is matching on the propensity score (Austin, 2007, 2008), which requires creating matched sets of treated and untreated subjects such that matched subjects have similar values of the propensity scores.

2 Rationale and Background

This project aims to assess the feasibility of developing a valid propensity score method to be used in matching cohorts in a potential hypothetical study to assess the association between H1N1 (2009) pandemic influenza vaccination and the risk of stroke in subjects 65 years old and above. The scope of the project reported here is therefore the propensity score. The results of the potential hypothetical study itself are out of the scope of this project. A summary of the design, population and setting of the potential hypothetical study are described in the following sections.

2.1 Design of the Potential Hypothetical Study

The potential hypothetical study would be a retrospective, propensity score matched-cohort study, would be conducted to compare the risk of stroke in the cohort of subjects that were exposed to H1N1 (2009) pandemic influenza vaccine with the risk of stroke in the cohort of subjects that were not exposed to any H1N1 (2009) pandemic influenza vaccine. Subjects exposed to H1N1 (2009) pandemic influenza vaccine would be identified and each matched to one unexposed subject on logit of propensity score. Propensity score matching would be used to further ensure that the exposed and unexposed cohorts are comparable in the propensity of being exposed, thus minimizing the effect of potential confounders and aiming to obtain an unbiased comparison between both cohorts. Upon matching on a valid propensity score, the distribution of observed covariates is expected to be balanced between exposed and unexposed matched cohorts which should result in similar probabilities of receiving the vaccine, conditional on the propensity score. The association between vaccination with H1N1 (2009) pandemic influenza vaccine and the risk of stroke would be assessed by estimating the hazard ratio, which is the ratio of the hazards of first ever stroke at any time during the study period for vaccinated subjects and non-vaccinated subjects. The adjusted hazard ratio and their 95% CI would be estimated from a Cox regression.

2.2 Population, Period, and Subjects of the Potential Hypothetical Study

Subjects aged 65 years or older on 1 October 2009, with available records in the database from 1 September 2008 (beginning of the seasonal influenza vaccination campaign in previous year; i.e. at least 13 months database active registration prior to the start of the study period), without any history of stroke, and with no record of immunisation with H1N1 (2009) pandemic influenza vaccine with an unbranded/unknown manufacturer. The potential hypothetical study period was from 1 October 2009, which is the beginning of the H1N1 (2009) pandemic influenza mass vaccination campaign, until 31 August 2010. Cohort is defined as subjects with a CPRD record of being vaccinated with H1N1 influenza vaccine during the H1N1 (2009) pandemic influenza mass vaccination campaign (i.e. October 2009 - March 2010). Unexposed cohort is defined as propensity score matched-subjects with no CPRD record of H1N1 (2009) pandemic influenza vaccination during the H1N1 (2009) pandemic influenza mass vaccination campaign between October 2009 and March 2010.

2.3 Covariates

The outcome of interest would be occurrence of first ever stroke (regardless of type) within 180 days following administration of H1N1 (2009) pandemic influenza vaccine among an exposed cohort and during an equivalent time period (up to 180 days) in the unexposed cohort. Covariates included in the potential hypothetical study would be demographic characteristics (age, gender, and region), cardiovascular risk factors (diabetes mellitus, myocardial infarction, and congestive heart failure), lifestyle risk factors (alcohol intake, tobacco use and smoking, and body mass index), medications (antiplatelet, anticoagulant, antihypertensive, diuretics, statin, antidiabetic, antipsychotic, antidepressant, and nonsteroidal anti-inflamatory drugs), influenza infection or influenza-like illness (ILI), seasonal influenza vaccination, Charlson comorbidity index, and number of consultations. Further definition and measurement scale for each covariates are presented in the Appendix A Table 6.

2.4 Data Source

The Clinical Practice Research Datalink (CPRD) is one of the world's largest computerised databases of linked anonymised longitudinal medical records from primary care (Williams et al., 2012). The CPRD contains data for more than 11 million research standard patients, drawn from approximately 680 practices throughout the United Kingdom. The CPRD population closely matches the age and gender distribution of the UK population as a whole. Mean follow-up is approximately 7 years (median 5 years). The data are drawn from the computer systems used by general practitioners (GPs) to manage the clinical records within their practices. The CPRD contains coded longitudinal medical records from general practices (i.e. demographic information, records of clinical events [medical diagnoses]), referrals to specialists and secondary care settings, prescriptions issued in primary care, records of immunisations/vaccinations, diagnostic testing, lifestyle information [smoking and alcohol status] and all other types of care administered as part of routine GP practice). Diagnoses are retrieved by means of the READ medical classification system; READ codes are a coded thesaurus of clinical terms, which are the basic means by which clinicians record patient findings and procedures in health and social care IT systems across primary and secondary care (e.g. GP surgeries and pathology reporting of results). More recently, it has been linked to certain key data from hospital-based care called The Hospital Episode Statistics (HES) and The Office for National Statistics (ONS) mortality data. Inclusion of HES and ONS linked mortality data is expected to increase the validity of the case definition for at least those subjects with available linked data.

3 Objectives of the Project

1. To develop a propensity score model for H1N1 (2009) pandemic influenza vaccine exposure.
2. To estimate propensity score for exposed and unexposed subject cohorts in the potential hypothetical study population using the measured confounders in CPRD.
3. To assess the balance of covariates between exposed and unexposed cohorts.

4 Methodology

4.1 Propensity Score

Suppose in a model for an observational study with overt but no hidden bias, there are M units available for study and each has a value of an observed covariate $\mathbf{x}$, which can contain several covariates. Often, covariate $\mathbf{x}$ are used to reorganize the data prior to analysis, for example, by matching or stratifying on $\mathbf{x}$. Number of M units $j = 1, \ldots, M$, so $\mathbf{x}_{[j]}$ is the covariate for the jth unit and the treatment assignment for this unit is $Z_{[j]}$. After the reorganization, a unit will have different subscript without a bracket. Next, unit j is assigned to treatment with probability $\pi_{[j]} = \text{prob}(Z_{[j]} = 1)$ and to control with probability $1 - \pi_{[j]} = \text{prob}(Z_{[j]} = 0)$, with assignments for distinct units is independent, and with $0 < \pi_{[j]} < 1$. The model says that treatments were assigned by flipping a biased coin, possibly a different coin with a different bias for each unit, where the biases of the coin to the π's are unknown (Rosenbaum, 2002b).

$$\text{prob}(Z_{[1]} = z_1, \ldots, Z_{[M]} = z_M) = \prod_{j=1}^{M} \pi_{[j]}^{z_j} (1 - \pi_{[j]})^{1-z_j} \tag{1}$$

In an observational study, $\pi_{[j]}$ is unknown, thus the distribution of treatment assignment $Z_{[1]} = z_1, \ldots, Z_{[M]} = z_M$ is unknown, and it is not possible to draw inference similar to randomized experiment, where randomization created a known distribution of the treatment assignment. An observational study is free of the hidden bias if the π's, though unknown, are known to depend only on the observed covariates $\mathbf{x}_{[j]}$, hence the two units with the same value of $\mathbf{x}$ have the same chance π of receiving the treatment. In a formal fashion, the study is free of hidden bias if there is a function $\lambda(.)$, whose form will be typically be unknown, such that $\pi_{[j]} = \lambda(\mathbf{x}_{[j]})$ for $j = 1, \ldots, M$. If it is free of hidden bias, then equation (1) becomes

$$\text{prob}(Z_{[1]} = z_1, \ldots, Z_{[M]} = z_M) = \prod_{j=1}^{M} \lambda(\mathbf{x}_{[j]})^{z_j} [1 - \lambda(\mathbf{x}_{[j]})]^{1-z_j} \tag{2}$$

Equation (2) is also called randomization on the basis of the covariate (Rubin, 1977).

Besides to obtain matched sets that are homogeneous in covariates $\mathbf{x}$, there are two other aims of matching or stratification. Firstly, if there is no hidden bias so that it suffices to adjust for covariates, then strata or matched sets are desired that permit use of conventional methods. Secondly, whether or not there is hidden bias, one would like to compare treated and control groups with similar distributions of covariates, even if the matched subjects have different values of covariates. In addition, it is called covariate balance. Propensity score is a tool for constructing match sets or strata when covariates $\mathbf{x}$ has many in numbers (Rosenbaum, 2002b). Propensity score is defined as the conditional probability of receiving treatment given the observed covariates available (Rosenbaum and Rubin, 1983; Rosenbaum, 2002a,b).

When the study is free of hidden bias, the function $\lambda(\mathbf{x})$ is called the propensity score. A study is free of hidden bias when the treatment assignment probabilities $\pi_{[j]}$ are given by the propensity score $\lambda(\mathbf{x}_{[j]})$ which is always a function of the observed covariates $\mathbf{x}_{[j]}$. To adjust for overt bias in such study, one has to address the fact that the true $\lambda(\mathbf{x})$ is unknown. If the true $\lambda(\mathbf{x})$ is known, both objectives above would be attained by matching or stratifying on the propensity score, a single covariate. If the formed match sets or strata are homogeneous in the propensity score, even if they are heterogeneous in covariates $\mathbf{x}$, then the conventional methods are appropriate in the absence of hidden bias and the observed covariates $\mathbf{x}$ will tend to balance whether or not there is a hidden bias (Rosenbaum, 2002b).

There are two useful properties of the propensity score. First, if there is no hidden bias, then one does not need to form strata or matched sets that are homogeneous in $\mathbf{x}$, thus it suffices to obtain strata or matched sets that are homogeneous in $\lambda(\mathbf{x})$. If there is no hidden bias and if the strata are homogeneous in $\lambda(\mathbf{x})$, then the conditional distribution of the treatment assignment is uniform and the statistical methods for a randomized experiment may be used. Since $\mathbf{x}$ may be of high dimension, but $\lambda(\mathbf{x})$ is a number, it is often much easier to find subjects with similar values of $\lambda(\mathbf{x})$ than similar values of $\mathbf{x}$. When there is no hidden bias, when there is only overt bias due to $\mathbf{x}$, it suffices to adjust for the propensity score $\lambda(\mathbf{x})$.

Second, propensity score applies whether or not there is hidden bias, that is $\pi_i \neq \lambda(\mathbf{x})$. Strata or matched sets that are homogeneous in $\lambda(\mathbf{x})$ tend to balance $\mathbf{x}$ in the sense that treated and control subjects in the same stratum or matched set tend to have the same distribution of $\mathbf{x}$. In an experiment, randomization tends to balance all covariates, observed and unobserved, in the sense that treated and control groups tend to have the same distribution of covariate values. In an observational study, strata or matched sets that are homogeneous in the propensity score $\lambda(\mathbf{x})$ tend to balance observed covariates $\mathbf{x}$, though there may be imbalances in unobserved covariates (Rosenbaum, 2002b). The propensity score is the coarsest function of the covariates that is a balancing score, where the balancing score, $b(\mathbf{x})$, is defined as a function of the observed covariates $\mathbf{x}$ such that the conditional distribution of $\mathbf{x}$ given $b(\mathbf{x})$ is the same for treated $(Z_{[j]} = 1)$ and control $(Z_{[j]} = 0)$ units (Rosenbaum and Rubin, 1983; Rosenbaum, 2002b).

4.1.1 Propensity Score Models

In this study, propensity score $\lambda(\mathbf{x})$ was estimated using a logit model and the estimate was used to place the true propensity score (Rosenbaum, 2002b; Hosmer et al., 2013; Guo and Fraser, 2015). In the absence of hidden bias, the distribution of treatment assignments $\text{prob}(\mathbf{Z} = \mathbf{z})$ was unknown because the propensity score $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_S)^T$ was unknown. However, by conditioning on the number of the treated subjects in each stratum $\mathbf{m}$, the conditional distribution of the treatment assignments prob($\mathbf{Z} = \mathbf{z}|\mathbf{m}$), was known, and, in fact was a uniform randomized experiment. Thus, the unknown parameter $\boldsymbol{\lambda}$, was eliminated by conditioning on sufficient statistic $\mathbf{m}$. Suppose a logit model accurately describes the propensity score,

$$\log\left(\frac{\lambda(x_s)}{1-\lambda(x_s)}\right) = \beta^T \mathbf{x}_s \tag{3}$$

where β is an unknown parameter and $\bar{\mathbf{m}} = \sum_{s=1}^{S} m_s \mathbf{x}_s$, the sum of the $\mathbf{x}_s$ weighted by the number of treated subjects m_s in stratum s. Under model (3), $\bar{\mathbf{m}}$ is sufficient for β, hence the prob($\mathbf{Z} = \mathbf{z}|\bar{\mathbf{m}}$ is a known distribution, free of the unknown parameter β (Rosenbaum, 2002b).

Well-documented or well-known risk factors for stroke as suggested by Goldstein et al. (2011) were age > 55 years, male gender, low birth weight, black race, positive family history of stroke or transient ischemic attack, hypertension, smoking, diabetes, dyslipidemia, atrial fibrillation, left atrial thrombus, primary cardiac tumors, vegetations, and prosthetic cardiac valves, dilated cardiomyopathy, coronary artery disease, valvular heart disease, endocarditis, asymptomatic carotid artery stenosis, sickle cell disease, postmenopausal hormone therapy, excess salt intake, low potassium intake, excess weight, high alcohol consumption, suboptimal dietary pattern, physical inactivity, and obesity. Less-well documented risk factors were migrain, metabolic syndrome, excessive consumption of alcohol, drug addiction, sleep-disordered breathing, hyperhomocysteinemia, elevated lipoprotein(a), hypercoagulability, inflammation (i.e. rheumatoid arthritis and systemic lupus erthematosus) and infection. In addition, O'Donnell et al. (2010) suggested psychosocial stress or depression as a risk factor, although the association was weak. In terms of covariate selection for propensity score model, these were considered as potential confounders or covariates that affect the outcome (Brookhart et al., 2006). Recommendation for vaccination with H1N1 (2009) pandemic influenza vaccine as suggested by Mereckiene et al. (2012) were all age groups, health care workers, chronic diseases and underlying conditions (i.e. respiratory, cardiovascular, renal, neurological/neuromuscular, metabolic including diabetes, hepatic, immunosupression due to disease or treatment, any condition compromising respiratory function, hematologic, heamoglobinopathies, body mass index > 40 kg/m^2, pregnant women) and residents of long term care facilities. To determine true confounders, that is, covariates that affect both treatment assignment and outcome (Brookhart et al., 2006), we considered age, gender, region, diabetes mellitus, myocardial infarction, congestive heart failure, body mass index, antiplatelet used, anticoagulant used, antihypertensive used, diuretics used, antidiabetic used, statin used, NSAIDs used, influenza infection or influenza-like illness, and Charlson comorbidity index. Therefore, three propensity score models were developed using the following as the considered covariates:

1. All measured baseline covariates (regardless of their effect on the treatment and outcome).

$$\textbf{Model 1}: \log\left(\frac{\lambda(x_j)}{1-\lambda(x_j)}\right) = \beta_0 + \beta_1 Age_j + \cdots + \beta_p CCI_j \tag{4}$$

   where p is total number of covariates, including the dummy covariates for categorical data.

2. All potential confounders (baseline covariates that affect the outcome), as suggested by Brookhart et al. (2006).

$$\textbf{Model 2}: \log\left(\frac{\lambda(x_j)}{1-\lambda(x_j)}\right) = \beta_0 + \beta_1 Age_j + \cdots + \beta_q CCI_j \tag{5}$$

   where $q < p$ and q is number of all potential confounders, including the dummy covariates for categorical data. The potential confounders considered were: age, gender, region, diabetes mellitus, myocardial infarction, congestive heart failure, alcohol intake, tobacco use and smoking, body mass index, antiplatelet use, anticoagulant use, antihypertensive use, diuretics use, statin use, antidiabetic use, antipsychotic use, antidepressant use, non-steroidal anti-inflammatory drug use, influenza infection or influenza-like illness, and Charlson comorbidity index.

3. Only true confounders (baseline covariates that affect both treatment assignment and outcome), as suggested by Brookhart et al. (2006).

$$\textbf{Model 3}: \log\left(\frac{\lambda(x_j)}{1-\lambda(x_j)}\right) = \beta_0 + \beta_1 Age_j + \cdots + \beta_r CCI_j \tag{6}$$

where $r < q$ and r is number of all true confounders, including the dummy covariates for categorical data. The true confounders considered were: age, gender, region, diabetes mellitus, myocardial infarction, congestive heart failure, body mass index, antiplatelet use, anticoagulant use, antihypertensive use, diuretics use, antidiabetic use, statin use, NSAIDs use, influenza infection or influenza-like illness, and Charlson comorbidity index.

4.1.2 Forming Propensity Score Matched Sets

Propensity score matching requires the formation of sets of treated and untreated subjects with similar propensity score. A matched sets is a set of at least one treated subject and at least one untreated with similar propensity score values. The most commonly used approach in the medical literature is to form pairs of treated and untreated subjects with similar propensity score by nearest neighbor (greedy) matching using calipers of a specified width. For a given treated subject which is randomly selected, the closest untreated subject within the specified caliper distance is selected for matching to this treated subject, even if the untreated subject would be better have served as a match for a different treated subject. If multiple untreated subjects has propensity score that are equally close to that of the treated subject, then one of these untreated subjects is selected at random (Rosenbaum, 2002b). Nearest neighbor (greedy) matching on the logit of propensity score using proportion or calipers of width 0.2 of the standard deviation of the logit of the propensity score as suggested by Austin (2009b), with a fixed number of untreated subjects was performed (Rosenbaum, 2002b). Each exposed subject was matched to one subject who did not receive any H1N1 (2009) pandemic influenza vaccine during the study period on the basis of the individual logit of propensity score.

4.1.3 Assessing Balance in Baseline Characteristics

Absolute standardized difference was calculated to assess the balance or comparability of the treated and untreated subjects (Rosenbaum and Rubin, 1985; Austin et al., 2007; Rosenbaum, 2002b; Austin et al., 2010). For continuous covariates, it is defined as:

$$d = \frac{|\bar{x}_T - \bar{x}_C|}{\sqrt{\frac{s_T^2 + s_C^2}{2}}} \tag{7}$$

where $\bar{x}_T$ and $\bar{x}_C$ denote the sample mean of the covariate in treated and untreated subjects, and s_T^2 and s_C^2 are the sample standard deviations of the covariate in treated and untreated subjects, respectively (Flury and Riedwyl, 1986). For dichotomous covariates, it is defined as:

$$d = \frac{|\hat{p}_T - \hat{p}_C|}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{2}}} \tag{8}$$

where $\hat{p}_T$ and $\hat{p}_C$ denote the proportion of the dichotomous covariate in treated and untreated subjects, respectively (Flury and Riedwyl, 1986). Standardized differences of less than 0.1 (10%) was considered as negligible imbalance between treated and untreated subjects (Austin and Mamdani, 2006; Austin et al., 2007, 2010).

While the mean of covariate between treated and untreated subjects could be compared by standardized difference, one might want to compare the distribution of continuous covariates between treated and untreated subjects in the matched sample. Boxplots, empirical cumulative distribution function, or non-parametric estimates of the probability density function could be used to achieve this (Austin, 2009a; Austin et al., 2010).

4.1.4 Propensity Score with Missing Data

Since propensity score method encourage use of many covariates and with the generally large number of missing values recorded in the observational studies, a large proportion of subjects have at least one missing covariate value. D'Agostino and Rubin (2000) and D'Agostino et al. (2001) described the notation for propensity score with missing data as follows: Let the response indicator be $R_{ij}, (j = 1, \ldots, T)$, which is 1 if the value of the jth covariate for the ith subject is observed, 0 if it is missing. By definition, R_{ij} is fully observed. Let $\mathbf{x} = (\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{miss}})$, where $\mathbf{x}_{\text{obs}} = \mathbf{x}_{ij} | R_{ij} = 1$ denotes the observed parts and $\mathbf{x}_{\text{miss}} = \mathbf{x}_{ij} | R_{ij} = 0$ denotes the missing parts. The propensity score $\lambda(\mathbf{x})$ is now called generalized propensity score λ_i^*, defined by Rosenbaum and Rubin (1984) as the probability of treatment assignment given both the observed covariates and the patterns of missing data.

$$\lambda_i^* = \lambda_i^*(\mathbf{x}_{\text{obs,i}}, R_i) = \text{prob}(Z_i = 1 | \mathbf{x}_{\text{obs,i}}, R_i) \tag{9}$$

It has been showed by Rosenbaum and Rubin (1985) that with missing covariate and strongly ignorable treatment assignment given $\mathbf{x}_{\text{obs}}$ and R_i, the generalized propensity score λ_i^* in equation (9) plays the same role as the ordinary propensity score in equation (2) with no missing covariate data. Treatment assignment is strongly ignorable given $(\mathbf{x}_{\text{obs}}, R_i)$ if $\text{prob}(Z|X, Y, R = \text{prob}(Z|\mathbf{x}_{\text{obs}}, R)$. If in addition, the missing data mechanism is such that $\text{prob}(R|\mathbf{x}_{\text{obs}})$, then $\text{prob}(Z|X, Y, R = \text{prob}(Z|\mathbf{x}_{\text{obs}})$, and R itself can be ignored in the modelling.

In a large study with only a few patterns of missing data, Rosenbaum and Rubin (1984) and D'Agostino et al. (2001) considered using separate logit model using the subset of covariates fully observed for each patterns of missing data to estimate generalized propensity score. In an ideal situations, one would estimate a separate propensity score model for each specific patterns of missing data which could be interpreted as fitting a stratified model that stratifies the overall propensity score model by the missing data pattern (D'Agostino et al., 2001). With missing data, balance in the $(\mathbf{x}_{\text{obs}}, R_i)$ distributions between treated and untreated groups is also assessed after matching (D'Agostino and Rubin, 2000; D'Agostino et al., 2001).

4.2 Analyses

Comparison of the distributions and calculation of the absolute standardized difference (%) for each covariate between exposed and unexposed cohorts were performed in original sample. Correlation matrices between covariates were also constructed. Proportions and absolute standardized differences of patterns of missing data between exposed and unexposed cohorts were compared. Propensity score were estimated using logit models with all observed covariates in the first model, potential confounders in the second model, and true confounders in the third model. Ordinary propensity score was estimated under complete case method, while generalized propensity score was estimated under missingness pattern method, where separate logit models were fitted using the subset of covariates fully observed for each patterns of missing data. With the second method, estimated generalized propensity score for all subjects were expected to be available.

Based on each models, variance ratio of the ordinary and generalized propensity scores were calculated. Comparison of the distributions of the ordinary and generalized propensity scores for exposed and unexposed cohorts were performed using boxplots and quantile-quantile plots. Assessment of the overlapping range of logit of ordinary and generalized propensity scores, between exposed and unexposed cohorts, were also performed based on the three models. The overlapping range was determined on the basis of the logit of ordinary and generalized propensity scores values between the 5th percentiles of exposed cohort and 95th percentiles of the unexposed cohort. The characteristics between cohorts with logit of ordinary and generalized propensity scores below the overlap, within the overlap and above the overlap were compared again.

After ordinary and generalized propensity scores were estimated, nearest neighbor (greedy), without replacement matching, on the logit of ordinary and generalized propensity scores using proportion or calipers of width 0.2 of the standard deviation of the logit of the ordinary and generalized propensity scores were performed based on the three propensity score models. Number of match sets based on each models were recorded. Absolute standardized difference (%) for each covariates between exposed and unexposed cohorts and variance ratio of ordinary and generalized propensity scores based on each models were calculated again in the matched samples.

Distributions of ordinary and generalized propensity scores, age, and number of consultations using boxplots and quantile-quantile plots were also constructed again in the match samples. Proportions and absolute standardized differences of patterns of missing data between exposed and unexposed cohorts were compared again. These assessments of the quality of the propensity score matching were aimed to determine the quality of the matching in term of the distribution of the differences of distances between the logit of ordinary and generalized propensity scores of the exposed and unexposed subjects in each matched sets and to assess the performance of developed propensity score models. In order to correct unbalanced covariates and patterns of missing data, further modification of the propensity score model was considered by including the unbalanced covariates into the model (if previously not included), adding higher order terms (i.e. quadratic or cubic terms) for continuous covariates and/or including interaction terms in the propensity score model (Rosenbaum and Rubin, 1984; Austin et al., 2010). All analyses were perfomed using SAS software, version 9.2 of the SAS System for Windows. Copyright 2002 - 2008 SAS Institute Inc. Several graphical presentations were constructed using R 3.1.2 (R Core Team, 2014).

5 Results

5.1 Descriptive Analysis

Baseline characteristics of exposed and unexposed cohorts in the original sample and their absolute standardized differences are reported in Table 1. Patterns of missing data and their frequencies are reported in Table 2. A total of 775,360 subjects were eligible for the study. Vaccination coverage or percentage of subjects exposed to H1N1 (2009) pandemic influenza vaccine was 23.05%. Some covariates were observed to have initial absolute standardized differences larger than 0.1 (10%). Correlation coefficient between continuous covariates, age and number of consultations, were assessed in exposed and unexposed cohorts using Pearson product-moment correlation coefficient. The correlation coefficient was 0.11 and 0.24, for exposed and unexposed cohorts, respectively. Hence, indicating slightly larger correlation in the unexposed cohorts. Missing values were observed only for considered lifestyle risk factors, thus 8 different patterns of missing data were observed with proportion and absolute standardized difference of pattern 1 (complete cases) and pattern 4 (only missing alcohol intake) in exposed cohort were larger compared with unexposed cohort, while proportion and absolute standardized difference of pattern 8 (complete missingness of lifestyle risk factors) in unexposed cohort was larger compared with exposed cohort. Therefore, considerable initial bias was observed due to different distributions of observed covariates and patterns of missing data. Hence, this could extent to a biased comparison of outcome between exposed and unexposed cohorts in the potential hypothetical study.

Table 1: CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts in the original sample, including their absolute standardized differences.

Covariates	Exposed ($N = 178,737$)	Unexposed ($N = 596,623$)	Total ($N = 775,360$)	Standardized difference
Age in years, mean ± SD	74.63 ± 7.31	74.08 ± 7.87	74.21 ± 7.75	0.072
Age category, N (%)				
64-69 years	52,321 (29.27)	212,361 (35.59)	264,682 (34.14)	0.135
70-74 years	42,911 (24.01)	134,442 (22.53)	177,353 (22.87)	0.035
75-79 years	37,613 (21.04)	102,978 (17.26)	140,591 (18.13)	0.096
80-84 years	26,286 (14.71)	74,470 (12.48)	100,756 (12.99)	0.065
$\geq$ 85 years	19,606 (10.97)	72,372 (12.13)	91,978 (11.86)	0.036
Number of consultations, median (IQR)	35 (25)	22 (23)	25 (25)	0.634
Male, N (%)	89,023 (49.81)	252,856 (42.38)	341,879 (44.09)	0.149
Region, N (%)				
North East	3,629 (2.03)	9,910 (1.66)	13,539 (1.75)	0.027
North West	19,922 (11.15)	68,319 (11.45)	88,241 (11.38)	0.010
Yorkshire and The Humber	5,418 (3.03)	14,031 (2.35)	19,449 (2.51)	0.042
East Midlands	6,275 (3.51)	20,081 (3.37)	26,370 (3.40)	0.008
West Midlands	14,451 (8.09)	52,286 (8.76)	66,737 (8.61)	0.024
East of England	13,306 (7.44)	55,821 (9.36)	69,127 (8.92)	0.069
South West	16,800 (9.40)	59,053 (9.90)	75,853 (9.78)	0.017
South Central	16,909 (9.46)	73,481 (12.32)	90,390 (11.66)	0.092
London	13,062 (7.31)	54,887 (9.20)	67,949 (8.76)	0.069
South East Coast	18,271 (10.22)	68,643 (11.51)	86,914 (11.21)	0.041
Northern Ireland	7,134 (3.99)	15,601 (2.61)	22,735 (2.93)	0.077
Scotland	24,550 (13.74)	48,272 (8.09)	72,822 (9.39)	0.182
Wales	19,010 (10.64)	56,238 (9.43)	75,248 (9.70)	0.040
Cardiovascular risk factors, N (%)				
Diabetes mellitus	35,586 (19.91)	39,934 (6.69)	75,520 (9.74)	0.397
Myocardial infarction	13,118 (7.34)	15,969 (2.68)	29,087 (3.75)	0.215
Congestive heart failure	4,639 (2.60)	7,888 (1.32)	12,527 (1.62)	0.092
Lifestyle risk factors, N (%)				
Alcohol intake				
No	16,807 (35.52)	33,919 (37.82)	50,726 (37.03)	0.048
Current	26,811 (56.66)	49,599 (55.31)	76,410 (55.77)	0.027
Former	3,705 (7.83)	6,159 (6.87)	9,864 (7.20)	0.037
Missing	131,414 (73.52)	506,946 (84.97)	638,360 (82.33)	0.285
Tobacco use and smoking				
No	59,504 (42.63)	163,223 (47.08)	222,727 (45.80)	0.090
Current	15,809 (11.33)	49,423 (14.26)	65,232 (13.42)	0.088
Former	64,267 (46.04)	134,031 (38.66)	198,298 (40.78)	0.150
Missing	39,157 (21.91)	249,946 (41.89)	289,103 (37.29)	0.439
Body mass index				
Normal	153 (0.13)	256 (0.11)	409 (0.12)	0.006
Underweight	22 (0.02)	28 (0.01)	50 (0.01)	0.005
Overweight	461 (0.39)	1,071 (0.46)	1,532 (0.43)	0.010
Unknown	118,054 (99.46)	233,619 (99.42)	351,673 (99.44)	0.005
Missing	60,047 (33.60)	361,649 (60.62)	421,696 (54.39)	0.562
Medications, N (%)				
Antiplatelet	83,711 (46.83)	136,819 (22.93)	220,530 (28.44)	0.518
Anticoagulant	14,782 (8.27)	23,495 (3.94)	38,277 (4.94)	0.182
Antihypertensive	124,299 (69.54)	266,421 (44.65)	390,720 (50.39)	0.518
Diuretics	68,792 (38.49)	159,648 (26.76)	228,440 (29.46)	0.252
Antidiabetic	88 (0.05)	90 (0.02)	178 (0.02)	0.019

Continued on next page

Table 1 – continued from previous page

Covariates	Exposed $(N = 178,737)$	Unexposed $(N = 596,623)$	Total $(N = 775,360)$	Standardized difference
Statin	102,937 (57.59)	177,980 (29.83)	280,917 (36.23)	0.583
Antipsychotic	5,594 (3.13)	15,263 (2.56)	20,857 (2.69)	0.034
Antidepressant	26,937 (15.07)	65,618 (11.00)	92,555 (11.94)	0.121
NSAIDs	24,974 (13.97)	71,452 (11.98)	96,426 (12.44)	0.059
Influenza infection or ILI, N (%)	575 (0.32)	1,185 (0.20)	1,760 (0.23)	0.024
Seasonal influenza vaccination, N (%)				
Year 2008	165,315 (92.49)	364,115 (61.03)	529,430 (68.28)	0.803
Year 2009	171,223 (95.80)	364,164 (61.04)	535,387 (69.05)	0.932
Charlson comorbidity index, N (%)				
0	29,337 (16.41)	317,395 (53.20)	346,732 (44.72)	0.837
1	118,753 (66.44)	196,417 (32.92)	315,170 (40.65)	0.712
2	30,465 (17.04)	82,066 (13.76)	112,531 (14.51)	0.091
3 or more	182 (0.10)	745 (0.12)	927 (0.12)	0.007

NSAIDs, Nonsteroidal anti-inflammatory drugs; ILI, Influenza-Like Illness

Table 2: CPRD Data. Overview of patterns of missing data and their frequencies. ”O” indicates observed and ”M” indicates missing.

	Covariates			Exposed	Unexposed	Standardized
Pattern	Alcohol intake	Smoking status	BMI status	$(N = 178,737)$	$(N = 596,623)$	difference
1	O	O	O	43,102 (24.11)	75,300 (12.62)	0.174
2	O	O	M	3,387 (1.89)	12,304 (2.06)	0.002
3	O	M	O	640 (0.36)	1,169 (0.20)	0.002
4	M	O	O	63,832 (35.71)	125,816 (21.09)	0.233
5	M	M	O	11,116 (6.22)	32,689 (5.48)	0.011
6	O	M	M	194 (0.11)	904 (0.15)	0.001
7	M	O	M	29,259 (16.37)	133,257 (22.34)	0.089
8	M	M	M	27,207 (15.22)	215,184 (36.07)	0.287

5.2 Estimating Propensity Score

Under complete case method, the ordinary propensity score was estimated using three logit models that had exposure to vaccination as the response covariate. Under missingness pattern method, the generalized propensity score was estimated using separate logit models for each patterns of missing data, where models with pattern 1 were equal with previous under complete case method. Boxplots and quantile-quantile plots of ordinary and generalized propensity scores of exposed and unexposed cohorts are displayed in Figure 1 and 2. Comparison of each covariates between exposed and unexposed cohorts based on overlapping range of logit of ordinary and generalized propensity scores were also performed based on all models and the results are presented in the Appendix C.1 and C.2.

Under complete case method, based on model 1, the estimated ordinary propensity score for exposed and unexposed cohorts ranged from 0.007 to 0.9812 and 0.004 to 0.9781. Based on model 2, the estimated ordinary propensity score for exposed and unexposed cohorts ranged from 0.0491 to 0.8478 and 0.0467 to 0.8410. Based on model 3, the estimated ordinary propensity score for exposed and unexposed cohorts ranged from 0.0512 to 0.8280 and 0.0471 to 0.8244. More variability was observed under missingness pattern method. Based on model 1, the estimated generalized propensity score for exposed and unexposed cohorts ranged from 0.028 to 1.0000 and 0.000 to 0.9781. Based on model 2, the estimated generalized propensity score for exposed and unexposed cohorts ranged from 0.0154 to 1.0000 and 0.0000 to 0.8895. Based on model 3, the estimated generalized propensity score for exposed and unexposed cohorts ranged from 0.0146 to 1.0000 and 0.0000 to 0.8617. Under both methods, based on boxplots and quantile-quantile plots of ordinary and generalized propensity scores, some overlapping and some differences in distribution were observed. There were more exposed and unexposed subjects observed within the overlapping range based one all models compared with below and above the overlapping range. Absolute standardized differences for most of the covariates were observed to be lesser than 0.1 (10%) in below, within, and above overlapping ranges, based on all models. Overall, there was sufficient rationale to peform matching to balance the covariates and the patterns of missing data based on estimated ordinary and generalized propensity scores.

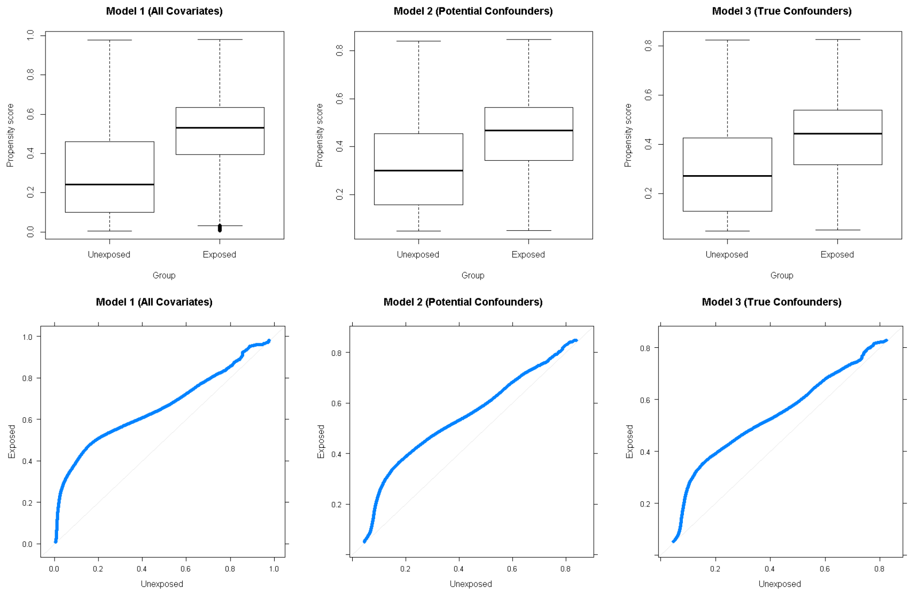

Figure 1: CPRD Data. Boxplots and quantile-quantile plots of ordinary propensity score of exposed and unexposed cohorts before matching based on all models under complete case method.

5.3 Propensity Score Matching and Balance of Covariates and Patterns of Missng Data Assessment

After ordinary and generalized propensity score were estimated based on logit models, subjects were then matched on the logit of the ordinary and generalized propensity scores, using a caliper of width of 0.2 standard deviation of the logit of the ordinary and generalized propensity scores. Standard deviation of the logit of ordinary and generalized propensity scores, variance ratio of ordinary and generalized propensity scores between exposed and unexposed cohorts, and absolute standardized difference of ordinary and generalized propensity scores between exposed and unexposed cohorts before matching based on all models, were calculated and recorded. Number of match sets or match pairs, variance ratio of ordinary and generalized propensity scores between exposed and unexposed cohorts, and absolute standardized difference of ordinary and generalized propensity scores between exposed and unexposed cohorts after matching based on all models, were also calculated and recorded. The previous terms were then compared and the results are presented in Table 3.

Comparison of absolute standardized difference of all covariates in original sample and after matching is presented in Table 4. Comparison of proportions and absolute standardized differences of patterns of missing data after matching based on all models under missingness pattern method are presented in

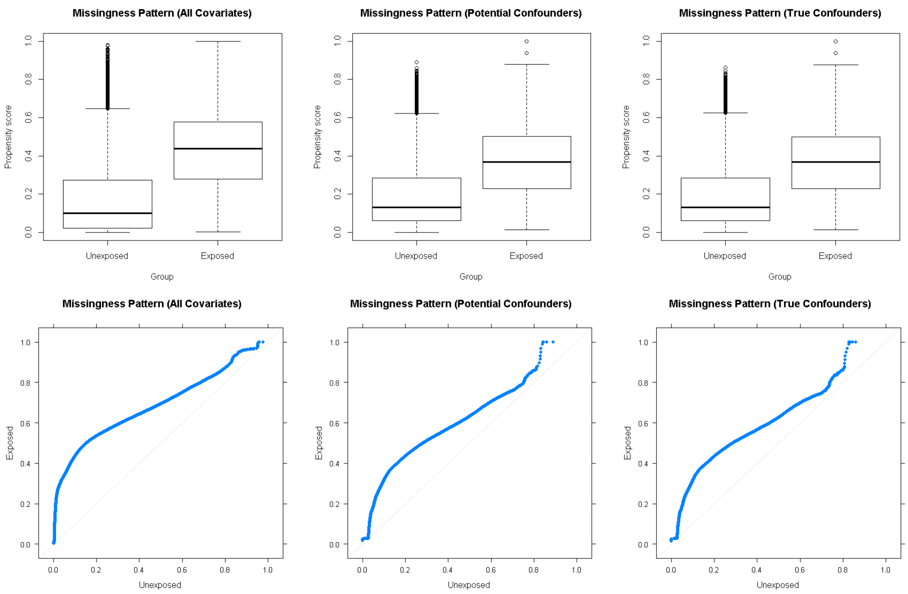

Figure 2: CPRD Data. Boxplots and quantile-quantile plots of generalized propensity score of exposed and unexposed cohorts before matching based on all models under missingness pattern method.

Table 5. Boxplots and quantile-quantile plots of the ordinary and generalized propensity scores, age, and number of consultations were calculated and constructed again to assess balanced of the covariates in exposed and unexposed cohorts. Boxplots and quantile-quantile plots of ordinary and generalized propensity score after matching under complete case and missingness methods are displayed in Figure 3 and 4, respectively. Boxplots and quantile-quantile plots of age and number of consultations, before and after matching, are displayed in the Appendix B.

Under complete case method, model 3 was observed to have the largest match sets since there was only one covariate with missing values included in the model and had large pool of unexposed subjects (approximately 2 times number of subjects in exposed cohort). Of the 118,690 subjects in exposed cohort in the initial sample, 109,992 (92.67%) were matched to one subject in unexposed cohort, while 8,698 (7.33%) subjects in exposed cohort were excluded from the matched sample because an appropriate subject in unexposed cohort was not identified. Similarly, 124,982 (53.19%) of subjets in unexposed cohort were excluded from the matched sample. Under missingness pattern method, proportions of subjects in exposed cohort matched to one subject in unexposed cohort were greater than 90%, based on all models since all subjects had their generalized propensity score estimated and had large pool of unexposed subjects (approximately 3 times number of subjects in exposed cohort).

Under both methods, variance ratio of ordinary and generalized propensity scores after matching were observed to be closed to 1 based on all models. Absolute standardized differences of ordinary and generalized propensity scores after matching were reduced to lesser than 0.1 (10%). Absolute standardized differences of covariates between exposed and unexposed cohorts after matching were lower than original sample based on all models. All covariates included in all models were observed to have absolute standardized differences lesser than 0.1 (10%). Although number of consultations and seasonal influenza vaccination were not included in model 2 and 3, some reduction of their standardized differences were observed after matching. The distribution of ordinary and generalized propensity scores between exposed and unexposed cohorts after matching were observed to be identical based on all models. The distribution of age and number of consultations between exposed and unexposed cohorts were observed to be more identical after matching. Under missingness pattern method, differences in patterns of missing data between exposed and unexposed cohorts were substantially reduced based on all models.

Propensity score matching based on all models under missingness pattern had good performance to balance the covariates between exposed and unexposed cohorts with respect to absolute standardized differences, variance ratio of generalized propensity score, and distributions of generalized propensity score, patterns of missing data, age, and number of consultations. Further modification of the propensity score model for model 2 and 3 would mean adding seasonal influenza vaccination and number of consultations into the model, thus the same as model 1. Therefore, model 1 under missingness pattern method appeared to be the best propensity score model that was fitted for producing balanced matched samples. This model reduced the bias on all covariates with large or moderate initial bias and on the patterns of missing data for lifestyle risk factors.

Table 3: CPRD Data. Number available cases before matching and match sets based on all models, under complete case and missingness pattern methods. Standard deviation of the logit of ordinary and generalized propensity scores before matching, variance ratio of ordinary and generalized propensity scores before and after matching, and absolute standardized differences of ordinary and generalized propensity scores before and after matching, are also included.

	Complete case							
	Before matching					After matching		
Model	Exposed	Unexposed	SD logit PS	Var. ratio	Stand. diff.	Match sets	Var. ratio	Stand. diff.
1	43,102	75,300	1.41	0.71	1.134	34,580	1.03	0.023
2	43,102	75,300	0.90	0.81	0.828	38,756	1.02	0.008
3	118,690	234,974	0.92	0.83	0.849	109,992	1.01	0.007
	Missingness pattern							
	Before matching					After matching		
Model	Exposed	Unexposed	SD logit PS	Var. ratio	Stand. diff.	Match sets	Var. ratio	Stand. diff.
1	178,737	596,623	1.78	1.16	1.353	158,737	1.04	0.023
2	178,737	596,623	1.17	1.21	1.059	169,683	1.02	0.007
3	178,737	596,623	1.17	1.19	1.054	170,330	1.02	0.007

Table 4: CPRD Data. Absolute standardized differences of covariates of exposed and unexposed cohorts in the original and propensity score-matched samples based on complete case and missingness pattern methods.

		Complete case			Missingness pattern		
Covariates	Original	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Age in years	0.072	0.032	0.023	0.020	0.025	0.017	0.013
Age category							
64-69 years	0.135	0.028	0.001	0.005	0.024	0.006	0.003
70-74 years	0.035	0.017	0.007	0.005	0.025	0.009	0.006
75-79 years	0.096	0.011	0.023	0.011	0.010	0.022	0.015
80-84 years	0.065	0.005	0.009	0.005	0.005	0.000	0.000
$\geq$ 85 years	0.036	0.024	0.031	0.026	0.017	0.032	0.023
Number of consultations	0.634	0.009	0.153	0.154	0.006	0.168	0.180
Male	0.149	0.020	0.017	0.014	0.011	0.014	0.018
Region							
North East	0.027	0.003	0.003	0.003	0.001	0.005	0.007
North West	0.010	0.013	0.011	0.022	0.007	0.010	0.017
Yorkshire and The Humber	0.042	0.002	0.002	0.009	0.003	0.005	0.011
East Midlands	0.008	0.008	0.004	0.007	0.000	0.005	0.006
West Midlands	0.024	0.004	0.007	0.002	0.001	0.002	0.000
East of England	0.069	0.004	0.002	0.006	0.004	0.007	0.007
South West	0.017	0.005	0.005	0.000	0.003	0.001	0.000
South Central	0.092	0.012	0.011	0.003	0.006	0.002	0.007

Continued on next page

Table 4 – continued from previous page

		Complete case			Missingness pattern		
Covariates	Original	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
London	0.069	0.000	0.003	0.010	0.010	0.008	0.005
South East Coast	0.041	0.000	0.002	0.002	0.008	0.008	0.002
Northern Ireland	0.077	0.005	0.006	0.014	0.009	0.017	0.017
Scotland	0.182	0.015	0.015	0.017	0.015	0.015	0.012
Wales	0.040	0.003	0.005	0.005	0.009	0.003	0.002
Cardiovascular risk factors							
Diabetes mellitus	0.397	0.020	0.016	0.007	0.024	0.019	0.018
Myocardial infarction	0.215	0.003	0.002	0.012	0.010	0.007	0.014
Congestive heart failure	0.092	0.003	0.005	0.013	0.001	0.008	0.016
Lifestyle risk factors							
Alcohol intake							
No	0.048	0.007	0.012	0.026	0.009	0.013	0.033
Current	0.027	0.009	0.010	0.021	0.009	0.010	0.024
Former	0.037	0.003	0.002	0.008	0.001	0.003	0.015
Tobacco use and smoking							
No	0.090	0.008	0.001	0.027	0.000	0.011	0.020
Current	0.088	0.000	0.001	0.096	0.006	0.009	.095
Former	0.150	0.008	0.000	0.090	0.004	0.005	0.083
Body mass index							
Normal	0.006	0.008	0.003	0.011	0.006	0.004	0.007
Underweight	0.005	0.003	0.003	0.001	0.001	0.002	0.001
Overweight	0.010	0.002	0.002	0.006	0.001	0.001	0.007
Unknown	0.005	0.006	0.003	0.009	0.002	0.002	0.010
Medications							
Antiplatelet	0.518	0.010	0.005	0.011	0.007	0.007	0.008
Anticoagulant	0.182	0.002	0.006	0.011	0.002	0.006	0.016
Antihypertensive	0.518	0.004	0.009	0.002	0.005	0.001	0.001
Diuretics	0.252	0.005	0.000	0.009	0.001	0.007	0.010
Antidiabetic	0.019	0.008	0.003	0.002	0.001	0.000	0.002
Statin	0.583	0.017	0.010	0.000	0.015	0.006	0.006
Antipsychotic	0.034	0.003	0.003	0.002	0.002	0.007	0.007
Antidepressant	0.121	0.000	0.005	0.038	0.001	0.008	0.045
NSAIDs	0.059	0.001	0.003	0.012	0.002	0.011	0.015
Influenza infection or ILI	0.024	0.001	0.000	0.004	0.000	0.005	0.008
Seasonal influenza vaccination							
Year 2008	0.803	0.012	0.534	0.545	0.003	0.577	0.583
Year 2009	0.932	0.006	0.687	0.712	0.002	0.742	0.751
Charlson comorbidity index							
0	0.837	0.029	0.018	0.010	0.015	0.022	0.016
1	0.712	0.001	0.003	0.008	0.008	0.012	0.009
2	0.091	0.028	0.013	0.001	0.025	0.007	0.005
3 or more	0.007	0.002	0.003	0.004	0.001	0.005	0.006

Table 5: CPRD Data. Proportion of patterns of missing data after matching based on all models under missingness pattern method.

	Model 1			Model 2			Model 3		
	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.
Pattern	$(N = 158,737)$	$(N = 158,737)$	diff.	$(N = 169,683)$	$(N = 169,683)$	diff.	$(N = 170,330)$	$(N = 170,330)$	diff.
1	35,415 (22.31)	34,598 (21.80)	0.012	39,089 (23.04)	37,332 (22.00)	0.025	39,386 (23.12)	38.499 (22.60)	0.012
2	3,207 (2.02)	3,279 (2.07)	0.003	3,332 (1.96)	3,058 (1.80)	0.012	3,334 (1.96)	3,050 (1.79)	0.012
3	501 (0.32)	480 (0.30)	0.002	559 (0.33)	511 (0.30)	0.005	553 (0.32)	473 (0.28)	0.009
4	54,230 (34.16)	53,474 (33.69)	0.010	59,423 (35.02)	59,307 (34.95)	0.001	59,770 (35.09)	59,768 (35.09)	0.000
5	9,964 (6.28)	10,016 (6.31)	0.001	10,763 (6.34)	10,668 (6.29)	0.002	10,757 (6.32)	10,400 (6.11)	0.009
6	182 (0.11)	193 (0.12)	0.002	192 (0.11)	158 (0.09)	0.006	191 (0.11)	150 (0.09)	0.008
7	28,579 (18.00)	29,473 (18.57)	0.015	29,182 (17.20)	30,149 (17.77)	0.015	29,186 (17.13)	30,143 (17.70)	0.015
8	26,659 (16.79)	27,224 (17.15)	0.009	27,143 (16.00)	28,500 (16.80)	0.022	27,153 (15.94)	27,847 (16.35)	0.011

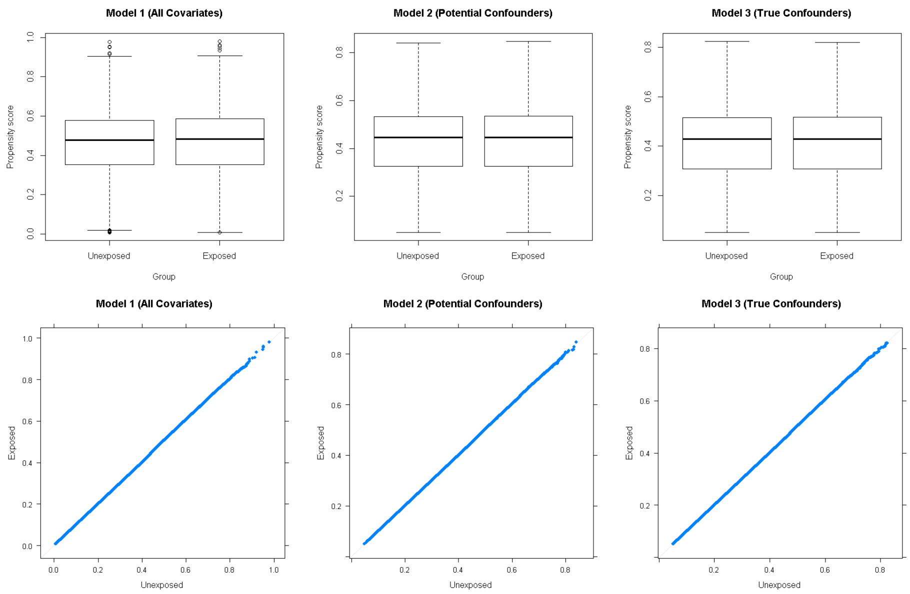

Figure 3: CPRD Data. Boxplots and quantile-quantile plots of ordinary propensity score of exposed and unexposed cohorts after matching based on all models under complete case method.

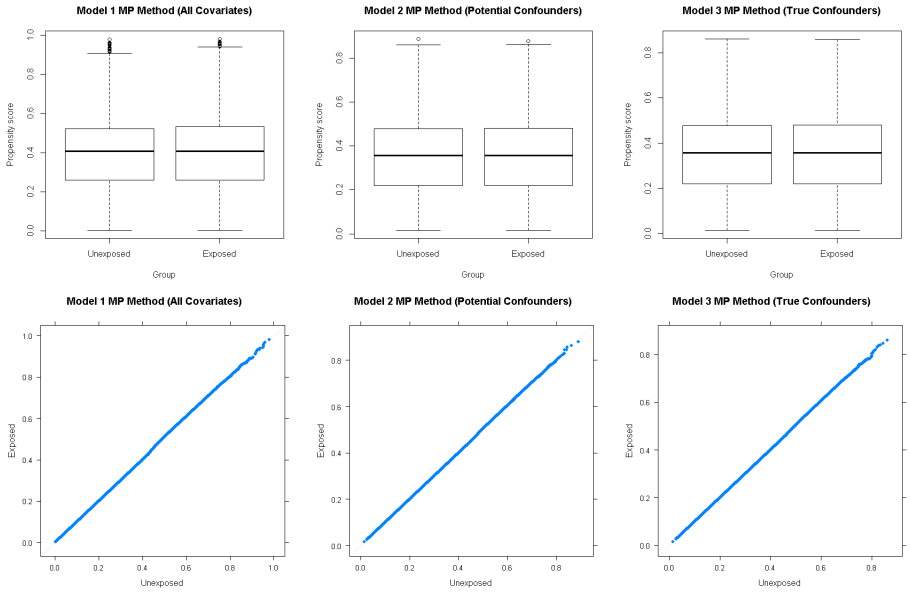

Figure 4: CPRD Data. Boxplots and quantile-quantile plots of generalized propensity score of exposed and unexposed cohorts after matching based on all models under missingness pattern method.

6 Discussion

The objectives of the project are to develop a propensity score model for H1N1 (2009) pandemic influenza vaccine exposure and to estimate the propensity score for exposed and unexposed subjects cohorts using observed covariates in the United Kingdom Clinical Practice Research Datalink (CPRD). Furthermore, to assess the balance of the covariates between exposed and unexposed cohorts. From descriptive statistics, it was observed that in the original sample, there was considerable initial bias due to different distributions of observed covariates. Charlson comorbidity index of 1, diabetes mellitus, antiplatelet used, antihypertensice used, statin used, diuretics used, antidepressant used, diabetes mellitus, myocardial infarction, practice based in Scotland, and former smoker were observed to have larger proportion in exposed group. This suggested confounding by indication (Walker, 1996), since most of them were indications recommended for H1N1 (2009) pandemic influenza vaccination. A larger proportion of subjects 85 years and older in unexposed cohort was observed, suggesting a confounding by frailty (Glynn et al., 2001). A healthy user or adherer effect (Simpson et al., 2006) was also suggested by larger proportion of seasonal influenza vaccination and median of number of consultations in exposed cohort. Three covariates for lifestyle risk factors were observed to have missing values with different proportion between exposed and unexposed cohorts, thus they were $2^3 = 8$ patterns of missing data. Exposed cohort had more complete cases and alcohol intake missing than unexposed cohort. Pattern with missing alcohol intake and BMI status and also missing in all lifestyle covariates were observed more in unexposed cohort. Thus, the missingness might be predictive about which exposure is received, in the sense that treatment assignment mechanism is ignorable given the observed covariates and the patterns of missing data (D'Agostino and Rubin, 2000; D'Agostino et al., 2001). All of these could lead to a biased comparison of outcome between the exposed and unexposed cohorts in the potential hypothetical study.

Under complete data method, there were very large numbers of subjects in exposed and unexposed cohorts that were excluded, thus lesser number of matched sets was formed. Matched sets based on model 1 and 2 were the formed from the completers subjects in total cohorts (15.27%) or subjects with pattern 1 only, in term of patterns of missing data. Although based on model 3 more than 90% of subjects in exposed cohort were matched, it actually comprised of subjects with some patterns of missing data (pattern 1, 3, 4 and 5). Thus, it was still ignoring missing data. Therefore, it could lead to a bias estimate of the exposure effect in the potential hypothetical study.

In order to minimize the bias on estimating the exposure effect in the potential hypothetical study, incorporating the information from missing data into propensity score estimation was important. Since large enough samples and sufficient subjects from each cohorts were obtained, separate logit models were fitted to estimate the generalized propensity score using the subset of the covariates fully observed for each patterns of missing data (Rosenbaum and Rubin, 1984; D'Agostino et al., 2001). Eight logit models based on each patterns were fitted to estimate the generalized propensity score, using all covariates, potential confounders, and true confounder, respectively. Models based on pattern 1 were similar with the previous complete case method. An issue was observed when using true confounders as covariates, since alcohol intake and tobacco used and smoking status were not included as covariates. Therefore, body mass index status was not included as covariates in the models based on pattern 2, 6,and 7. One benefit of this approach, as compared to the previous, was that the covariates were allowed to have different estimated parameters depending on the patterns of missing data. Thus, the other fully observed covariates, had 8 different estimates for its relationship with the exposure, each conditional on only the set of observed covariates within a separate missing data pattern (D'Agostino et al., 2001). Hence, all subjects had estimated generalized propensity score. Based on all models, more than 90% of subjects in exposed cohort were matched. The drawback from missingness pattern approach was that it increased the variability of the estimated generalized propensity score because only a subset of subjects was included in the model (Qu and Lipkovich, 2009, 2010).

To balance the observed covariates and patterns of missing data in exposed and unexposed subjects, matching on the logit of generalized propensity score were performed. Both empirical studies and Monte Carlo simulations have found that matching on propensity score eliminates a greater degree of the systematic differences in the observed covariates between treated and untreated subjets compare to stratification on the propensity score (Austin, 2007; Austin et al., 2007). Matching on the logit of the propensity score using calipers of width 0.2 of the standard deviation of the logit of the propensity score has been shown to result estimates of treatment effect with lower mean squarred error compared to other methods that are commonly used in the medical literature (Austin, 2009b; Austin et al., 2010).

Despite it is the most commonly reported in literatures, balance diagnostics based on the distribution of the estimated propensity score in exposed and unexposed cohorts should be handled with caution since it can be similar despite a miss-specified propensity score model (Austin, 2009a). Distributions of observed continuous covariates, such as age and number of consultations, in exposed and unexposed cohorts could be assessed in this regard (Austin, 2009a). Unbalanced of seasonal influenza vaccination and number of consultations were observed for model 2 and 3, under complete case and missingness pattern methods, since they were not included in the models. However, small reduction in their absolute standardized differences were still observed. Apart from assessing balance of observed covariates between exposed and unexposed cohorts after matching, balance in the distributions of patterns of missing data was also very important aspect since propensity score should condition both on observed covariates and patterns of missing data. (D'Agostino and Rubin, 2000; D'Agostino et al., 2001). Both were best obtained by propensity score model 1 (all covariates) under missingness pattern method.

This report has several limitations. Cardiovascular risk factors for stroke were not directly available such that proxies were used instead of true diagnosis. Diabetes, myocardial infarction, and congestive heart failure were derived from Charlson score for diabetes and diabetes with complications, myocardial infarction, and congestive heart failure, respectively. More important cardiovascular risk factors such as hypertension, hyperlipidemia, and atrial fibrillation were also not available. Due to the nature of the data source, not all risk factors could be controlled in this project. For instance, family history of stroke, education level and socio-economic status were not included in the analyses. Several risk factors were also not included in the analysis because they were not recorded in the CPRD (e.g. physical inactivity, diet and nutrition, ethnicity). Obstructive sleep apnea, an independent risk factor for stroke (Yaggi et al., 2005), was also not included. Although the information about obstructive sleep apnea was available in the CPRD, but seemed to be not comprehensively recorded. Such risk factors should be considered in the implementation of the potential hypothetical study. The data selection for this project was selected based on the most likely informative data for the propensity score.

Only one caliper width was considered while perhaps different widths could give a better performance in terms of obtaining more matched sets and much better balance in the distributions of observed covariates and patterns of missing data. When nearest neighbor (greedy) matching within specified caliper widths algorithm was applied, not all exposed subjects were being matched to unexposed subjects. For some exposed subjects, there might not be any unexposed subjects who were unmatched and whose ordinary propensity score or generalized propensity score lied between the specified caliper distance of that of the exposed subject.

There are still many aspects for further investigation of propensity score method. Different approaches for estimating propensity score could be explored, such as generalized boosted models (McCaffrey et al., 2004), neural network, linear classifiers (support vector machines), and classification and regression trees (CART) (Westreich et al., 2010). In term of handling the confounders by matching using propensity score, there are severals different settings that one could apply in terms of the caliper width settings (Lunt, 2014) and method, such as nearest neighbor (greedy) matching or optimal matching (Rosenbaum, 2002b; Guo and Fraser, 2015), where the average within-pair difference in the propensity score is minimized. Different algorithms for matching on the propensity score with their performances are further discussed in Austin (2014). Besides matching, one might also use stratification or subclassification (Rosenbaum and Rubin, 1984; Lunceford and Davidian, 2004; Rosenbaum, 2002b; Brookhart et al., 2013; Guo and Fraser, 2015) and inverse probability weighting (IPW) (Lunceford and Davidian, 2004; Cavuto et al., 2006; Austin et al., 2010; Brookhart et al., 2013; Guo and Fraser, 2015). For the potential hypothetical study, missing covariates would be appropriately handled by performing multiple imputation (MI) method or multiple imputation missingness pattern (MIMP) method, which utilizes multiple imputation method and missing pattern method (Qu and Lipkovich, 2009, 2010). For obtaining best results, it is also recommended to include the outcome covariate in the imputation procedure (Crowe et al., 2010). More detailed lifestyle risk factors information may be ascertained for another set of study subjects external to the potential hypothetical study. Such supplemental information can thus be used as ”validation data” to correct for confounding bias resulting from incomplete lifestyle risk factors information in the potential hypothetical study data. Regression calibration (RC) method by Stürmer et al. (2005), Bayesian propensity score approach by McCandless et al. (2012), or two-stage calibration (TCS) method by Lin et al. (2014) could be implemented. However, all calibration methods require that the covariates are measured in the same fashion between the 2 studies combined. Therefore, a necessary criterion for choosing an external data sample is the consistency in the definition and measurement instrument of the covariates between the external and the potential hypothetical study.

7 Conclusions

Propensity score was developed using logit models to handle confounding by matching in a potential hypothetical study to assess the association between H1N1 (2009) pandemic influenza vaccination and the risk of stroke in elderly based on electronic medical records databases. The propensity score model with all observed covariates included, under missingness pattern method, appeared to be the best propensity score model that was fitted for forming balanced matched sample on observed covariates and patterns of missing data. Thus, the bias observed between exposed and unexposed cohorts on some covariates and their patterns of missing data was substantially reduced by the matching. Therefore, it is feasibile to use a propensity score matching method to minimize the bias on estimating measure of association between H1N1 (2009) pandemic influenza vaccination and the risk of stroke in elderly.

References

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons, Hoboken, New Jersey.

Austin, P. C. (2007). Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery*, **134**(5):1128–35.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, **27**(12):2037–49.

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, **28**(25):3083–107.

Austin, P. C. (2009b). Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal. Biometrische Zeitschrift*, **51**(1):171–84.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, **33**(6):1057–69.

Austin, P. C., Chiu, M., Ko, D. T., Goeree, R., and Tu, J. V. (2010). Propensity score matching for estimating treatment effect. In Faries, D. E., Leon, A. C., Haro, J. M., and Obenchan, R. L., editors, *Analysis of Observational Health Care Data Using SAS*, pages 51–82. SAS Institute, Inc., Cary, North Carolina.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics In Medicine*, **26**(4):734–53.

Austin, P. C. and Mamdani, M. M. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-ami statin use. *Statistics In Medicine*, **25**(12):2084–106.

Bova, I. Y., Bornstein, N. M., and Korczyn, A. D. (1996). Acute infection as a risk factor for ischemic stroke. *Stroke*, **27**(12):2204–6.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, **163**(12):1149–56.

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J., and Schneeweiss, S. (2010). Confounding control in healthcare database research: Challenges and potential approaches. *Medical Care*, **48**(6):S114–20.

Brookhart, M. A., Wyss, R., Layton, J. B., and Stürmer, T. (2013). Propensity score methods for confounding control in nonexperimental research. *Circulation: Cardiovascular Quality and Outcomes*, **6**(5):604–11.

Cavuto, S., Bravi, F., Grassi, M., and Apolone, G. (2006). Propensity score for the analysis of observational data: an introduction and an illustrative example. *Drug Development Research*, **67**(3):208–16.

Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, **158**(3):280–7.

Cox, N. J. and Subbarao, K. (1999). Influenza. *Lancet*, **354**(9186):1277–82.

Crowe, B. J., Lipkovich, I. A., and Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical Statistics*, **9**(4):269–79.

D'Agostino, R., Lang, W., Walkup, M., Morgan, T., and Karter, A. (2001). Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG). *Health Services and Outcomes Research Methodology*, **2**(3-4):291–315.

D'Agostino, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, **95**(451):749–59.

Flury, B. K. and Riedwyl, H. (1986). Standard distance in univariate and multivariate analysis. *The American Statistician*, **40**(3):249–51.

Glynn, R. J., Knight, E. L., Levin, R., and Avorn, J. (2001). Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology*, **12**(6):682–9.

Glynn, R. J., Schneeweiss, S., and Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology*, **98**(3):253–9.

Goldstein, L. B., Bushnell, C. D., Adams, R. J., Appel, L. J., Braun, L. T., Chaturvedi, S., Creager, M. A., Culebras, A., Eckel, R. H., Hart, R. G., Hinchey, J. A., Howard, V. J., Jauch, E. C., Levine, S. R., Meschia, J. F., Moore, W. S., Nixon, J., and Pearson, T. A. (2011). Guidelines for the primary prevention of stroke: A guideline for healthcare professionals from the american heart association/american stroke association. *Stroke*, **42**(2):517–84.

Grau, A. J., Boddy, A. W., Dukovic, D. A., Buggle, F., Lichy, C., Brandt, T., Hacke, W., and for the CAPRIE Investigators (2004). Leukocyte count as an independent predictor of recurrent ischemic events. *Stroke*, **35**(5):1147–52.

Grau, A. J., Buggle, F., Becher, H., Zimmermann, E., Spiel, M., Fent, T., Maiwald, M., Werle, E., Zorn, M., Hengel, H., and Hacke, W. (1998). Recent bacterial and viral infection is a risk factor for cerebrovascular ischemia: clinical and biochemical studies. *Neurology*, **50**(1):196–203.

Grau, A. J., Buggle, F., Ziegler, C., Schwarz, W., Meuser, J., Tasman, A. J., Bühler, A., Benesch, C., Becher, H., and Hacke, W. (1997). Association between acute cerebrovascular ischemia and chronic and recurrent infection. *Stroke*, **28**(9):1724–29.

Grau, A. J., Fischer, B., Barth, C., Ling, P., Lichy, C., and Buggle, F. (2005). Influenza vaccination is associated with a reduced risk of stroke. *Stroke*, **36**(7):1501–6.

Guo, S. and Fraser, M. W. (2015). *Propensity Score Analysis: Statistical Methods and Applications*. SAGE Publications, Inc., Thousand Oaks, California.

Harder, V. S., Stuart, E. A., and Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, **15**(3):234–49.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, **150**(4):327–33.

Lavallée, P., Perchaud, V., Gautier-Bertrand, M., Grabli, D., and Amarenco, P. (2002). Association between influenza vaccination and reduced risk of brain infarction. *Stroke*, **33**(2):513–8.

Lin, H. C., Chiu, H. F., Ho, S. C., and Yang, C. Y. (2014). Association of influenza vaccination and reduced risk of stroke hospitalization among the elderly: A population-based case-control study. *International Journal of Environmental Research and Public Health*, **11**(4):3639–49.

Lindsberg, P. J. and Grau, A. J. (2003). Inflammation and infections as risk factors for ischemic stroke. *Stroke*, **34**(10):2518–32.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23**(19):2937–60.

Lunt, M. (2014). Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American Journal of Epidemiology*, **179**(2):226–35.

Luo, Z., Gardiner, J. C., and Bradley, C. J. (2010). Applying propensity score methods in medical research: Pitfalls and prospects. *Medical Care Research and Review*, **67**(5):528–54.

Macko, R. F., Ameriso, S. F., Barndt, R., Clough, W., Weiner, J. M., and Fisher, M. (1996). Precipitants of brain infarction: Roles of preceding infection/inflammation and recent psychological stress. *Stroke*, **27**(11):1999–2004.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, **9**:403–25.

McCandless, L. C., Richardson, S., and Best, N. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association*, **107**(497):40–51.

Mereckiene, J., Cotter, S., Weber, J. T., Nicoll, A., D'Ancona, F., Lopalco, P. L., Johansen, K., Wasley, A. M., Jorgensen, P., Levy-Bruhl, D., Giambi, C., Stefanoff, P., Dematte, L., and O'Flanagan, D. (2012). Influenza a(h1n1)pdm09 vaccination policies and coverage in europe. *Euro Surveill*, **17**(4).

Nichol, K. L., Nordin, J., Mullooly, J., Lask, R., Fillbrandt, K., and Iwane, M. (2003). Influenza vaccination and reduction in hospitalizations for cardiac disease and stroke among the elderly. *New England Journal of Medicine*, **348**(14):1322–32.

O'Donnell, M. J., Xavier, D., Liu, L., Zhang, H., Chin, S. L., Rao-Melacini, P., Rangarajan, S., Islam, S., Pais, P., McQueen, M., Mondo, C., Damasceno, A., Lopez-Jaramillo, P., Hankey, G. J., Dans, A. L., Yusoff, K., Truelsen, T., Diener, H. C., Sacco, R. L., Ryglewicz, D., Czlonkowska, A., Weimar, C., Wang, X., and Yusuf, S. (2010). Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the interstroke study): a case-control study. *The Lancet*, **376**(9735):112–23.

Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine*, **28**(9):1402–14.

Qu, Y. and Lipkovich, I. (2010). Propensity scoring with missing values. In Faries, D. E., Leon, A. C., Haro, J. M., and Obenchan, R. L., editors, *Analysis of Observational Health Care Data Using SAS*, pages 105–28. SAS Institute, Inc., Cary, North Carolina.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, **17**(3):286–327.

Rosenbaum, P. R. (2002b). *Observational Studies*. Springer-Verlag, New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, **79**(387):516–24.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, **39**(1):33–8.

Ross, R. (1999). Atherosclerosis — an inflammatory disease. *New England Journal of Medicine*, **340**(2):115–26.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics*, **29**(1):159–83.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**(1):185–203.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral Statistics*, **2**(1):1–26.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*, **127**(8):757–63.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, **2**(3-4):169–88.

Seeger, J. D., Williams, P. L., and Walker, A. M. (2005). An application of propensity score matching using claims data. *Pharmacoepidemiology and Drug Safety*, **14**(7):465–76.

Simpson, S. H., Eurich, D. T., Majumdar, S. R., Padwal, R. S., Tsuyuki, R. T., Varney, J., and Johnson, J. A. (2006). A meta-analysis of the association between adherence to drug therapy and mortality. *BMJ*, **333**(7557):15.

Smeeth, L., Thomas, S. L., Hall, A. J., Hubbard, R., Farrington, P., and Vallance, P. (2004). Risk of myocardial infarction and stroke after acute infection or vaccination. *New England Journal of Medicine*, **351**(25):2611–8.

Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology*, **162**(3):279–89.

Syrjänen, J., Valtonen, V. V., Iivanainen, M., Kaste, M., and Huttunen, J. K. (1988). Preceding infection as an important risk factor for ischaemic brain infarction in young and middle aged patients. *British Medical Journal*, **296**(6630):1156–60.

Thoemmes, F. J. and Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, **46**(1):90–118.

Walker, A. M. (1996). Confounding by indication. *Epidemiology*, **7**(4):335–6.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., and Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, **13**(12):841–53.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *J Clin Epidemiol*, **63**(8):826–33.

Williams, T., Van Staa, T., Puri, S., and Eaton, S. (2012). Recent advances in the utility and use of the general practice research database as an example of a uk primary care data resource. *Therapeutic Advances in Drug Safety*, **3**(2):89–99.

Yaggi, H. K., Concato, J., Kernan, W. N., Lichtman, J. H., Brass, L. M., and Mohsenin, V. (2005). Obstructive sleep apnea as a risk factor for stroke and death. *New England Journal of Medicine*, **353**(19):2034–41.

A List of Extracted Covariates and Code List

Table 6: CPRD Data. List of extracted covariates and code list used.

Number	Column name	Field name	Description	Code list
1.	Patient identifier	patid	Unique identifier given to a patient in the CPRD	-
2.	Exposure to vaccine	H1N1cohort	Indentfier for receiving influenza vaccine	Boolean
3.	Patient age	age2009	Patient's gender in years	Integer
4.	Patient gender	gender	Patient's gender	Boolean
5.	Number of consultations	N_consult12	Number of GP visits and consultations in the 12 months prior to potential hypothetical study	Integer
6.	Region	region	Value to indicate where in the UK the practice is based	Enumeration
7.	Diabetes mellitus	diabetes	Charlson score weight for diabetes and diabetes with complications as proxy	Boolean
8.	Myocardial infarction	myocardial	Charlson score weight for myocardial infarction as proxy	Boolean
9.	Congestive heart failure	chf	Charlson score weight for congestive heart disease as proxy	Boolean
10.	Alcohol intake	alcohol	Patient's alcohol intake status	Enumeration
11.	Tobacco use and smoking	smoking	Patient's tobacco and smoking status	Enumeration
12.	Body mass index	bmi	Patient's body mass index level	Enumeration
13.	Antiplatelet used	antiplatelet	Received more than 1 prescription in the previous 13 months for aspirin and other antiplatelets	Boolean
14.	Anticoagulant used	anticoagulant	Received more than 1 prescription in the previous 13 months of oral anticoagulants	Boolean
15.	Antihypertensive used	antihypert	Received more than 1 prescription in the previous 13 months of antihypertensives	Boolean
16.	Diuretics used	diuretics	Received more than 1 prescription in the previous 13 months of diuretics	Boolean
17.	Antidiabetic used	antidiabetic	Received more than 1 prescription in the previous 13 months of antidiabetics	Boolean
18.	Statin used	statin	Received more than 1 prescription in the previous 13 months of statins	Boolean
19.	Antipsychotic used	antipsychotic	Received more than 1 prescription in the previous 13 months of antipsychotics	Boolean
20.	Antidepresant used	antidepressant	Received more than 1 prescription in the previous 13 months of antidepressants	Boolean
21.	NSAIDs used	nsaids	Received more than 1 prescription in the previous 13 months of non-steroidal anti-inflammatory drugs	Boolean
22.	Influenza infection or ILI	flag_ILI	Influenza infection or Influenza-like illness in the 90 days prior to potential hypothetical study	Boolean
23.	Seasonal influenza vaccination 2008	flag_FLU2008	Seasonal influenza vaccination taken by a patient in 2008	Boolean
24.	Seasonal influenza vaccination 2009	flag_FLU2009	Seasonal influenza vaccination taken by a patient in 2009	Boolean
25.	Charlson comordity index	CCI	Total score of Charlson score for acquired immunodeficiency syndrome, chronic pulmonary disease, congestive heart disease, dementia, diabetes, diabetes with complications, hemiplegia and paraplegia, mild liver disease, moderate or severe liver disease, myocardial infarction, peptic ulcer disease, peripheral vascular disease, renal disease, rheumatological disease, cancer, metastatic tumour	Integer

B Figures

B.1 Distributions of Continuous Covariates under Complete Case Method

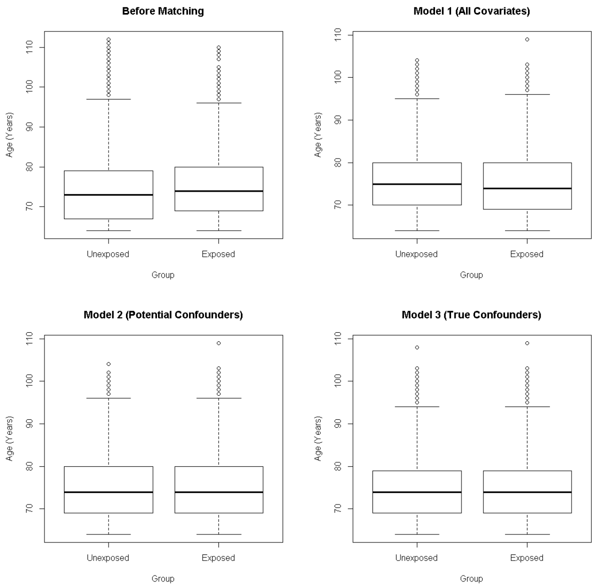

Figure 5: CPRD Data. Boxplots of age (years) of exposed and unexposed cohorts before and after matching based on all models under complete case method.

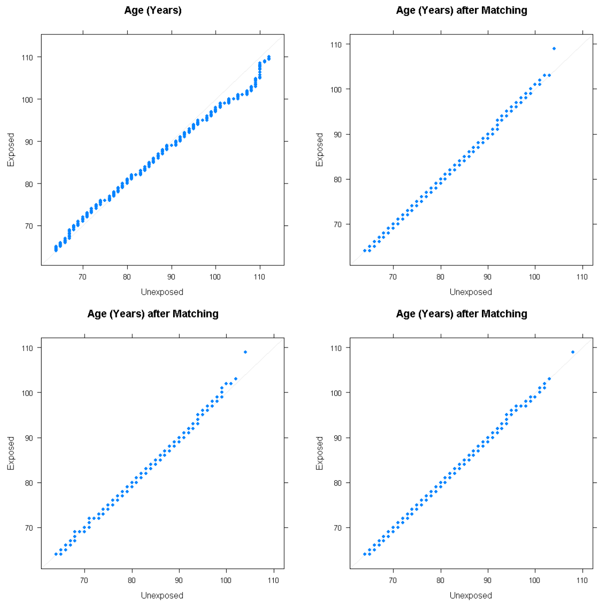

Figure 6: CPRD Data. Quantile-quantile plots of age (years) of exposed and unexposed cohorts before and after matching based on all models under complete case method.

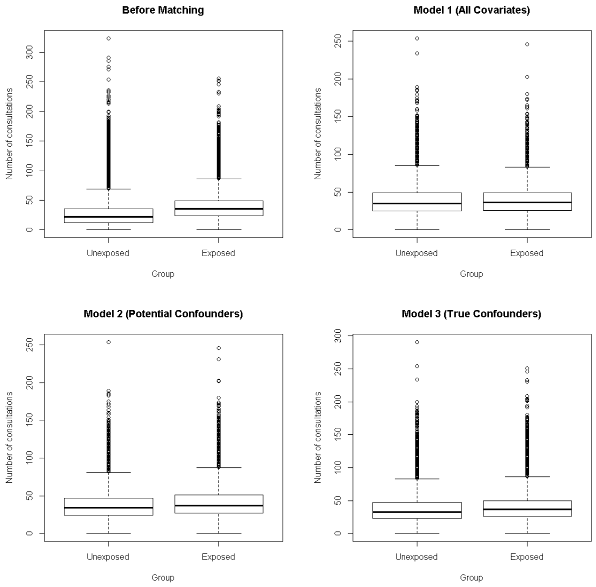

Figure 7: CPRD Data. Boxplots of number of consultations of exposed and unexposed cohorts before and after matching based on all models under complete case method.

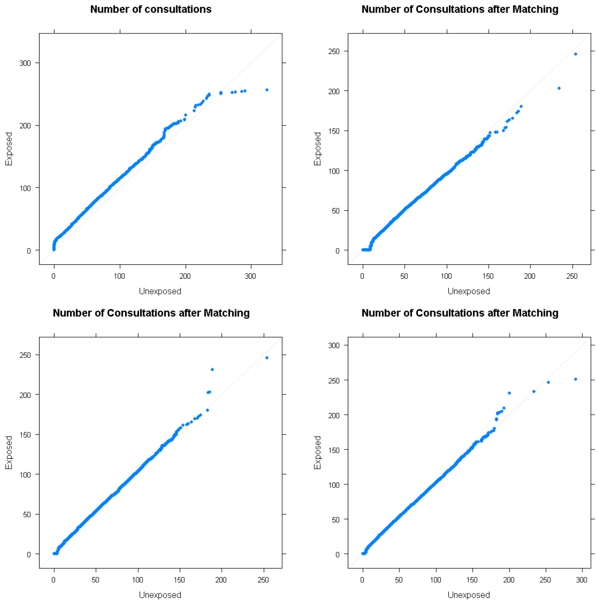

Figure 8: CPRD Data. Quantile-quantile plots of number of consultations of exposed and unexposed cohorts before and after matching based on all models under complete case method.

B.2 Distributions of Continuous Covariates under Missingness Pattern Method

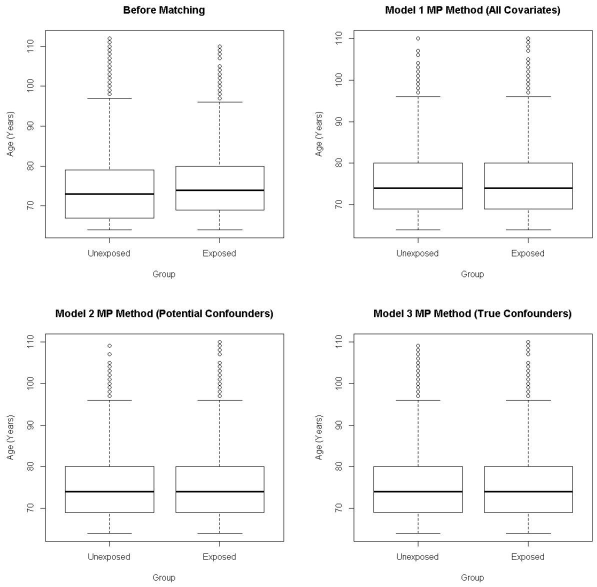

Figure 9: CPRD Data. Boxplots of age (years) of exposed and unexposed cohorts before and after matching based on all models under missngness pattern method.

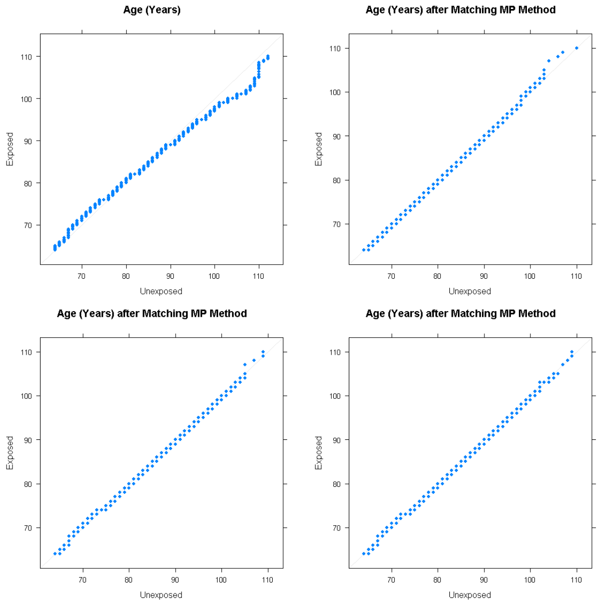

Figure 10: CPRD Data. Quantile-quantile plots of age (years) of exposed and unexposed cohorts before and after matching based on all models under missingness pattern method.

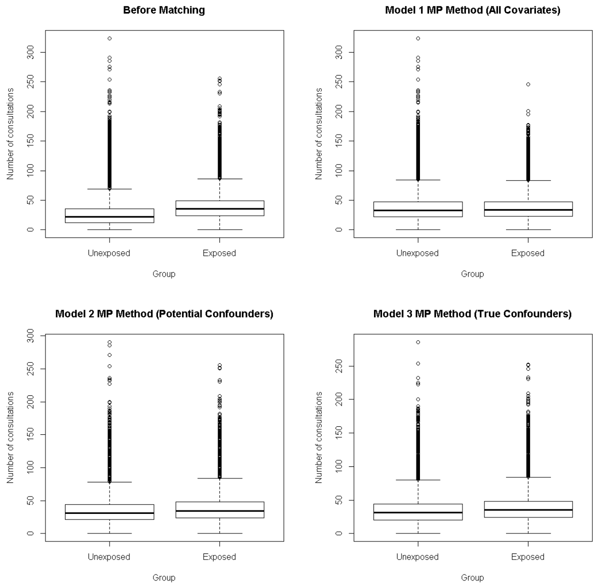

Figure 11: CPRD Data. Boxplots of number of consultations of exposed and unexposed cohorts before and after matching based on all models under missingness pattern method.

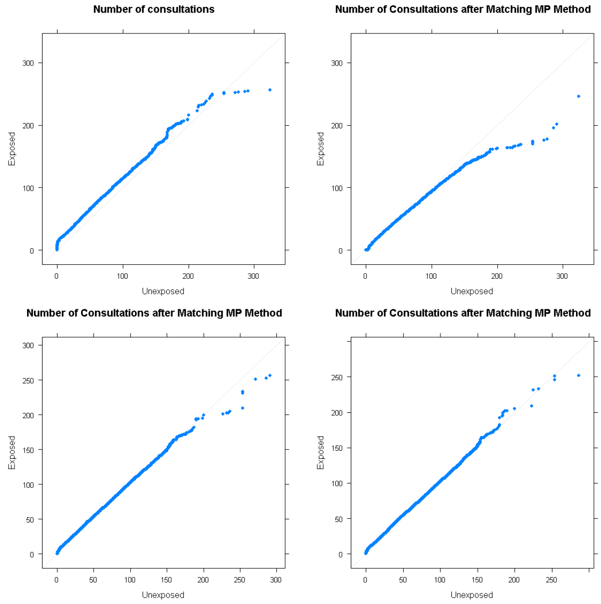

Figure 12: CPRD Data. Quantile-quantile plots of number of consultations of exposed and unexposed cohorts before and after matching based on all models under missingness pattern method.

C Tables

C.1 Overlapping Range under Complete Case Method

Table 7: CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of ordinary propensity score using model 1 (all covariates) under complete case method.

	Below the range			Within the range			Above the range		
	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.
Covariates	($N = 2,155$)	($N = 29,023$)	diff.	($N = 31,584$)	($N = 42,512$)	diff.	($N = 9,363$)	($N = 3,765$)	diff.
Age in years, mean±SD	74.72 ± 7.88	73.85 ± 7.66	0.112	75.27 ± 7.00	75.50 ± 7.19	0.031	72.09 ± 5.85	72.45 ± 5.93	0.062
Age category, N (%)									
64-69 years	692 (32.11)	10,514 (36.23)	0.087	8,804 (25.84)	13,172 (25.95)	0.000	3,654 (39.03)	1,401 (37.21)	0.037
70-74 years	435 (20.19)	6,287 (21.66)	0.036	8,256 (24.23)	11,606 (22.87)	0.006	2,698 (28.82)	1,109 (29.46)	0.014
75-79 years	432 (20.05)	5,230 (18.02)	0.052	7,882 (23.14)	10,920 (21.52)	0.034	1,857 (19.83)	725 (19.26)	0.015
80-84 years	321 (14.90)	3,804 (13.11)	0.052	5,569 (16.35)	8,433 (16.62)	0.010	876 (9.36)	386 (10.25)	0.030
$\geq$ 85 years	275 (12.76)	3,188 (10.98)	0.055	3,557 (10.44)	6,622 (13.05)	0.049	278 (2.97)	144 (3.82)	0.047
Number of consultations, median (IQR)	26 (13)	24 (12)	0.067	36 (22)	32 (22)	0.175	49 (31)	49 (34)	0.046
Male, N (%)	783 (36.33)	9,897 (34.10)	0.047	15,067 (44.23)	20,330 (40.06)	0.066	6,350 (67.82)	2,505 (66.53)	0.027
Region, N (%)									
North East	48 (2.23)	607 (2.09)	0.009	820 (2.41)	1,185 (2.33)	0.004	195 (2.08)	90 (2.39)	0.021
North West	399 (18.52)	4,577 (15.77)	0.073	5,260 (15.44)	8,185 (16.13)	0.001	561 (5.99)	311 (8.26)	0.088
Yorkshire and The Humber	67 (3.11)	764 (2.63)	0.029	1,014 (2.98)	1,427 (2.81)	0.004	178 (1.90)	100 (2.66)	0.051
East Midlands	55 (2.55)	758 (2.61)	0.004	1,228 (3.60)	1,618 (3.19)	0.023	346 (3.70)	177 (4.70)	0.050
West Midlands	183 (8.49)	2,441 (8.41)	0.003	2,922 (8.58)	4,278 (8.43)	0.017	365 (3.90)	206 (5.47)	0.074
East of England	189 (8.77)	2,534 (8.73)	0.001	2,741 (8.05)	4,062 (8.00)	0.008	264 (2.82)	151 (4.01)	0.066
South West	110 (5.10)	1,990 (6.86)	0.074	2,376 (6.97)	3,320 (6.54)	0.020	411 (4.39)	187 (4.97)	0.027
South Central	243 (11.28)	2,978 (10.26)	0.033	3,137 (9.21)	5,017 (9.89)	0.014	145 (1.55)	90 (2.39)	0.061
London	185 (8.58)	2,794 (9.63)	0.036	2,539 (7.45)	4,130 (8.14)	0.011	213 (2.27)	133 (3.53)	0.075
South East Coast	240 (11.14)	3,006 (10.36)	0.025	2,723 (7.99)	4,279 (8.43)	0.013	228 (2.44)	137 (3.64)	0.070
Northern Ireland	53 (2.46)	510 (1.76)	0.049	815 (2.39)	1,197 (2.36)	0.016	730 (7.80)	254 (6.75)	0.040
Scotland	163 (7.56)	2,768 (9.54)	0.071	4,429 (13.00)	6,479 (12.77)	0.028	4,525 (48.33)	1,434 (38.09)	0.208
Wales	220 (10.21)	3,296 (11.36)	0.037	4,064 (11.93)	5,576 (10.99)	0.017	1,202 (12.84)	495 (13.15)	0.009
Cardiovascular risk factors, N (%)									
Diabetes mellitus	257 (11.93)	2,764 (9.52)	0.078	8,939 (26.24)	9,718 (19.15)	0.186	4,649 (49.65)	1,864 (49.51)	0.003
Myocardial infarction	96 (4.45)	882 (3.04)	0.075	3,083 (9.05)	3,516 (6.93)	0.088	1,199 (12.81)	399 (10.60)	0.069
Congestive heart failure	29 (1.35)	317 (1.09)	0.023	892 (2.62)	1,196 (2.36)	0.032	241 (2.57)	113 (3.00)	0.026
Lifestyle risk factors, N (%)									
Alcohol intake									
No	937 (43.48)	11,113 (38.29)	0.106	12,477 (36.62)	19,183 (37.80)	0.008	2,809 (30.00)	1,235 (32.80)	0.060
Current	1,084 (50.30)	16,103 (55.48)	0.104	19,072 (55.98)	28,001 (55.17)	0.003	5,666 (60.51)	2,152 (57.16)	0.068
Former	134 (6.22)	1,807 (6.23)	0.000	2,519 (7.39)	3,569 (7.03)	0.022	888 (9.48)	378 (10.04)	0.019
Tobacco use and smoking									
No	1,194 (55.41)	16,937 (58.36)	0.060	17,206 (50.50)	26,933 (53.07)	0.057	3,234 (34.54)	1,297 (34.45)	0.002
Current	317 (14.71)	4,148 (14.29)	0.012	3,524 (10.34)	5,798 (11.42)	0.005	819 (8.75)	353 (9.38)	0.022
Former	644 (29.88)	7,938 (27.35)	0.056	13,338 (39.15)	18,022 (35.51)	0.062	5,310 (56.71)	2,115 (56.18)	0.011
Body mass index									
Normal	1 (0.05)	19 (0.07)	0.008	33 (0.10)	48 (0.09)	0.009	11 (0.12)	10 (0.27)	0.034
Underweight	1 (0.05)	0 (0.00)	0.030	4 (0.01)	2 (0.00)	0.006	0 (0.00)	2 (0.05)	0.033
Overweight	7 (0.32)	102 (0.35)	0.005	113 (0.33)	176 (0.35)	0.001	30 (0.32)	9 (0.24)	0.015
Unknown	2,146 (99.58)	28,902 (99.58)	0.000	33,918 (99.56)	50,527 (99.55)	0.005	9,322 (99.56)	3,744 (99.44)	0.017
Medications, N (%)									
Antiplatelet	583 (27.05)	6,621 (22.81)	0.098	18,490 (54.27)	23,023 (45.36)	0.169	7,815 (83.47)	3,160 (83.93)	0.013
Anticoagulant	117 (5.43)	837 (2.88)	0.128	2,494 (7.32)	2,998 (5.91)	0.055	1,244 (13.29)	574 (15.25)	0.056
Antihypertensive	1,212 (56.24)	15,684 (54.04)	0.044	26,768 (78.57)	37,177 (73.25)	0.119	8,655 (92.44)	3,485 (92.56)	0.005
Diuretics	778 (36.10)	9,753 (33.60)	0.052	14,767 (43.35)	21,398 (42.16)	0.024	4,312 (46.05)	1,746 (46.37)	0.006
Antidiabetic	0 (0.00)	11 (0.04)	0.028	21 (0.06)	30 (0.06)	0.008	5 (0.05)	1 (0.03)	0.013
Statin	782 (36.29)	9,170 (31.60)	0.099	23,202 (68.10)	29,322 (57.77)	0.190	8,737 (93.31)	3,517 (93.41)	0.004
Antipsychotic	75 (3.48)	931 (3.21)	0.015	1,159 (3.40)	1,793 (3.53)	0.001	319 (3.41)	146 (3.88)	0.025
Antidepressant	266 (12.34)	3,064 (10.56)	0.056	5,080 (14.91)	7,089 (13.97)	0.030	1,912(20.42)	844 (22.42)	0.049
NSAIDs	286 (13.27)	3,436 (11.84)	0.043	4,430 (13.00)	6,588 (12.98)	0.003	1,363 (14.56)	534 (14.18)	0.011
Influenza infection or ILI, N (%)	3 (0.14)	58 (0.20)	0.015	98 (0.29)	135 (0.27)	0.004	49 (0.52)	17 (0.45)	0.010
Seasonal influenza vaccination, N (%)									
Year 2008	1,214 (56.33)	11,506 (39.64)	0.339	31,954 (93.79)	37,807 (74.49)	0.089	9,275 (99.06)	3,727 (98.99)	0.007
Year 2009	993 (46.08)	9,302 (32.05)	0.291	32,893 (96.55)	37,162 (73.22)	0.159	9,363 (100.00)	3,764 (99.97)	0.023
Charlson comorbidity index, N (%)									
0	1,178 (54.66)	17,767 (61.22)	0.133	3,095 (9.08)	8,104 (15.97)	0.331	1 (0.01)	2 (0.05)	0.024
1	734 (34.06)	8,289 (28.56)	0.119	25,111 (73.71)	31,644 (62.35)	0.312	8,850 (94.52)	3,550 (94.29)	0.010
2	243 (11.28)	2,944 (10.14)	0.037	5,830 (17.11)	10,938 (21.55)	0.054	512 (5.47)	212 (5.63)	0.007
3 or more	0 (0.00)	23 (0.08)	0.040	32 (0.09)	67 (0.13)	0.005	0 (0.00)	1 (0.03)	0.023

Table 8: CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of ordinary propensity score using model 2 (potential confounders) under complete case method.

	Below the range			Within the range			Above the range		
	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.
Covariates	($N = 2,155$)	($N = 20,784$)	diff.	($N = 34,068$)	($N = 50,753$)	diff.	($N = 6,879$)	($N = 3,763$)	diff.
Age in years, mean±SD	74.67 ± 7.49	73.43 ± 7.26	0.168	74.99 ± 7.01	75.40 ± 7.41	0.057	72.38 ± 5.93	72.55 ± 6.08	0.027
Age category, N (%)									
65-69 years	639 (29.65)	7,616 (36.64)	0.149	12,832 (25.31)	9,181 (21.64)	0.003	2550 (37.07)	1,411 (37.50)	0.009
70-74 years	503 (23.34)	5,054 (24.32)	0.023	13,200 (16.04)	10,790 (25.43)	0.032	1,969 (28.62)	1,066 (28.33)	0.007
75-79 years	449 (20.84)	3,762 (18.10)	0.069	12,575 (24.81)	10,654 (25.11)	0.039	1,442 (20.96)	740 (19.67)	0.032
80-84 years	320 (14.85)	2,463 (11.85)	0.088	8,004 (15.79)	7,419 (17.49)	0.007	688 (10.00)	374 (9.94)	0.002
$\geq$ 85 years	5244 (11.32)	1,889 (9.09)	0.074	4,079 (8.05)	4,379 (10.32)	0.081	230 (3.34)	172 (4.57)	0.063
Number of consultations, median (IQR)	28 (20)	22 (17)	0.429	38 (24)	32 (12)	0.245	42 (29)	40 (28)	0.084
Male, N (%)	703 (32.62)	6,339 (30.50)	0.046	27,116 (53.49)	20,446 (48.20)	0.085	4,973 (72.29)	2,663 (70.77)	0.034
Region, N (%)									
North East	59 (2.74)	427 (2.05)	0.045	1,328 (2.62)	978 (2.31)	0.005	118 (1.72)	74 (1.97)	0.019
North West	407 (18.89)	3,243 (15.60)	0.087	7,242 (14.29)	6,869 (16.19)	0.019	296 (4.30)	214 (5.69)	0.064
Yorkshire and The Humber	61 (2.83)	626 (3.01)	0.011	1,657 (3.27)	1,143 (2.69)	0.010	152 (2.21)	101 (2.68)	0.031
East Midlands	54 (2.51)	476 (2.29)	0.014	2,166 (4.27)	1,556 (3.67)	0.023	240 (3.49)	176 (4.68)	0.060
West Midlands	190 (8.82)	1,750 (8.42)	0.014	4,866 (9.60)	4,029 (9.50)	0.005	211 (3.07)	152 (4.04)	0.053
East of England	182 (8.45)	1,954 (9.40)	0.034	4,129 (8.15)	3,916 (9.23)	0.002	170 (2.47)	125 (3.32)	0.051
South West	100 (4.64)	1,416 (6.81)	0.094	5,180 (10.22)	3,962 (9.34)	0.017	228 (3.31)	164 (4.36)	0.054
South Central	192 (8.91)	2,205 (10.61)	0.057	5,361 (10.58)	5,122 (12.07)	0.023	46 (0.67)	46 (1.22)	0.057
London	220 (10.21)	2,185 (10.51)	0.010	4,073 (8.04)	4,143 (9.77)	0.026	120 (1.74)	83 (2.21)	0.033
South East Coast	211 (9.79)	2,378 (11.44)	0.054	5,446 (10.74)	4,662 (10.99)	0.016	92 (1.34)	76 (2.02)	0.053
Northern Ireland	51 (2.37)	309 (1.49)	0.064	1,083 (2.14)	688 (1.62)	0.002	625 (9.09)	313 (8.32)	0.027
Scotland	216 (10.02)	1,532 (7.37)	0.094	2,787 (5.50)	1,741 (4.10)	0.007	3,938 (57.25)	1,830 (48.63)	0.173
Wales	212 (9.84)	2,283 (10.98)	0.038	5,372 (10.60)	3,614 (8.52)	0.030	643 (9.35)	409 (10.87)	0.050
Cardiovascular risk factors, N (%)									
Diabetes mellitus	0 (0.00)	0 (0.00)	-	18,442 (36.38)	13,929 (32.83)	0.170	3,565 (51.82)	1,962 (52.14)	0.006
Myocardial infarction	0 (0.00)	3 (0.01)	0.017	6,398 (12.62)	5,138 (12.11)	0.078	874 (12.71)	378 (10.05)	0.084
Congestive heart failure	1 (0.05)	4 (0.02)	0.015	1,471 (2.90)	1,239 (2.92)	0.017	168 (2.44)	95 (2.52)	0.005
Lifestyle risk factors, N (%)									
Alcohol intake									
No	896 (41.58)	7,719 (37.14)	0.091	6,961 (35.42)	6,077 (37.84)	0.024	1,864 (27.10)	1,082 (28.75)	0.037
Current	1,128 (52.34)	11,995 (57.71)	0.108	11,183 (56.91)	8,789 (54.73)	0.016	4,381 (63.69)	2,323 (61.73)	0.040
Former	131 (6.08)	1,070 (5.15)	0.040	1,506 (7.66)	1,194 (7.43)	0.014	634 (9.22)	358 (9.51)	0.010
Tobacco use and smoking									
No	1,202 (55.78)	13,076 (62.91)	0.146	20,015 (43.25)	17,389 (45.17)	0.051	2,169 (31.53)	1,170 (31.09)	0.009
Current	367 (17.03)	2,744 (13.20)	0.107	4,147 (8.96)	4,137 (10.75)	0.035	382 (5.55)	247 (6.56)	0.042
Former	586 (27.19)	4,964 (23.88)	0.076	22,112 (47.78)	16,968 (44.08)	0.075	4,328 (62.92)	2,346 (62.34)	0.012
Body mass index									
Normal	0 (0.00)	19 (0.09)	0.043	58 (0.11)	28 (0.07)	0.001	24 (0.35)	9 (0.24)	0.020
Underweight	0 (0.00)	0 (0.00)	-	4 (0.01)	4 (0.01)	0.009	0 (0.03)	2 (0.05)	0.033
Overweight	7 (0.32)	65 (0.31)	0.002	197 (0.39)	183 (0.43)	0.003	20 (0.29)	10 (0.27)	0.005
Unknown	2,148 (99.68)	20,700 (99.60)	0.013	50,431 (99.49)	42,208 (99.49)	0.001	6,835 (99.36)	3,742 (99.44)	0.011
Medications, N (%)									
Antiplatelet	393 (18.24)	1,978 (9.52)	0.254	33,791 (97.76)	26,294 (97.04)	0.179	5,955 (86.57)	3,331 (88.52)	0.059
Anticoagulant	72 (3.34)	287 (1.38)	0.129	5,010 (97.49)	3,663 (97.06)	0.057	1,111 (16.15)	662 (17.59)	0.039
Antihypertensive	1,010 (46.87)	9,101 (43.79)	0.062	48,433 (99.30)	40,116 (99.20)	0.125	6,498 (94.46)	3,557 (94.53)	0.003
Diuretics	691 (32.06)	6,191 (29.79)	0.049	25,203 (96.00)	21,394 (95.83)	0.024	3,273 (47.58)	1,757 (46.69)	0.018
Antidiabetic	0 (0.00)	0 (0.00)	-	40 (78.43)	45 (78.95)	0.001	3 (0.04)	0 (0.00)	0.030
Statin	517 (23.99)	3,907 (18.80)	0.127	49,883 (98.41)	41,638 (98.15)	0.215	6,671 (96.98)	3,640 (96.73)	0.014
Antipsychotic	83 (3.85)	595 (2.86)	0.055	1,368 (48.96)	1,142 (49.70)	0.007	238 (3.46)	141 (3.75)	0.015
Antidepressant	247 (11.46)	1,892 (9.10)	0.078	7,726 (84.96)	6,161 (84.54)	0.027	1,624 (23.61)	894 (23.76)	0.004
NSAIDs	300 (13.92)	2,457 (11.82)	0.063	6,607 (59.65)	5,450 (59.98)	0.001	1,085 (15.77)	574 (15.25)	0.014
Influenza infection or ILI, N (%)	2 (0.09)	35 (0.17)	0.021	170 (0.34)	103 (0.24)	0.004	42 (0.61)	17 (0.45)	0.022
Seasonal influenza vaccination, N (%)									
Year 2008	1,938 (89.93)	14,056 (67.63)	0.567	49,090 (96.84)	40,774 (96.11)	0.548	0 (0.00)	0 (0.00)	-
Year 2009	2,055 (95.36)	14,424 (69.40)	0.725	50,632 (99.89)	42,395 (99.93)	0.689	6,698 (97.37)	2,642 (70.21)	0.793
Charlson comorbidity index, N (%)									
0	2,113 (98.05)	20,546 (98.85)	0.065	43,594 (86.00)	35,097 (82.73)	0.209	0 (0.00)	0 (0.00)	-
1	2 (0.09)	11 (0.05)	0.015	43,594 (86.00)	35,097 (82.73)	0.245	6,586 (95.74)	3,601 (95.69)	0.002
2	39 (1.81)	219 (1.05)	0.064	7,085 (13.98)	7,309 (17.23)	0.113	293 (4.26)	162 (4.31)	0.002
3 or more	1 (0.05)	8 (0.04)	0.004	11 (0.02)	17 (0.04)	0.011	0 (0.00)	0 (0.00)	-

Table 9: CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of ordinary propensity score using model 3 (true confounders) under complete case method.

	Below the range			Within the range			Above the range		
Covariates	Exposed ($N = 5,934$)	Unexposed ($N = 65,382$)	Stand. diff.	Exposed ($N = 93,452$)	Unexposed ($N = 157,833$)	Stand. diff.	Exposed ($N = 19,304$)	Unexposed ($N = 11,759$)	Stand. diff.
Age in years, mean±SD	73.77 ± 7.15	72.76 ± 7.02	0.142	74.63 ± 7.05	74.82 ± 7.49	0.026	72.18 ± 5.97	72.41 ± 6.14	0.038
Age category, N (%)									
65-69 years	1,982 (33.40)	25,994 (39,76)	0.132	26,181 (28.02)	46,488 (29.45)	0.032	7,477 (38.73)	4,510 (38.35)	0.008
70-74 years	1,519 (25.60)	16,605 (25.40)	0.005	22,820 (24.42)	36,700 (23.29)	0.026	5,551 (28.76)	3,294 (28.01)	0.016
75-79 years	1,131 (19.06)	11,201 (17.13)	0.050	20,780 (22.24)	31,717 (20.10)	0.052	3,751 (19.43)	2,290 (19.47)	0.001
80-84 years	757 (12.76)	6,562 (10.04)	0.086	14,343 (15.35)	23,607 (14.96)	0.011	1,878 (9.73)	1,149 (9.77)	0.001
$\geq$ 85 years	545 (9.18)	5,020 (7.68)	0.054	9,328 (9.98)	19,261 (12.20)	0.071	647 (3.35)	516 (4.39)	0.054
Number of consultations, median (IQR)	27 (20)	20 (17)	0.420	37 (24)	31 (22)	0.261	41 (27)	38 (27)	0.100
Male, N (%)	2,251 (37.93)	22,078 (38.77)		45,257 (48.43)	70,561 (44.71)	0.075	14,506 (75.15)	8,705 (74.03)	0.026
Region, N (%)									
North East	131 (2.21)	1,105 (1.69)	0.037	2,147 (2.30)	3,449 (2.19)	0.008	486 (2.52)	367 (3.12)	0.036
North West	929 (15.66)	8,738 (13.36)	0.065	12,718 (13.61)	21,645 (13.71)	0.003	585 (2.93)	491 (4.18)	0.068
Yorkshire and The Humber	152 (2.56)	1,672 (2.56)	0.000	2,682 (2.87)	4,230 (2.68)	0.012	807 (4.18)	525 (4.46)	0.014
East Midlands	180 (3.03)	1,742 (2.66)	0.022	3,489 (3.73)	5,410 (3.43)	0.016	730 (3.78)	506 (4.30)	0.026
West Midlands	527 (8.88)	5,455 (8.34)	0.019	8,053 (8.62)	12,813 (8.12)	0.018	775 (4.01)	620 (5.27)	0.060
East of England	560 (9.44)	6,855 (10.48)	0.035	7,677 (8.21)	13,343 (8.45)	0.009	156 (0.81)	125 (1.06)	0.026
South West	520 (8.76)	6,262 (9.58)	0.028	8,913 (9.54)	14,222 (9.01)	0.018	1,018 (5.27)	787 (6.69)	0.060
South Central	678 (11.43)	9,622 (14.72)	0.098	10,004 (10.70)	17,613 (11.16)	0.015	235 (1.22)	172 (1.46)	0.021
London	712 (12.00)	7,923 (12.12)	0.004	7,846 (8.40)	14,500 (9.19)	0.028	76 (0.39)	75 (0.64)	0.034
South East Coast	721 (12.15)	8,485 (12.98)	0.025	9,535 (10.20)	16,036 (10.16)	0.001	675 (3.50)	471 (4.01)	0.027
Northern Ireland	134 (2.26)	1,384 (2.12)	0.010	2,991 (3.20)	5,524 (3.50)	0.017	2,097 (10.86)	1,092 (9.29)	0.052
Scotland	181 (3.05)	1,168 (1.79)	0.082	8,163 (8.73)	14,687 (9.31)	0.020	8,573 (44.41)	4,547 (38.67)	0.117
Wales	509 (8.58)	4,971 (7.60)	0.036	9,234 (9.88)	14,361 (9.10)	0.027	3,111 (16.12)	1,981 (16.85)	0.020
Cardiovascular risk factors, N (%)									
Diabetes mellitus	0 (0.00)	0 (0.00)	-	23,046 (24.66)	26,539 (16.81)	0.194	10,229 (52.99)	6,112 (51.98)	0.020
Myocardial infarction	0 (0.00)	0 (0.00)	-	7,110 (7.61)	8,318 (5.27)	0.095	2,448 (12.68)	1,250(10.63)	0.064
Congestive heart failure	0 (0.00)	0 (0.00)	-	2,280 (2.44)	3,159 (2.00)	0.030	537 (2.78)	340 (2.89)	0.007
Lifestyle risk factors, N (%)									
Alcohol intake[a]									
No	721 (41.16)	6,723 (37.28)	0.079	12,224 (36.30)	20,006 (37.42)	0.023	2,473 (29.67)	1,593 (32.06)	0.052
Current	897 (51.85)	10,321 (57.24)	0.108	18,931 (56.21)	29,761 (55.66)	0.011	5,176 (62.11)	2,981 (59.99)	0.043
Former	121 (6.99)	988 (5.48)	0.063	2,523 (7.49)	3,701 (6.92)	0.022	685 (8.22)	395 (7.95)	0.010
Tobacco use and smoking[a]									
No	2,174 (43.65)	29,431 (55.60)	0.241	37,510 (44.77)	65,469 (47.71)	0.059	6,502 (35.77)	3,934 (35.85)	0.002
Current	781 (15.68)	6,146 (11.61)	0.119	8,248 (9.85)	17,455 (12.72)	0.091	2,274 (12.51)	1,884 (17.17)	0.131
Former	2,025 (40.66)	17,357 (32.79)	0.164	38,017 (45.38)	54,286 (39.56)	0.118	9,403 (51.72)	5,154 (46.97)	0.095
Body mass index									
Normal	4 (0.07)	51 (0.08)	0.004	91 (0.10)	181 (0.11)	0.005	58 (0.30)	24 (0.20)	0.019
Underweight	0 (0.00)	4 (0.01)	0.011	14 (0.01)	19 (0.01)	0.003	8 (0.04)	5 (0.04)	0.001
Overweight	32 (0.54)	326 (0.50)	0.006	383 (0.41)	719 (0.46)	0.007	46 (0.24)	26 (0.22)	0.004
Unknown	5,898 (99.39)	65,001 (99.42)	0.003	92,964 (99.48)	156,914 (99.42)	0.008	19,192 (99.42)	11,704 (99.53)	0.016
Medications, N (%)									
Antiplatelet	642 (10.82)	4,091 (6.26)	0.164	45,234 (48.40)	62,488 (39.59)	0.178	16,246 (84.16)	10,064 (85.59)	0.040
Anticoagulant	131 (2.21)	657 (1.00)	0.096	6,620 (7.08)	9,107 (5.77)	0.054	3,272 (16.95)	2,066 (17.57)	0.016
Antihypertensive	2,116 (35.66)	22,839 (34.93)	0.015	68,843 (73.67)	106,076 (67.21)	0.142	18,127 (93.90)	11,036 (93.85)	0.002
Diuretics	1,517 (25.56)	15,178 (23.21)	0.055	37,507 (40.14)	60,263 (38.18)	0.040	8,999 (46.62)	5,413 (46.03)	0.012
Antidiabetic	0 (0.00)	0 (0.00)	-	(0.06)	76 (0.05)	0.005	28 (0.15)	11 (0.09)	0.015
Statin	868 (14.63)	8,420 (12.88)	0.051	58,333 (62.42)	82,270 (52.12)	0.209	18,601 (96.36)	11,322 (96.28)	0.004
Antipsychotic	222 (3.74)	1,627 (2.49)	0.072	2,889 (3.09)	5,015 (3.18)	0.005	455 (2.36)	305 (2.59)	0.015
Antidepressant	852 (14.36)	6,734 (10.30)	0.124	14,222 (15.22)	21,423 (13.57)	0.047	2,925 (15.15)	1,655 (14.07)	0.031
NSAIDs	882 (13.85)	8,221 (12.57)	0.038	12,470 (13.34)	21,200 (13.43)	0.003	3,086 (15.99)	1,775 (15.09)	0.025
Influenza infection or ILI, N (%)	14 (0.24)	102 (0.16)	0.018	249 (0.27)	404 (0.26)	0.002	132 (0.68)	63 (0.54)	0.019
Seasonal influenza vaccination, N (%)									
Year 2008	5,301 (89.33)	43,755 (66.92)	0.563	87,508 (93.64)	116,098 (73.56)	0.563	18,256 (94.57)	8,396 (71.40)	0.648
Year 2009	5,684 (95.79)	45,020 (68.86)	0.754	90,292 (96.62)	114,263 (72.39)	0.710	18,732 (97.04)	8,089 (68.79)	0.810
Charlson comorbidity index, N (%)									
0	5,930 (99.93)	65,344 (99.94)	0.004	9,125 (9.76)	31,404 (19.90)	0.288	0 (0.00)	0 (0.00)	-
1	0 (0.00)	0 (0.00)	-	68,305 (73.09)	93,111 (58.99)	0.301	18,636 (96.54)	11,355 (96.56)	0.001
2	0 (0.00)	10 (0.02)	0.017	15,942 (17.06)	33,078 (20.96)	0.099	668 (3.46)	404 (3.44)	0.001
3 or more	4 (0.07)	28 (0.04)	0.010	80 (0.09)	240 (0.15)	0.019	0 (0.00)	0 (0.00)	-

[a] Covariate was not included in the model

C.2 Overlapping Range under Missingness Pattern Method

Table 10: CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of generalized propensity score using model 1 (all covariates) under missingness pattern method.

	Below the range			Within the range			Above the range		
	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.
Covariates	$(N = 8,937)$	$(N = 273,957)$	diff.	$(N = 117,941)$	$(N = 292,834)$	diff.	$(N = 51,859)$	$(N = 29,832)$	diff.
Age in years, mean±SD	74.68 ± 8.58	73.04 ± 8.23	0.194	75.50 ± 7.51	75.12 ± 7.53	0.051	72.69 ± 6.13	73.41 ± 6.32	0.125
Age category, N (%)									
64-69 years	3,169 (35.46)	121,156 (44.22)	0.180	30,339 (25.72)	81,617 (27.87)	0.049	18,813 (36.28)	9,588 (32.14)	0.087
70-74 years	1,703 (19.06)	53,867 (19.66)	0.015	26,523 (22.49)	72,139 (24.63)	0.051	14,685 (28.32)	8,436 (28.28)	0.001
75-79 years	1,509 (16.88)	38,264 (14.10)	0.077	25,443 (21.57)	57,967 (19.80)	0.044	10,661 (20.56)	6,387 (21.41)	0.021
80-84 years	1,154 (12.91)	28,558 (10.42)	0.078	19,612 (16.63)	42,244 (14.43)	0.061	5,520 (10.64)	3,668 (12.30)	0.052
$\geq$ 85 years	1,402 (15.69)	31,752 (11.59)	0.120	16,024 (13.59)	38,867 (13.27)	0.009	2,180 (4.20)	1,753 (5.88)	0.077
Number of consultations, median (IQR)	21 (21)	15 (20)	0.345	32 (21)	27 (21)	0.238	45 (29)	44 (31)	0.028
Male, N (%)	3,362 (37.62)	112,024 (40.89)	0.067	51,450 (43.62)	121,826 (41.60)	0.041	34,211 (65.97)	19,006 (63.71)	0.047
Region, N (%)									
North East	117 (1.31)	4,708 (1.49)	0.015	2,199 (1.86)	5,044 (1.72)	0.011	1,313 (2.53)	788 (2.64)	0.007
North West	1,119 (12.52)	31,169 (11.38)	0.035	14,458 (12.26)	34,000 (11.61)	0.020	4,345 (8.38)	3,150 (10.56)	0.075
Yorkshire and The Humber	168 (1.88)	5,138 (1.88)	0.000	3,285 (2.79)	7,777 (2.66)	0.008	1,965 (3.79)	1,116 (3.74)	0.003
East Midlands	260 (2.91)	9,339 (3.41)	0.029	3,961 (3.36)	9,527 (3.25)	0.006	2,054 (3.96)	1,215 (4.07)	0.006
West Midlands	719 (8.05)	24,811 (9.06)	0.036	10,289 (8.72)	25,184 (8.60)	0.004	3,443 (6.64)	2,291 (7.68)	0.040
East of England	879 (9.84)	27,966 (10.21)	0.012	10,427 (8.84)	26,419 (9.02)	0.006	2,000 (3.86)	1,436 (4.81)	0.047
South West	711 (7.96)	27,222 (9.94)	0.069	11,869 (10.06)	29,163 (9.96)	0.003	4,220 (8.14)	2,668 (8.94)	0.029
South Central	1,337 (14.96)	38,316 (13.99)	0.028	12,957 (10.99)	33,253 (11.36)	0.012	2,615 (5.04)	1,912 (6.41)	0.059
London	881 (9.86)	27,602 (10.08)	0.007	10,398 (8.82)	26,005 (8.88)	0.002	1,783 (3.44)	1,280 (4.29)	0.044
South East Coast	1,305 (14.60)	31,353 (11.44)	0.094	13,382 (11.35)	34,898 (11.92)	0.018	3,584 (6.91)	2,392 (8.02)	0.042
Northern Ireland	177 (1.98)	5,031 (1.84)	0.011	3,473 (2.94)	9,005 (3.08)	0.008	3,484 (6.72)	1,565 (5.25)	0.062
Scotland	492 (5.51)	15,959 (5.83)	0.014	10,758 (9.12)	26,476 (9.04)	0.003	13,300 (25.65)	5,837 (19.57)	0.146
Wales	772 (8.64)	25,973 (9.48)	0.029	10,485 (8.89)	26,083 (8.91)	0.001	7,753 (14.95)	4,182 (14.02)	0.026
Cardiovascular risk factors, N (%)									
Diabetes mellitus	443 (4.96)	7,632 (2.79)	0.113	13,378 (11.34)	20,313 (6.94)	0.153	21,765 (41.97)	11,989 (40.19)	0.036
Myocardial infarction	171 (1.91)	3,103 (1.13)	0.064	5,928 (5.03)	9,056 (3.09)	0.098	7,019 (13.53)	3,810 (12.77)	0.023
Congestive heart failure	104 (1.16)	2,049 (0.75)	0.043	2,900 (2.46)	4,841 (1.65)	0.057	1,635 (3.15)	998 (3.35)	0.011
Lifestyle risk factors, N (%)									
Alcohol intake[a]									
No	482 (45.39)	8,895 (39.55)	0.118	10,040 (38.27)	21,235 (37.81)	0.010	6,285 (31.38)	3,789 (34.39)	0.064
Current	520 (48.96)	12,139 (53.97)	0.100	14,338 (54.65)	31,175 (55.51)	0.017	11,953 (59.68)	6,285 (57.04)	0.054
Former	60 (5.65)	1,459 (6.49)	0.035	1,856 (7.07)	3,756 (6.69)	0.015	1,789 (8.93)	944 (8.57)	0.013
Tobacco use and smoking[a]									
No	2,141 (45.39)	55,018 (47.63)	0.045	41,109 (46.66)	99,111 (48.44)	0.035	16,254 (34.75)	9,094 (34.25)	0.011
Current	898 (19.04)	21,856 (18.92)	0.003	10,385 (11.79)	24,740 (12.09)	0.009	4,526 (9.68)	2,827 (10.65)	0.032
Former	1,678 (35.57)	38,626 (33.44)	0.045	36,601 (41.55)	80,772 (39.47)	0.042	25,988 (55.57)	14,633 (55.11)	0.009
Body mass index[a]									
Normal	1 (0.04)	28 (0.05)	0.004	101 (0.15)	188 (0.13)	0.005	51 (0.11)	40 (0.15)	0.012
Underweight	0 (0.00)	4 (0.01)	0.011	12 (0.02)	16 (0.01)	0.006	10 (0.02)	8 (0.03)	0.006
Overweight	10 (0.38)	350 (0.57)	0.027	295 (0.43)	632 (0.43)	0.000	156 (0.33)	89 (0.33)	0.001
Unknown	2,620 (99.58)	61,130 (99.38)	0.028	68,015 (99.40)	145,852 (99.43)	0.003	47,419 (99.54)	26,637 (99.49)	0.008
Medications, N (%)									
Antiplatelet	1,486 (16.63)	30,035 (10.96)	0.165	43,781 (37.12)	84,683 (28.92)	0.175	38,444 (74.13)	22,101 (74.08)	0.001
Anticoagulant	275 (3.08)	4,763 (1.74)	0.087	7,974 (6.76)	14,719 (5.03)	0.074	6,533 (12.60)	4,013 (13.45)	0.025
Antihypertensive	3,145 (35.19)	76,591 (27.96)	0.156	75,461 (63.98)	163,692 (55.90)	0.166	45,693 (88.11)	26,138 (87.62)	0.015
Diuretics	2,065 (23.11)	48,588 (17.74)	0.134	44,006 (37.31)	97,932 (33.44)	0.081	22,721 (43.81)	13,128 (44.01)	0.004
Antidiabetic	0 (0.00)	13 (0.00)	0.010	34 (0.03)	53 (0.02)	0.007	54 (0.10)	24 (0.08)	0.008
Statin	1,909 (21.36)	40,296 (14.71)	0.174	55,841 (47.35)	111,936 (38.23)	0.185	45,187 (87.13)	25,748 (86.31)	0.024
Antipsychotic	300 (3.36)	5,402 (1.97)	0.086	3,774 (3.20)	8,856 (3.02)	0.010	1,520 (2.93)	1,005 (3.37)	0.025
Antidepressant	1,052 (11.77)	21,785 (7.95)	0.128	16,702 (14.16)	38,502 (13.15)	0.030	9,183(17.71)	5,331 (17.87)	0.004
NSAIDs	1,081 (12.10)	26,472 (9.66)	0.078	16,515 (14.00)	40,738 (13.91)	0.003	7,378 (14.23)	4,242 (14.22)	0.000
Influenza infection or ILI, N (%)	14 (0.16)	353 (0.13)	0.007	309 (0.26)	705 (0.24)	0.004	252 (0.49)	127 (0.43)	0.009
Seasonal influenza vaccination, N (%)									
Year 2008	4,306 (48.18)	71,487 (26.09)	0.470	109,845 (93.14)	263,298 (89.91)	0.116	51,164 (98.66)	29,330 (98.32)	0.028
Year 2009	3,812 (42.65)	58,590 (21.39)	0.468	115,567 (97.99)	275,753 (94.17)	0.198	51,844 (99.97)	29,821 (99.96)	0.004
Charlson comorbidity index, N (%)									
0	5,413 (60.57)	197,358 (72.04)	0.244	23,890 (20.26)	120,008 (40.98)	0.461	34 (0.07)	29 (0.10)	0.011
1	2,734 (26.56)	50,704 (18.51)	0.194	68,313 (57.92)	118,268 (40.39)	0.356	48,066 (92.69)	27,445 (92.00)	0.026
2	1,141 (12.77)	25,600 (9.34)	0.109	25,578 (21.69)	54,115 (18.48)	0.080	3,746 (7.22)	2,351 (7.88)	0.025
3 or more	9 (0.10)	295 (0.11)	0.002	160 (0.14)	443 (0.15)	0.004	13 (0.03)	7 (0.02)	0.001

[a] Covariate with missing value

Table 11: CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of generalized propensity score using model 2 (potential confounders) under missingness pattern method.

	Below the range			Within the range			Above the range		
	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.	Exposed	Unexposed	Stand.
Covariates	($N = 8,937$)	($N = 196,762$)	diff.	($N = 129,253$)	($N = 370,029$)	diff.	($N = 40,547$)	($N = 29,832$)	diff.
Age in years, mean±SD	73.99 ± 7.82	72.68 ± 7.69	0.169	75.21 ± 7.50	74.88 ± 7.96	0.043	72.88 ± 6.22	73.35 ± 6.44	0.073
Age category, N (%)									
64-69 years	3,199 (35.80)	86,420 (43.92)	0.167	34,976 (27.06)	116,071 (31.37)	0.095	14,146 (34.89)	9,870 (33.09)	0.038
70-74 years	2,061 (23.06)	43,907 (22.31)	0.018	29,538 (22.85)	82,404 (22.27)	0.014	11,312 (27.90)	8,131 (27.26)	0.014
75-79 years	1,550 (17.34)	29,150 (14.81)	0.069	27,517 (21.29)	67,526 (18.25)	0.076	8,546 (21.08)	6,302 (21.12)	0.001
80-84 years	1,082 (12.11)	18,961 (9.64)	0.079	20,501 (15.86)	51,779 (13.99)	0.052	4,703 (11.60)	3,730 (12.50)	0.028
$\geq$ 85 years	1,045 (11.69)	18,324 (9.31)	0.078	16,721 (12.94)	52,249 (14.12)	0.035	1,840 (4.54)	1,799 (6.03)	0.067
Number of consultations, median (IQR)	21 (19)	12 (18)	0.629	34 (23)	26 (22)	0.378	41 (27)	38 (26)	0.100
Male, N (%)	3,269 (36.58)	77,299 (39.29)	0.056	57,589 (44.56)	155,346 (41.98)	0.052	28,165 (69.46)	20,211 (67.75)	0.037
Region, N (%)									
North East	127 (1.42)	2,484 (1.26)	0.014	2,465 (1.91)	6,586 (1.78)	0.009	1,037 (2.56)	840 (2.82)	0.016
North West	1,137 (12.72)	21,979 (11.17)	0.048	15,778 (12.21)	43,557 (11.77)	0.013	3,007 (7.42)	2,783 (9.33)	0.069
Yorkshire and The Humber	180 (2.01)	3,060 (1.56)	0.035	3,667 (2.84)	9,826 (2.66)	0.011	1,571 (3.87)	1,145 (3.84)	0.002
East Midlands	299 (3.35)	6,703 (3.41)	0.003	4,239 (3.28)	12,073 (3.26)	0.001	1,737 (4.28)	1,305 (4.37)	0.004
West Midlands	788 (8.82)	18,400 (9.35)	0.019	10,997 (8.51)	31,604 (8.54)	0.001	2,666 (6.58)	2,282 (7.65)	0.042
East of England	941 (10.53)	22,052 (11.21)	0.022	10,881 (8.42)	32,485 (8.78)	0.013	1,484 (3.66)	1,284 (4.30)	0.033
South West	765 (8.56)	20,012 (10.17)	0.055	12,945 (10.02)	36,513 (9.87)	0.005	3,090 (7.62)	2,528 (8.47)	0.031
South Central	1,216 (13.61)	29,394 (14.94)	0.038	13,922 (10.77)	42,555 (11.50)	0.023	1,771 (4.37)	1,532 (5.14)	0.036
London	1,101 (12.32)	20,957 (10.65)	0.052	10,843 (8.39)	32,888 (8.89)	0.018	1,118 (2.76)	1,042 (3.49)	0.042
South East Coast	1,104 (11.35)	23,934 (12.16)	0.025	14,811 (11.46)	42,705 (11.54)	0.003	2,446 (6.03)	2,004 (6.72)	0.028
Northern Ireland	119 (1.33)	2,418 (1.23)	0.009	3,970 (3.07)	11,320 (3.06)	0.001	3,045 (7.51)	1,863 (6.24)	0.050
Scotland	478 (5.35)	7,985 (4.06)	0.061	12,252 (9.48)	33,241 (8.98)	0.017	11,820 (29.15)	7,046 (23.62)	0.126
Wales	772 (8.64)	17,384 (8.84)	0.007	12,483 (9.66)	34,676 (9.37)	0.010	5,755 (14.19)	4,178 (14.01)	0.005
Cardiovascular risk factors, N (%)									
Diabetes mellitus	23 (0.26)	225 (0.11)	0.033	17,167 (13.28)	26,494 (7.16)	0.203	18,396 (45.37)	13,215 (44.30)	0.022
Myocardial infarction	0 (0.00)	0 (0.00)	-	7,602 (5.88)	12,303 (3.32)	0.122	5,516 (13.60)	3,666 (12.29)	0.039
Congestive heart failure	0 (0.00)	0 (0.00)	-	3,458 (2.68)	7,025 (1.90)	0.052	1,181 (2.91)	863 (2.89)	0.001
Lifestyle risk factors, N (%)									
Alcohol intake[a]									
No	190 (47.86)	2,995 (46.80)	0.021	11,431 (38.38)	26,950 (37.94)	0.009	5,186 (30.25)	3,974 (32.46)	0.048
Current	184 (46.35)	3,111 (48.61)	0.045	16,169 (54.29)	39,319 (55.35)	0.021	10,458 (61.00)	7,169 (58.55)	0.050
Former	23 (5.79)	294 (4.59)	0.054	2,181 (7.32)	4,764 (6.71)	0.024	1,501 (8.75)	1,101 (8.99)	0.008
Tobacco use and smoking[a]									
No	1,556 (42.09)	31,376 (50.31)	0.165	45,256 (46.14)	122,779 (47.83)	0.034	12,692 (33.58)	9,068 (32.85)	0.016
Current	990 (26.78)	13,104 (21.01)	0.136	12,332 (12.57)	34,223 (13.33)	0.023	2,487 (6.58)	2,096 (7.59)	0.039
Former	1,151 (31.13)	17,891 (28.68)	0.053	40,495 (41.29)	99,696 (38.84)	0.050	22,621 (59.84)	16,444 (59.56)	0.006
Body mass index[a]									
Normal	0 (0.00)	0 (0.00)	-	71 (0.09)	222 (0.12)	0.009	82 (0.21)	34 (0.12)	0.022
Underweight	0 (0.00)	0 (0.00)	-	11 (0.01)	21 (0.01)	0.003	11 (0.03)	7 (0.02)	0.002
Overweight	9 (0.67)	140 (0.74)	0.008	349 (0.45)	846 (0.45)	0.001	103 (0.26)	85 (0.30)	0.007
Unknown	1,333 (99.33)	18,809 (99.26)	0.008	77,973 (99.45)	185,570 (99.42)	0.004	38,748 (99.50)	28,240 (99.56)	0.009
Medications, N (%)									
Antiplatelet	564 (6.31)	6,145 (3.12)	0.008	50,598 (39.15)	106,651 (28.80)	0.220	32,549 (80.27)	24,113 (80.83)	0.014
Anticoagulant	102 (1.14)	874 (0.44)	0.079	8,876 (6.87)	18,316 (4.95)	0.081	5,804 (14.31)	4,305 (14.43)	0.003
Antihypertensive	1,876 (20.99)	31,520 (16.02)	0.128	85,329 (66.02)	207,747 (56.14)	0.204	8,655 (92.44)	3,485 (92.56)	0.016
Diuretics	1,483 (16.59)	22,720 (11.55)	0.146	48,803 (37.76)	123,546 (33.39)	0.091	318,506 (45.64)	13,382 (44.86)	0.016
Antidiabetic	0 (0.00)	1 (0.00)	0.003	39 (0.03)	66 (0.02)	0.008	49 (0.12)	23 (0.08)	0.014
Statin	788 (8.82)	11,042 (5.61)	0.124	64,327 (49.77)	139,140 (37.60)	0.247	37,822 (93.28)	27,798 (93.18)	0.004
Antipsychotic	251 (2.81)	2,902 (1.47)	0.092	4,222 (3.27)	11,444 (3.09)	0.010	1,121 (2.76)	917 (3.07)	0.018
Antidepressant	892 (9.98)	12,107 (6.15)	0.141	18,457 (14.28)	47,944 (12.96)	0.039	7,588 (18.71)	5,567 (18.66)	0.001
NSAIDs	990 (11.08)	16,847 (8.56)	0.085	17,894 (13.84)	50,177 (13.56)	0.008	6,090 (15.02)	4,428 (14.84)	0.005
Influenza infection or ILI, N (%)	12 (0.13)	182 (0.09)	0.012	347 (0.27)	867 (0.23)	0.007	216 (0.53)	136 (0.46)	0.011
Seasonal influenza vaccination, N (%)									
Year 2008	7,597 (85.01)	92,080 (46.80)	0.881	119,362 (92.35)	249,624 (67.46)	0.653	38,356 (94.60)	22,411 (75.12)	0.564
Year 2009	8,186 (91.60)	95,351 (48.46)	1.067	123,695 (95.70)	247,187 (66.80)	0.797	39,342 (97.03)	21,626 (72.49)	0.726
Charlson comorbidity index, N (%)									
0	8,913 (99.73)	196,508 (99.87)	0.031	20,421 (15.80)	120,885 (32.67)	0.402	3 (0.01)	2 (0.01)	0.001
1	22 (0.25)	234 (0.12)	0.030	80,383 (62.19)	167,965 (45.39)	0.342	38,348 (94.58)	28,218 (94.59)	0.001
2	2 (0.02)	10 (0.01)	0.015	28,270 (21.87)	80,448 (21.74)	0.003	2,193 (5.41)	1,608 (5.39)	0.001
3 or more	0 (0.00)	10 (0.01)	0.010	179 (0.14)	731 (0.20)	0.014	3 (0.01)	4 (0.01)	0.006

[a] Covariate with missing value

Table 12: CPRD Data. Comparison of baseline characteristics between exposed and unexposed cohorts based on overlapping range of logit of generalized propensity score using model 3 (true confounders) under missingness pattern method.

	Below the range			Within the range			Above the range		
Covariates	Exposed ($N = 8,935$)	Unexposed ($N = 196,528$)	Stand. diff.	Exposed ($N = 130,008$)	Unexposed ($N = 370,264$)	Stand. diff.	Exposed ($N = 39,794$)	Unexposed ($N = 29,831$)	Stand. diff.
Age in years, mean±SD	74.02 ± 7.82	72.70 ± 7.70	0.170	75.22 ± 7.50	74.88 ± 7.96	0.043	72.82 ± 6.17	73.21 ± 6.38	0.061
Age category, N (%)									
65-69 years	3,173 (35.51)	86,063 (43,79)	0.170	35,205 (27.08)	116,279 (31.40)	0.095	13,943 (35.04)	10,019 (33.59)	0.031
70-74 years	2,061 (23.07)	44,018 (22.40)	0.016	29,696 (22.84)	82,170 (22.19)	0.016	11,154 (28.03)	8,254 (27.67)	0.008
75-79 years	1,569 (17.56)	29,068 (14.79)	0.075	27,582 (21.22)	67,647 (18.27)	0.074	8,462 (21.26)	6,263 (20.99)	0.007
80-84 years	1,079 (12.08)	18,933 (9.63)	0.079	20,695 (15.92)	51,931 (14.03)	0.053	4,512 (11.34)	3,606 (12.09)	0.023
≥ 85 years	51,053 (11.79)	18,446 (9.39)	0.078	16,830 (12.95)	52,237 (14.11)	0.034	1,723 (4.33)	1,689 (5.66)	0.061
Number of consultations, median (IQR)	22 (20)	12 (18)	0.635	34 (24)	26 (22)	0.385	40 (27)	37 (25)	0.115
Male, N (%)	3,221 (36.05)	76,860 (39.11)	0.063	57,977 (44.59)	155,617 (42.03)	0.052	27,825 (69.92)	20,379 (68.31)	0.035
Region, N (%)									
North East	123 (1.38)	2,460 (1.25)	0.011	2,468 (1.90)	6,582 (1.78)	0.009	1,038 (2.61)	868 (2.91)	0.018
North West	1,135 (12.70)	21,855 (11.12)	0.049	16,010 (12.31)	43,819 (11.83)	0.015	2,777 (6.98)	2,645 (8.87)	0.070
Yorkshire and The Humber	184 (2.06)	3,129 (1.59)	0.035	3,651 (2.81)	9,738 (2.63)	0.011	1,583 (3.98)	1,164 (3.90)	0.004
East Midlands	302 (3.38)	6,724 (3.42)	0.002	4,242 (3.26)	12,045 (3.25)	0.001	1,731 (4.35)	1,312 (4.40)	0.002
West Midlands	786 (8.80)	18,348 (9.34)	0.019	11,004 (8.46)	31,626 (8.54)	0.003	2,661 (6.69)	2,312 (7.75)	0.041
East of England	939 (10.51)	22,025 (11.21)	0.022	10,959 (8.43)	32,518 (8.78)	0.013	1,408 (3.54)	1,278 (4.28)	0.038
South West	772 (8.64)	19,934 (10.14)	0.052	12,934 (9.95)	36,527 (9.87)	0.003	3,094 (7.78)	2,592 (8.69)	0.033
South Central	1,207 (13.51)	29,479 (15.00)	0.043	14,077 (10.83)	42,517 (11.48)	0.021	1,625 (4.08)	1,485 (4.98)	0.043
London	1,117 (12.50)	21,183 (10.78)	0.054	10,926 (8.40)	32,726 (8.84)	0.015	1,019 (2.56)	978 (3.28)	0.043
South East Coast	1,006 (11.26)	23,858 (12.14)	0.027	14,921 (11.48)	42,853 (11.57)	0.003	2,344 (5.89)	1,932 (6.48)	0.024
Northern Ireland	120 (1.34)	2,307 (1.17)	0.015	3,976 (3.06)	11,451 (3.09)	0.002	3,038 (7.63)	1,843 (6.18)	0.057
Scotland	454 (5.08)	7,722 (3.93)	0.056	12,372 (9.52)	33,375 (9.01)	0.017	11,724 (29.46)	7,175 (24.05)	0.122
Wales	790 (8.84)	17,504 (8.91)	0.002	12,468 (9.59)	34,487 (9.31)	0.009	5,752 (14.45)	4,247 (14.24)	0.006
Cardiovascular risk factors, N (%)									
Diabetes mellitus	20 (0.22)	149 (0.08)	0.038	17,129 (13.18)	26,248 (7.09)	0.203	18,437 (46.33)	13,537 (45.38)	0.019
Myocardial infarction	0 (0.00)	0 (0.00)	-	7,818 (6.01)	12,366 (3.34)	0.127	5,300 (13.32)	3,603(12.08)	0.037
Congestive heart failure	0 (0.00)	0 (0.00)	-	3,481 (2.68)	7,065 (1.91)	0.051	1,158 (2.91)	823 (2.76)	0.009
Lifestyle risk factors, N (%)									
Alcohol intake[a]									
No	159 (42.63)	2,539 (41.65)	0.020	11,231 (37.49)	27,001 (37.95)	0.009	5,417 (31.88)	4,379 (35.25)	0.071
Current	190 (50.94)	3,190 (52.33)	0.028	16,468 (54.97)	39,372 (55.33)	0.007	10,153 (59.76)	7,037 (56.64)	0.063
Former	24 (6.43)	367 (6.02)	0.017	2,261 (7.55)	4,785 (6.72)	0.032	1,420 (8.36)	1,007 (8.11)	0.009
Tobacco use and smoking[a]									
No	1,516 (41.59)	30,593 (49.50)	0.159	44,076 (44.58)	122,264 (47.52)	0.059	13,912 (37.54)	10,366 (37.56)	0.000
Current	627 (17.20)	9,372 (15.16)	0.055	10,858 (10.98)	35,663 (13.86)	0.087	4,324 (11.67)	4,388 (15.90)	0.123
Former	1,502 (41.21)	21,845 (35.34)	0.121	43,946 (44.44)	99,340 (38.61)	0.119	18,819 (50.79)	12,846 (46.54)	0.085
Body mass index									
Normal	0 (0.00)	0 (0.00)	-	71 (0.09)	222 (0.12)	0.009	82 (0.21)	34 (0.12)	0.023
Underweight	0 (0.00)	0 (0.00)	-	13 (0.02)	19 (0.01)	0.006	9 (0.02)	9 (0.03)	0.005
Overweight	10 (0.78)	148 (0.82)	0.004	356 (0.45)	843 (0.45)	0.000	95 (0.25)	80 (0.28)	0.006
Unknown	1,275 (99.22)	17,986 (99.18)	0.004	78,669 (99.44)	187,332 (99.42)	0.003	38,110 (99.51)	28,301 (99.57)	0.008
Medications, N (%)									
Antiplatelet	580 (6.49)	6,157 (3.13)	0.157	50,780 (39.06)	106,197 (28.68)	0.221	32,351 (81.30)	24,465 (82.01)	0.018
Anticoagulant	103 (1.15)	865 (0.44)	0.080	8,880 (6.83)	18,273 (4.94)	0.081	5,799 (14.57)	4,357 (14.61)	0.001
Antihypertensive	1,876 (21.00)	31,483 (16.02)	0.128	85,810 (66.00)	207,594 (56.07)	0.205	36,613 (92.01)	27,344 (91.66)	0.013
Diuretics	1,477 (16.53)	22,378 (11.39)	0.149	49,158 (37.81)	123,898 (33.46)	0.091	18,157 (45.63)	13,372 (44.83)	0.016
Antidiabetic	0 (0.00)	1 (0.00)	0.003	38 (0.03)	64 (0.02)	0.008	50 (0.13)	25 (0.08)	0.013
Statin	732 (8.19)	10,785 (5.49)	0.107	64,800 (49.84)	139,214 (37.60)	0.249	37,405 (94.00)	27,981 (93.80)	0.008
Antipsychotic	282 (3.16)	3,102 (1.58)	0.104	4,291 (3.30)	11,329 (3.06)	0.014	1,021 (2.57)	832 (2.79)	0.014
Antidepressant	1,122 (12.56)	14,296 (7.27)	0.177	19,819 (15.24)	47,171 (12.74)	0.072	5,996 (15.07)	4,151 (13.92)	0.033
NSAIDs	1,032 (11.55)	17,093 (8.70)	0.095	17,928 (13.79)	49,987 (13.50)	0.008	6,014 (15.11)	4,372 (14.66)	0.013
Influenza infection or ILI, N (%)	12 (0.13)	177 (0.09)	0.013	339 (0.26)	858 (0.23)	0.006	224 (0.56)	150 (0.50)	0.008
Seasonal influenza vaccination, N (%)									
Year 2008	7,601 (85.07)	92,412 (47.02)	0.877	120,093 (92.37)	249,595 (67.41)	0.655	37,621 (94.54)	22,108 (74.11)	0.585
Year 2009	8,183 (91.58)	95,635 (48.66)	1.062	124,450 (95.72)	247,226 (66.77)	0.799	38,590 (96.97)	21,303 (71.41)	0.748
Charlson comorbidity index, N (%)									
0	8,914 (99.76)	196,362 (99.92)	0.038	20,421 (15.71)	121,030 (32.69)	0.404	2 (0.01)	3 (0.01)	0.006
1	19 (0.21)	155 (0.08)	0.035	80,954 (62.27)	167,938 (45.36)	0.344	37,780 (94.94)	28,324 (94.95)	0.000
2	2 (0.02)	7 (0.00)	0.017	28,451 (21.88)	80,558 (21.76)	0.003	2,012 (5.06)	1,501 (5.03)	0.001
3 or more	0 (0.00)	4 (0.00)	0.006	182 (0.14)	738 (0.20)	0.014	0 (0.00)	3 (0.01)	0.014

[a] Covariate was not included in the model

D SAS Code

D.1 Fitting the Propensity Score Model under Complete Case Method

```
libname new "Z:\Biostat Sub-Teams\Epi Stat Team\student thesis\Kurnia 2014-15\Thesis CPRD\DATA";
data cprd2; set new.clear;
run;
/*Estimating Propensity Score under Complete Case Method*/
proc logistic data = cprd2 descending;
model  H1N1_cohort = age2009 N_consult12 male region2 region3 region4 region5 region6
region7 region8 region9 region10 region11 region12 region13
diabetes myocardial chf currentalcohol formeralcohol currentsmoking formersmoking
underweight overweight unknown antiplatelet anticoagulant antihypert diuretics
antidiabetic statin antipsychotic antidepressant nsaids
flag_ili flag_FLU2008 flag_FLU2009 CCIcat1 CCIcat2 CCIcat3;
output out=out_ps2 prob=ps xbeta=logit_ps;
run;
```

D.2 Fitting the Propensity Score Model under Missingness Pattern Method

```
libname new "Z:\Biostat Sub-Teams\Epi Stat Team\student thesis\Kurnia 2014-15\Thesis CPRD\DATA";
data pattern; set new.clear;

/*To obtain patterns of missing data*/
proc mi data=pattern nimpute=0;
ods select misspattern;
run;

/*Adding Missing Indicators to Each Covariates with Missing Values*/
data pattern; set pattern;
if noalcohol = . then miss_noalcohol = 1;
if noalcohol ge 0 then miss_noalcohol = 0;
if currentalcohol = . then miss_currentalcohol = 1;
if currentalcohol ge 0 then miss_currentalcohol = 0;
if formeralcohol = . then miss_formeralcohol = 1;
if formeralcohol ge 0 then miss_formeralcohol = 0;
if nosmoking = . then miss_nosmoking = 1;
if nosmoking ge 0 then miss_nosmoking = 0;
if currentsmoking = . then miss_currentsmoking = 1;
if currentsmoking ge 0 then miss_currentsmoking = 0;
if formersmoking =  . then miss_formersmoking = 1;
if formersmoking ge 0 then miss_formersmoking = 0;
if normal = . then miss_normal = 1;
if normal ge 0 then miss_normal = 0;
if underweight = . then miss_underweight = 1;
if underweight ge 0 then miss_underweight = 0;
if overweight = . then miss_overweight = 1;
if overweight ge 0 then miss_overweight = 0;
if unknown = . then miss_unknown = 1;
if unknown ge 0 then miss_unknown = 0;
run;

/*Adding Patterns of Missing Data Indicators*/
data pattern; set pattern;
if miss_noalcohol ne 1 and miss_currentalcohol ne 1 and miss_formeralcohol ne 1 and
miss_nosmoking ne 1 and miss_currentsmoking ne 1 and miss_formersmoking ne 1 and
miss_normal ne 1 and miss_underweight ne 1 and miss_overweight ne 1 and miss_unknown ne 1 then MP = 1;
if miss_normal = 1 and miss_underweight = 1 and miss_overweight = 1 and miss_unknown = 1 then MP = 2;
if miss_nosmoking = 1 and miss_currentsmoking = 1 and miss_formersmoking = 1 then MP = 3;
if miss_noalcohol = 1 and miss_currentalcohol = 1 and miss_formeralcohol = 1 then MP = 4;
if miss_noalcohol = 1 and miss_currentalcohol = 1 and miss_formeralcohol = 1 and
miss_nosmoking = 1 and miss_currentsmoking = 1 and miss_formersmoking = 1 then MP = 5;
if miss_nosmoking = 1 and miss_currentsmoking = 1 and miss_formersmoking = 1 and
miss_normal = 1 and miss_underweight = 1 and miss_overweight = 1 and miss_unknown = 1 then MP = 6;
if miss_noalcohol = 1 and miss_currentalcohol = 1 and miss_formeralcohol = 1 and
miss_normal = 1 and miss_underweight = 1 and miss_overweight = 1 and miss_unknown = 1 then MP = 7;
if miss_noalcohol = 1 and miss_currentalcohol = 1 and miss_formeralcohol = 1 and
miss_nosmoking = 1 and miss_currentsmoking = 1 and miss_formersmoking = 1 and
miss_normal = 1 and miss_underweight = 1 and miss_overweight = 1 and miss_unknown = 1 then MP = 8;
run;
```

```
/*Frequencies of Patterns Missing Data in Both Cohorts*/
proc freq data=pattern;
table MP*cohort;
run;

/*Estimating Propensity Score with Separate Logit Models based on Model 1 (All Covariates)*/
proc logistic data = pattern descending;
model  H1N1_cohort = age2009 N_consult12
male region2 region3 region4 region5 region6
region7 region8 region9 region10 region11 region12 region13
diabetes myocardial chf
currentalcohol formeralcohol
currentsmoking formersmoking
underweight overweight unknown
antiplatelet anticoagulant antihypert diuretics antidiabetic statin
antipsychotic antidepressant nsaids
flag_ili flag_FLU2008 flag_FLU2009 CCIcat1 CCIcat2 CCIcat3;
output out=out_ps_mp1 prob=ps xbeta=logit_ps;
where MP = 1;
run;

proc logistic data = pattern descending;
model  H1N1_cohort = age2009 N_consult12
male region2 region3 region4 region5 region6
region7 region8 region9 region10 region11 region12 region13
diabetes myocardial chf
currentalcohol formeralcohol
currentsmoking formersmoking
/*underweight overweight unknown*/
antiplatelet anticoagulant antihypert diuretics antidiabetic statin antipsychotic
antidepressant nsaids
flag_ili flag_FLU2008 flag_FLU2009 CCIcat1 CCIcat2 CCIcat3;
output out=out_ps_mp2 prob=ps xbeta=logit_ps;
where MP = 2;
run;

proc logistic data = pattern descending;
model  H1N1_cohort = age2009 N_consult12
male region2 region3 region4 region5 region6
region7 region8 region9 region10 region11 region12 region13
diabetes myocardial chf
currentalcohol formeralcohol
/*currentsmoking formersmoking*/
underweight overweight unknown
antiplatelet anticoagulant antihypert diuretics antidiabetic statin antipsychotic
antidepressant nsaids
flag_ili flag_FLU2008 flag_FLU2009 CCIcat1 CCIcat2 CCIcat3;
output out=out_ps_mp3 prob=ps xbeta=logit_ps;
where MP = 3;
run;

proc logistic data = pattern descending;
model  H1N1_cohort = age2009 N_consult12
male region2 region3 region4 region5 region6
region7 region8 region9 region10 region11 region12 region13
diabetes myocardial chf
/*currentalcohol formeralcohol*/
currentsmoking formersmoking
underweight overweight unknown
antiplatelet anticoagulant antihypert diuretics antidiabetic statin antipsychotic
antidepressant nsaids
flag_ili flag_FLU2008 flag_FLU2009 CCIcat1 CCIcat2 CCIcat3;
output out=out_ps_mp4 prob=ps xbeta=logit_ps;
where MP = 4;
run;

proc logistic data = pattern descending;
model  H1N1_cohort = age2009 N_consult12
male region2 region3 region4 region5 region6
region7 region8 region9 region10 region11 region12 region13
diabetes myocardial chf
```

```
/*currentalcohol formeralcohol
currentsmoking formersmoking*/
underweight overweight unknown
antiplatelet anticoagulant antihypert diuretics antidiabetic statin antipsychotic
antidepressant nsaids
flag_ili flag_FLU2008 flag_FLU2009 CCIcat1 CCIcat2 CCIcat3;
output out=out_ps_mp5 prob=ps xbeta=logit_ps;
where MP = 5;
run;

proc logistic data = pattern descending;
model  H1N1_cohort = age2009 N_consult12
male region2 region3 region4 region5 region6
region7 region8 region9 region10 region11 region12 region13
diabetes myocardial chf
currentalcohol formeralcohol
/*currentsmoking formersmoking
underweight overweight unknown*/
antiplatelet anticoagulant antihypert diuretics antidiabetic statin antipsychotic
antidepressant nsaids
flag_ili flag_FLU2008 flag_FLU2009 CCIcat1 CCIcat2 CCIcat3;
output out=out_ps_mp6 prob=ps xbeta=logit_ps;
where MP = 6;
run;

proc logistic data = pattern descending;
model  H1N1_cohort = age2009 N_consult12
male region2 region3 region4 region5 region6
region7 region8 region9 region10 region11 region12 region13
diabetes myocardial chf
/*currentalcohol formeralcohol*/
currentsmoking formersmoking
/*underweight overweight unknown*/
antiplatelet anticoagulant antihypert diuretics antidiabetic statin antipsychotic
antidepressant nsaids
flag_ili flag_FLU2008 flag_FLU2009 CCIcat1 CCIcat2 CCIcat3;
output out=out_ps_mp7 prob=ps xbeta=logit_ps;
where MP = 7;
run;

proc logistic data = pattern descending;
model  H1N1_cohort = age2009 N_consult12
male region2 region3 region4 region5 region6
region7 region8 region9 region10 region11 region12 region13
diabetes myocardial chf
/*currentalcohol formeralcohol
currentsmoking formersmoking
underweight overweight unknown*/
antiplatelet anticoagulant antihypert diuretics antidiabetic statin antipsychotic
antidepressant nsaids
flag_ili flag_FLU2008 flag_FLU2009 CCIcat1 CCIcat2 CCIcat3/ridging=none;
output out=out_ps_mp8 prob=ps xbeta=logit_ps;
where MP = 8;
run;

/*Combining the Outputs from Previous Logit Models*/
proc append base=out_ps_mp1 data=out_ps_mp2 force; run;
proc append base=out_ps_mp1 data=out_ps_mp3 force; run;
proc append base=out_ps_mp1 data=out_ps_mp4 force; run;
proc append base=out_ps_mp1 data=out_ps_mp5 force; run;
proc append base=out_ps_mp1 data=out_ps_mp6 force; run;
proc append base=out_ps_mp1 data=out_ps_mp7 force; run;
proc append base=out_ps_mp1 data=out_ps_mp8 force; run;

data combine; set out_ps_mp1; run;

/*Obtaining Variance of PS to Calculate Variance Ratio*/
proc means data=combine var;
class cohort;
var ps;
run;
```

D.3 Forming Propensity Score Matched Sample for Complete Case Method

```
/*Forming Propensity Score Matched Sample with 1:1 Ratio.
  For Missingness Pattern Method, Similar Codes were Used*/
/* 1. Computing standard deviation of the logit of the propensity score*/
proc means data=out_ps2;
var logit_ps;
output out=stddata (keep=std) std=std;
run;

data stddata;
set stddata;
std=0.2*std;
run;

/* 2. Creating macro variable that contains the width of the caliper for matching */
data _null_;
set stddata;
call symput (’stdcal’,std);
run;

/* 3. Matching subjects on the logit of the propensity score*/
proc sort data=out_ps2;
by H1N1_cohort;
run;

data out_ps2;
set out_ps2;
id = _N_;
run;

%MACRO GMATCH(DATA=,GROUP=,ID=,
             MVARS=,WTS=,DMAXK=,DMAX=,DIST=1,
             NCONTLS=1, TIME=,TRANSF=0,
             SEEDCA=,SEEDCO=,PRINT=y,
             OUT=__OUT,OUTNMCA=__NMCA,OUTNMCO=__NMCO);

   %LET BAD=0;

   %IF %LENGTH(&DATA)=0 %THEN %DO;
      %PUT ERROR: NO DATASET SUPPLIED;
      %LET BAD=1;
   %END;

   %IF %LENGTH(&ID)=0 %THEN %DO;
      %PUT ERROR: NO ID VARIABLE SUPPLIED;
      %LET BAD=1;
   %END;

   %IF %LENGTH(&GROUP)=0 %THEN %DO;
      %PUT ERROR: NO CASE(1)/CONTROL(0) GROUP VARIABLE SUPPLIED;
      %LET BAD=1;
   %END;

   %IF %LENGTH(&MVARS)=0 %THEN %DO;
      %PUT ERROR: NO MATCHING VARIABLES SUPPLIED;
      %LET BAD=1;
   %END;

  %IF %LENGTH(&WTS)=0 %THEN %DO;
      %PUT ERROR: NO WEIGHTS SUPPLIED;
      %LET BAD=1;
   %END;

   %LET NVAR=0;
   %DO %UNTIL(%SCAN(&MVARS,&NVAR+1,’ ’)= );
      %LET NVAR=%EVAL(&NVAR+1);
   %END;
   %LET NWTS=0;
   %DO %UNTIL(%QSCAN(&WTS,&NWTS+1,’ ’)= );
      %LET NWTS=%EVAL(&NWTS+1);
   %END;
```

```
 %IF &NVARˆ= &NWTS %THEN %DO;
    %PUT ERROR: #VARS MUST EQUAL #WTS;
    %LET BAD=1;
 %END;

%LET NK=0;
 %IF %QUOTE(&DMAXK)ˆ=  %THEN %DO %UNTIL(%QSCAN(&DMAXK,&NK+1,’ ’)= );
    %LET NK=%EVAL(&NK+1);
 %END;
 %IF &NK>&NVAR %THEN %LET NK=&NVAR;
 %DO I=1 %TO &NVAR;
    %LET V&I=%SCAN(&MVARS,&I,’ ’);
 %END;

%IF &NWTS>0 %THEN %DO;
      DATA _NULL_;
      %DO I=1 %TO &NWTS;
           %LET W&I=%SCAN(&WTS,&I,’ ’);
           IF &&W&I<0 THEN DO;
                PUT ’ERROR: WEIGHTS MUST BE NON-NEGATIVE’;
                CALL SYMPUT(’BAD’,’1’);
           END;
      %END;
      RUN;
 %END;

%IF &NK>0 %THEN %DO;
      DATA _NULL_;
      %DO I=1 %TO &NK;
           %LET K&I=%SCAN(&DMAXK,&I,’ ’);
           IF &&K&I<0 THEN DO;
                PUT ’ERROR: DMAXK VALUES MUST BE NON-NEGATIVE’;
                CALL SYMPUT(’BAD’,’1’);
           END;
      %END;
      RUN;
 %END;

  %MACRO MAX1;
    %IF &DMAXˆ= %THEN %DO;
       & __D<=&DMAX
    %END;
    %DO I=1 %TO &NK;
       & ABS(__CA&I-__CO&I)<=&&K&I
    %END;
  %MEND MAX1;

 %macro greedy;
  %GLOBAL BAD2;

    data __CHECK; set &DATA;
        __id=&id;
        if __id="" then delete;
        %DO I=1 %TO &NVAR;
              IF %scan(&mvars,&i)=. THEN DELETE;
         %END;
         %IF &TIMEˆ= %THEN %DO;
              IF &TIME=. THEN DELETE;
         %END;
     run;

    *** transform data if requested/separate cases & controls;
    %if &transf=1 %then %do;
    proc standard data=__check m=0 s=1 out=_stdzd; var &mvars;
    data _caco;
      set _stdzd;
    %end;

    %if &transf=2 %then %do;
    proc rank data=__check out=_ranks; var &mvars;
    data _caco;
  set _ranks;
%end;

%if &transf=0 %then %do;
data _caco;
  set __check;
%end;


DATA __CASE; SET _caco;
     if &group=1;
DATA __CASE; SET __CASE END=EOF;
 KEEP __IDCA __CA1-__CA&NVAR __R &mvars
   %if &time^= %then %do;
       __catime
   %end;
    ;
   __IDCA=&ID;
   %if &time^= %then %do;
      __catime=&time;
   %end;
   %DO I=1 %TO &NVAR;
      __CA&I=&&V&I;
   %END;
   %if &seedca^= %then %do;
   SEED=&SEEDCA;
   __R=RANUNI( SEED  );
   %end;
   %else %do;
   __R=1;
   %end;

   IF EOF THEN CALL SYMPUT('NCA',_N_);
PROC SORT; BY __R __IDCA;

DATA __CONT; SET _caco;
   if &group=0;
DATA __CONT; SET __CONT END=EOF;
 KEEP __IDCO __CO1-__CO&NVAR __R &mvars
  %if &time^= %then %do;
     __cotime
  %end;
  ;
   __IDCO=&ID;
   %if &time^= %then %do;
      __cotime=&time;
   %end;
   %DO I=1 %TO &NVAR;
      __CO&I=&&V&I;
   %END;
   %if &seedco^= %then %do;
   SEED=&SEEDCo;
   __R=RANUNI( SEED  );
   %end;
   %else %do;
   __R=1;
   %end;

   IF EOF THEN CALL SYMPUT('NCO',_N_);
RUN;
%LET BAD2=0;
%IF &NCO < %EVAL(&NCA*&NCONTLS) %THEN %DO;
   %PUT ERROR: NOT ENOUGH CONTROLS TO MAKE REQUESTED MATCHES;
   %LET BAD2=1;
%END;

%IF &BAD2=0 %THEN %DO;
   PROC SORT; BY __R __IDCO;
   DATA __MATCH;
    KEEP __IDCA __CA1-__CA&NVAR __DIJ __MATCH __CONT_N
    %if &time^= %then %do;
    __catime __cotime
 %end;
 ;
 ARRAY __USED(&NCO) $ 1 _TEMPORARY_;
   DO __I=1 TO &NCO;
      __USED(__I)=’0’;
   END;
   DO __I=1 TO &NCONTLS;
      DO __J=1 TO &NCA;
         SET __CASE POINT=__J;
         __SMALL=.;
         __MATCH=.;
         DO __K=1 TO &NCO;
            IF __USED(__K)=’0’ THEN DO;
               SET __CONT POINT=__K;

              %if &dist=2 %then %do;
               **wtd euclidian dist;
                __D= sqrt(
                %do k=1 %to &nvar;
                %scan(&wts,&k)*(__ca&k - __co&k)**2
                %if &k<&nvar %then + ;
               %end;
                );
              %end;
              %else %do;
               **wtd sum absolute diff;
                __D=
               %do k=1 %to &nvar;
               %scan(&wts,&k)*abs(__ca&k - __co&k )
               %if &k<&nvar %then + ;
               %end;
                 ;
              %end;

               IF __d^=. & (__SMALL=. | __D<__SMALL) %MAX1
               %if &time^= %then %do;
                  & __cotime > __catime
               %end;
               THEN DO;
                  __SMALL=__D;
                  __MATCH=__K;
                  __DIJ=__D;
                  __CONT_N=__I;
               END;
            END;
         END;
         IF __MATCH^=. THEN DO;
            __USED(__MATCH)=’1’;
            OUTPUT;
         END;
      END;
   END;
   STOP;
DATA &OUT;
 SET __MATCH;
 SET __CONT POINT=__MATCH;
 KEEP __IDCA __IDCO __CONT_N __DIJ __CA1-__CA&NVAR
      __CO1-__CO&NVAR __d1-__d&nvar __absd1-__absd&nvar  __WT1-__WT&NVAR
         __catime __cotime __dtime;

 %if &time= %then %do;
     __cotime=.; __catime=.;
 %end;
 LABEL
          __catime="&time/CASE"
          __cotime="&time/CONTROL"
          __dtime="&time/ABS. DIFF"
       __CONT_N=’CONTROL/NUMBER’
       __DIJ=’DISTANCE/D_IJ’
      %DO I=1 %TO &NVAR;
```

```
         __CA&I="&&V&I/CASE"
         __CO&I="&&V&I/CONTROL"
         __absd&I="&&V&I/ABS. DIFF "
         __d&I="&&V&I/DIFF "
         __WT&I="&&V&I/WEIGHT"
       %END;
         ;
      %DO I=1 %TO &NVAR;
         __d&i= (__CA&I-__CO&I);      **raw diff;
         __absd&I=abs(__CA&I-__CO&I); **abs diff;
         __WT&I=&&W&I;
      %END;
         __dtime=__cotime-__catime;

  PROC SORT DATA=&OUT; BY __IDCA __CONT_N;
  proc sort data=__case; by __IDCA;
  data &outnmca; merge __case
       &out(in=__inout where=(__cont_n=1)); by __idca;
       if __inout=0; **non-matches;

  proc sort data=__cont; by __IDCO;
  proc sort data=&out; by __IDCO;
  data &outnmco; merge __cont
       &out(in=__inout); by __idco;
       if __inout=0; **non-matched controls;
  proc sort data=&out; by __IDCA; **re-sort by case id;

%if %upcase(&print)=Y %then %do;
  PROC PRINT data=&out LABEL SPLIT='/';
   VAR __IDCA __IDCO __CONT_N

    __DIJ
   %DO I=1 %TO &NVAR;
    __absd&I
   %END;
   %if &time^= %then %do;
    __dtime
   %end;
   %DO I=1 %TO &NVAR;
    __CA&I __CO&I
   %END;
   %if &time^= %then %do;
    __catime __cotime
   %end;
    ;
   sum __dij;

  title9'Data listing for matched cases and controls';
  footnote"Greedy matching(gmatch) macro: data=&data group=&group id=&id    ";
  footnote2"   mvars=&mvars  wts=&wts dmaxk=&dmaxk dmax=&dmax ncontls=&ncontls";
  footnote3"   transf=&transf dist=&dist time=&time seedca=&seedca  seedco=&seedco";
  footnote4"   out=&out   outnmca=&outnmca  outnmco=&outnmco";
  run;
  title9'Summary data for matched cases and controls--one obs/control';
   %if &sysver ge 8 %then %do;
  proc means data=&out  maxdec=3 fw=8
    n mean median min p10 p25 p75 p90 max sum;
  %end;
  %else %do;
  proc means data=&out maxdec=3
   n mean min max sum;
  %end;
  class __cont_n;
   var __dij

       %do I=1 %TO &NVAR;
           __absd&I
       %end;
       %if &time^= %then %do;
           __dtime
       %end;
```

```
        %do I=1 %TO &NVAR;
            __ca&I
        %end;
        %if &time^= %then %do;
            __catime
        %end;
        %do I=1 %TO &NVAR;
            __co&I
        %end;
        %if &time^= %then %do;
            __cotime
        %end;
           ;
   run;
   *** estimate matching var means within matched sets for controls;
   proc means data=&out  n mean noprint; by __idca;
    var __dij
   %do i=1 %to &nvar;
      __co&i
   %end;
        __cotime
      ;
   output out=_mcont n=n_co mean=__dijm
   %do i=1 %to &nvar;
     __com&i
   %end;
       __tcom
     ;
   data _onecase; set &out; by __idca; if first.__idca;
   data __camcon; merge _onecase _mcont; by __idca;

   keep __idca n_co __dijm
       __dtime __catime  __tcom
    %do i=1 %to &nvar;
     __ca&i __com&i  __actd&i __absd&i
    %end;
   ;


   %do i=1 %to &nvar;
   __absd&i=abs(__ca&i - __com&i);
   __actd&i=(__ca&i - __com&i);
  %end;
   __dtime=__tcom-__catime
    ;

 label
  n_co="No./CONTROLS"
  __dijm="Average/Dij"
  __dtime="&time/Mean Time DIFF"
  __tcom="&time/Mean CONT TIME"

 %do i=1 %to &nvar; %let vvar=%scan(&mvars,&i);
   __absd&i="&vvar/Mean ABS. DIFF"
   __com&i="&vvar/Mean CONTROL"
 %end;
   ;
title9’Summary data for matched cases and controls--one obs/case(using average control value)’;
%if &sysver ge 8 %then %do;
proc means data=__camcon maxdec=3 fw=8
  n mean median min p10 p25 p75 p90 max sum;
%end;
%else %do;
proc means data=__camcon maxdec=3
  n mean min max sum;
%end;
var n_co __dijm
%do i=1 %to &nvar;
 __absd&i
%end;
%if &time^= %then %do;
```

```
       __dtime
      %end;
      %do i=1 %to &nvar;
      __ca&i
      %end;
      %if &time^= %then %do;
       __catime
      %end;
      %do i=1 %to &nvar;
      __com&i
      %end;
      %if &time^= %then %do;
      __tcom
      %end;
          ;
    %end; **end of print=y loop**;
   %END;  **end of bad2=0 loop**;
   run;
   title9; footnote;
   run;

   %mend greedy;

   %IF &BAD=0 %THEN %DO;
        %GREEDY
   %END;
%MEND GMATCH;

%gmatch(
data=out_ps2,
group=H1N1_cohort,
id=id,
mvars=logit_ps,
wts=1,
dist=1,
dmaxk=&stdcal,
ncontls=1,
seedca=83322,
seedco=14582,
out=matchpairs,
print=F
);

data matchpairs;
set matchpairs;
pair_id= _N_;
run;

/* 4. Creating a Dataset Containing the Matched Unexposed Subjects */
data control_match;
set matchpairs;
control_id = __IDCO;
logit_ps = __CO1;
keep pair_id control_id logit_ps;
run;

/* 5. Creating a Dataset Containing the Matched Exposed Subjects */
data case_match;
set matchpairs;
case_id = __IDCA;
logit_ps = __CA1;
keep pair_id case_id logit_ps;
run;

proc sort data=control_match;
by control_id;
run;

proc sort data=case_match;
by case_id;
run;
```

```
data exposed;
set out_ps2;
if H1N1_cohort=1;
case_id=id;
run;

data unexposed;
set out_ps2;
if H1N1_cohort=0;
control_id=id;
run;

proc sort data=exposed;
by case_id;
run;

proc sort data=unexposed;
by control_id;
run;

data unexposed_match;
merge control_match (in=f1) unexposed (in=f2);
by control_id;
if f1 and f2;
run;

data exposed_match;
merge case_match (in=f1) exposed (in=f2);
by case_id;
if f1 and f2;
run;

/*PS-Matched Sets are Created */
data long2;
set unexposed_match exposed_match;
prop_score=exp(logit_ps)/(1+exp(logit_ps));
run;
```

D.4 Absolute Standardized Differences for Matched Sample

```
/*Standardized difference for continuous variables for matched sample*/
proc sort data=long2;
by H1N1_cohort;
run;

%macro cont (var=,label=);
proc means data=long2 mean stddev data=cprd noprint;
var &var;
by H1N1_cohort;
output out=outmean (keep=H1N1_cohort mean stddev) mean=mean stddev=stddev;
run;

data H1N1_cohort0;
set outmean;
if H1N1_cohort=0;
mean_0=mean;
s_0=stddev;
keep mean_0 s_0;
run;

data H1N1_cohort1;
set outmean;
if H1N1_cohort=1;
mean_1=mean;
s_1=stddev;
keep mean_1 s_1;
run;

data newdata;
length label $ 35;
merge H1N1_cohort0 H1N1_cohort1;
d=(mean_1 - mean_0)/sqrt((s_1*s_1 + s_0*s_0)/2);
d=round(abs(d),0.001);
label=&label;
keep d label;
run;

proc append data=newdata base=standiff force;
run;
%mend cont;

/*Standardized difference for categorical variables for matched sample*/
%macro binary (var=,label=);
proc means data=long2 noprint;
var &var;
by H1N1_cohort;
output out=outmean (keep=H1N1_cohort mean) mean=mean;
run;

data H1N1_cohort0;
set outmean;
if H1N1_cohort=0;
mean_0=mean;
keep mean_0;
run;

data H1N1_cohort1;
set outmean;
if H1N1_cohort=1;
mean_1=mean;
keep mean_1;
run;

data newdata;
length label $ 35;
merge H1N1_cohort0 H1N1_cohort1;
d=(mean_1 - mean_0)/sqrt((mean_1*(1-mean_1) + mean_0*(1-mean_0))/2);
d=round(abs(d),0.001);
label=&label;
keep d label;
```

```
run;

proc append data=newdata base=standiff force;
run;

%mend binary;
%cont(var=age2009,label="Age");
%cont(var=N_consult12,label="Number of Consultation");

%binary(var=age1,label="64-69 Years");
%binary(var=age2,label="70-74 Years");
%binary(var=age3,label="75-79 Years");
%binary(var=age4,label="80-84 Years");
%binary(var=age5,label="85 Years or Older");
%binary(var=male,label="Male gender");
%binary(var=region1,label="North East");
%binary(var=region2,label="North West");
%binary(var=region3,label="Yorkshire and The Humber");
%binary(var=region4,label="East Midlands");
%binary(var=region5,label="West Midlands");
%binary(var=region6,label="East of England");
%binary(var=region7,label="Sout West");
%binary(var=region8,label="South Central");
%binary(var=region9,label="London");
%binary(var=region10,label="Sout East Coast");
%binary(var=region11,label="Northern Ireland");
%binary(var=region12,label="Scotland");
%binary(var=region13,label="Wales");
%binary(var=diabetes,label="Diabetes");
%binary(var=myocardial,label="Myocardial infarction");
%binary(var=chf,label="Congestive Heart Failure");
%binary(var=noalcohol,label="Not drinker");
%binary(var=currentalcohol,label="Current drinker");
%binary(var=formeralcohol,label="Former drinker");
%binary(var=nosmoking,label="Not smoker");
%binary(var=currentsmoking,label="Current smoker");
%binary(var=formersmoking,label="Former smoker");
%binary(var=normal,label="Normal");
%binary(var=underweight,label="Underweight");
%binary(var=overweight,label="Overweight");
%binary(var=unknown,label="Unknown");
%binary(var=antiplatelet,label="Antiplatelet Used");
%binary(var=anticoagulant,label="Anticoagulant Used");
%binary(var=antihypert,label="Antihypertensive Used");
%binary(var=diuretics,label="Diuretics Used");
%binary(var=antidiabetic,label="Antidiabetic Used");
%binary(var=statin,label="Statin Used");
%binary(var=antipsychotic,label="Antipsychotic Used");
%binary(var=antidepressant,label="Antidepressant Used");
%binary(var=nsaids,label="NSAIDS Used");
%binary(var=flag_ili,label="ILI Events");
%binary(var=flag_FLU2008,label="Seasonal 2008");
%binary(var=flag_FLU2009,label="Seasonal 2009");
%binary(var=CCIcat0,label="CCI = 0");
%binary(var=CCIcat1,label="CCI = 1");
%binary(var=CCIcat2,label="CCI = 2");
%binary(var=CCIcat3,label="CCI = 3 or more");

proc print data=standiff;
run;
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Development of a propensity score to handle confounding of the association between H1N1 (2009) pandemic influenza vaccination and the risk of stroke in elderly

Richting: **Master of Statistics-Biostatistics**
Jaar: **2015**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Wahyudi, Kurnia

Datum: **23/01/2015**