Pneumococcal vaccination prevented severe LRTIs in adults: a causal inference framework applied in registry data

Peer-reviewed author version

Mamouris, Pavlos; Henrard, Severine; MOLENBERGHS, Geert; Verhaegen, Jan; Lin , Guohao & Vaes , Bert (2022) Pneumococcal vaccination prevented severe LRTIs in adults: a causal inference framework applied in registry data. In: JOURNAL OF CLINICAL EPIDEMIOLOGY, 143 , p. 118 -127.

DOI: 10.1016/j.jclinepi.2021.12.008
Handle: http://hdl.handle.net/1942/36883

The effect of pneumococcal vaccination on preventing lower respiratory tract infections in adults: A causal inference methodological framework based on registry data

Pavlos Mamouris[1*]; Severine Henrard[2,3]; Geert Molenberghs[4,5]; Jan Verhaegen[6,7]; Guohao Lin[1]; Bert Vaes[1]

*Correspondence

Pavlos Mamouris

Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium

Kapucijnenvoer 33, J building, 3000 Leuven, Belgium

E-mail: pavlos.mamouris@kuleuven.be

Affiliations

[1] Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium

[2] Clinical Pharmacy Research Group, Louvain Drug Research Institute (LDRI), Université catholique de Louvain, Brussels, Belgium.

[3] Institute of Health and Society (IRSS), Université catholique de Louvain, Brussels, Belgium

[4] I-BioStat, KU Leuven University of Leuven, Leuven, Belgium

[5] I-BioStat, Hasselt University, Diepenbeek, Belgium

[6] National Reference Centre for Streptococcus pneumoniae, University Hospitals Leuven, Leuven, Belgium

[7] Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium.

Abstract

Objectives: We estimated the effect of pneumococcal vaccination (PV) on acute lower respiratory tract infections (LRTIs) in various age and risk groups using different methods within a causal inference methodological framework.

Study Design and Setting: We used data from a general practitioners' morbidity registry for the year 2019. Both traditional statistical methods (regression-based and propensity score methods) and machine learning techniques were deployed. Multiple imputation was used to account for missing data. Relative risks (RRs) with 95% confidence intervals were estimated. Sensitivity analyses were performed to account for the severity of LRTIs and differences in vaccination registration.

Results: All methods showed a standardized mean difference below 0.1 for each covariate. No method was found to be superior to another. PV (combination of conjugate and polysaccharide vaccine) had an overall protective effect for severe LRTIs. PV was protective in different age and risk groups, especially in people aged 50-84 years with an intermediate risk group.

Conclusion: Using several techniques, PV was found to prevent severe LRTIs and confirmed the recommendations of the Belgian Superior Health Council.

Keywords:

Pneumococcal vaccine; Relative risk; Causal inference; Propensity score; Registry data; Machine learning

Running title:

Pneumococcal vaccination prevents severe LRTIs in adults: A unified causal inference framework

What is new?

- Registry data was used to estimate the effect of pneumococcal vaccination to prevent lower respiratory tract infections
- Relative risks were calculated using different methods (regression-based methods, propensity score methods and machine learning techniques) to build confidence in the robustness of the conclusions
- A combination of a conjugate and polysaccharide vaccine was found to prevent severe lower respiratory tract infections in the global adult population and in different age and risk groups
- These results confirmed the recent recommendations of the Belgian Superior Health Council

1. Introduction

Acute lower respiratory tract infections (LRTIs) are a major cause of morbidity and mortality [1]. Almost 2.38 million deaths worldwide resulted from LRTIs in 2016 [2], and LRTIs are the fourth leading cause of global disability-adjusted life-years [3]. Streptococcus pneumoniae was found to be responsible for at least 5% of the severe LRTIs in the adult population in primary care [4]. To prevent pneumococcal diseases in adults, two types of vaccines are available: the 13-valent pneumococcal conjugate vaccine (PCV13) and the 23-valent pneumococcal polysaccharide vaccine (PPV23).

In Belgium, the Superior Health Council recommends vaccinating adults aged from 16-85 years old with a high risk of pneumococcal infection, adults aged from 50-85 years old who have certain comorbidities and healthy people aged 65-85 years with PCV13 followed by PPV23 [5]. However, the efficacy of these vaccines in the prevention of LRTIs in adults remains the subject of debate [6-8].

Randomized controlled trials (RCTs) are the gold standard approach for assessing the effect of treatments or interventions. Randomization ensures that the treatment effect can be directly estimated [9]. However, occasionally, a trial might be unethical, time-consuming and infeasible [10]. In addition, when the outcome is a rare disease or the aim is to investigate effects on patients with polymedication and multimorbidity, observational data are the only alternative [11]. In observational studies, treatment selection is often influenced by subject characteristics, which often differ systematically between treated and untreated subjects [12]. Therefore, methodologies and strategies that consider those systematic differences should be deployed. These methods vary and include general statistical methods, such as regression-based methods (RBMs) and propensity score (PS) methods. Recently, machine learning (ML) techniques, namely, Bayesian additive regression trees and generalized boosted modelling, have been increasingly used; they automatically detect the best model for balancing the covariates and for capturing nonlinearities, polynomial terms and interactions [13].

Therefore, there were two objectives of this study. First, a causal inference methodological framework is developed for data from registries. Both traditional statistical methods and ML techniques are used, and the differences in terms of which method achieves the best balance is examined. Second, the association between pneumococcal vaccination (PV) and LRTIs in different age and risk groups was investigated.

2. Materials and methods

2.1 Data source

We used the Intego registry, a Flemish general practice morbidity registry, which was described in more detail elsewhere [14]. Briefly, in 2019, Intego comprised approximately 285,000 people from 104 general practice centres, accounting for 4.3% of the Flemish

population. Medication and vaccines were classified according to the WHO's Anatomical Therapeutic Chemical classification system, and diagnoses were linked to the International Classification of Primary Care and International Statistical Classification of Diseases and Related Health Problems 10th Revision.

2.2 Study design and study population

2.2.1 Study population

For the current study, we included only practices that coded more than 80% of their registered diagnoses in 2019 (n = 86). The study was performed on the population aged 16 and older in 2019.

2.2.2 Intervention

PV consists of PPV23 (ATC code = J07AL01) and PCV13 (ATC code = J07AL02). We considered two strategies: (i) a patient was administered either vaccine, and (ii) a patient was administered both vaccines. We further categorized the treatment according to the years since the last vaccination of each type (0-5 years, 6-10 years, and ≥ 10 years), starting from the LRTIs date in 2019 or 31 December 2019. Last, to identify the PV in Intego, we used two sources: PV registrations and PV prescriptions. Nevertheless, when a vaccine was only prescribed, we cannot be certain that the vaccination occurred. Therefore, to make our conclusions robust, we investigated the effect of the intervention on (i) vaccination registration only and (ii) vaccination registration and vaccination prescription together.

2.2.3 Outcome definition

The outcome of interest was LRTIs in 2019. We made a distinction between LRTIs with or without antibiotics since the prescription of antibiotics might indicate a more severe LRTIs. Specifically, we considered (i) LRTIs without a prescription of antibiotics and (ii) LRTIs with a prescription of antibiotics 1 month before or after the LRTIs.

2.2.4 Main analysis and sensitivity analyses

In total, we performed eight analyses (one main and seven sensitivity analyses) on the effect of PV to prevent LRTIs. In addition, we estimated the effect of PV in different age groups (16-49, 16-84, 50-64, 50-84, 65-74, 74-84, 65-84, 85-plus, 50-plus, 65-plus, and 75-plus) and risk groups (low, intermediate and high risk).

2.3 Covariate selection

We selected appropriate confounding factors using previous research evidence from the literature, expert advice and the recommendations of the Superior Health Council in Belgium [15, 16]. The covariates used for adjustment were the baseline characteristics, the risk group for LRTIs infection, lab tests and comorbidities as detailed in Supplementary Tables S1 and S2.

2.4 Treatment effect estimators

The estimands of interest differ according to the research question at hand and the target population to be compared. The two most common estimands are the following: the average treatment effect, which is the effect on the entire population; and the average effect of the treatment on the treated, which is the effect for those in the treatment group [17]. The average treatment effect is of more interest if every treatment potentially might be offered to every subject, namely, if the entire population was moved from the control to the treated group [18]. The average effect of the treatment on the treated is preferable when patients' characteristics are more likely to determine the treatment received [19]. In our study, PV was more likely to be administered to patients belonging to specific age groups and with specific comorbidities. Therefore, for our research question, the average effect of the treatment on the treated estimand is of interest.

2.5 Modelling framework

2.5.1 Regression-based methods

Multivariable logistic regression is used to compute odds ratios. However, in our study we aim to compute relative risks (RRs), thus a log-binomial model will be deployed. The difference between the multivariable logistic regression and log-binomial models is the link function: in the multivariable logistic regression, the logit function is used; and in the log-binomial model, the log function is used [20].

2.5.2 Propensity score methods

The PS of a subject is defined as the probability of treatment assignment T conditional on a vector of observed baseline covariates X [12],

$$e(X) = Pr\,(T = 1 \mid X)$$

In this way, all the baseline covariates X are summarized into one single variable. In RCTs, when the outcome is binary, the PS is approximately 0.5 since we expect a balance of covariates between the intervention and control groups. In observational studies, due to the imbalance of the covariates, the PS differs between subjects and therefore needs to be estimated. The most popular method to estimate the PS is the logistic regression, where the outcome is the intervention conditional on all covariates. Once the PS is calculated, we can use PS-based methods including matching, stratification, inverse probability of treatment weighting and overlap weighting to balance our covariates [21, 22]. We performed nearest neighbour matching with a calliper of 0.2 combined with exact matching for sex, age group, risk, socioeconomic status and smoking status.

2.5.3 Machine learning methods

ML techniques offer an alternative approach when calculating PSs [23, 24]. It differs from the logistic regression in terms of automatically incorporating quadratic, polynomial, or interaction terms and does not require any parametric assumptions [25, 26]. In this work, we deployed the Bayesian additive regression trees and generalized boosted modelling for binary outcomes [27, 28].

2.6 PSs after multiple imputation

An additional difficulty arises when data are incomplete. Multiple imputation is a methodology to "fill in" the missing data multiple times with plausible values that reflect the uncertainty in predicting the true unobserved values [29]. The values are typically drawn from the conditional distribution of a subject's missing measurements given the observed ones. We performed longitudinal imputation for the missing covariates since the previous and earlier observations of the same patient can be considered. The partially missing variables were smoking status, body mass index, estimated glomerular filtration rate, systolic blood pressure and diastolic blood pressure. We drew 20 imputations, which is prudent since the percentage of missing values was substantial. In the context of PSs, there are two strategies for estimating the effects after multiple imputation. The first strategy is the within approach [30], where the effects are calculated within each dataset and then the results are pooled together. The second strategy is the across approach [31], where the PSs are averaged across imputed datasets, and the effects are calculated using this average PS. We used the within approach since it was demonstrated to have superior statistical performance [32].

2.7 Balance diagnostics

The standardized difference, which is a comparison of the means of continuous covariates and the distribution of their categorical counterparts divided by the pooled standard

deviation between treated and untreated subjects, was used to investigate the covariate balance between the intervention groups [33]. This metric lies between 0 and 1, and the typical threshold is 0.1 [34, 35]. In addition, the bias reduction is given as:

$$bias\ reduction\ = \left(1 - \frac{|d_{after}|}{|d_{before}|} * 100\right),$$

where d_{before} and d_{after} denote the standardized difference before and after PS matching, respectively. The bias reduction provides an alternative and intuitive way to investigate how the bias is reduced by using this method.

2.8 Statistical analysis

We deployed eight different methods to calculate the RRs and the 95% Confidence Intervals. Further details for these methods can be found in Supplementary Methods. We used robust standard errors to account for potential sources of uncertainty when using weighting techniques. For our analysis, we used the R software [36] and different packages as described in Supplementary R packages.

3. **Results**

3.1 Baseline covariate balance

Fig. 1 displays the flowchart of the study population as described in the materials section.

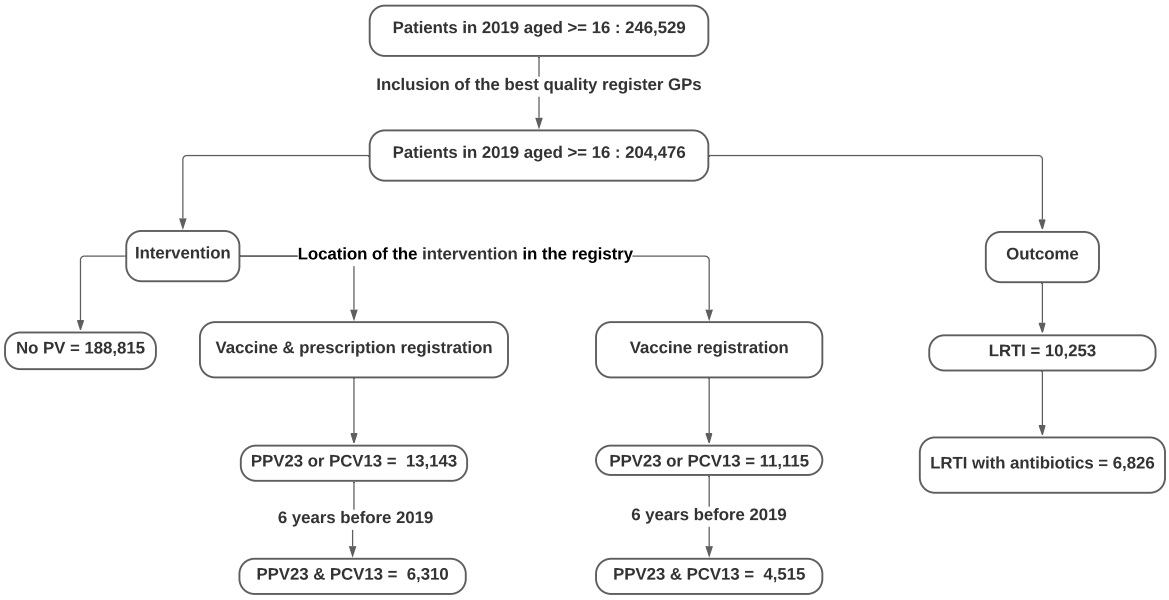

Fig. 1: Flowchart of the study population in 2019

Table 1 reports the absolute standardized difference before and after PS matching when the treatment was PPV23 & PCV13 located in the vaccine and prescription data source, and the outcome was LRTIs with antibiotics. We observe that bias was present in our data since the standardized difference of the distance measure in the original sample was 1.38, and most of the variables had standardized differences largely above the 0.10 threshold. This was further supported by the bias reduction metric, which reached 99% for several variables.

Table 1: Baseline covariates before and after 1:1 PS matching for PCV13 & PPV23 (0-5 years)

Variables	Original sample			PS matching 1:1			
	Not vaccinated (N = 188815)	Vaccinated (N = 6310)	SMD/ SPD*	Not vaccinated (N = 6022)	Vaccinated (N = 6022)	SMD/ SPD	Bias reduction
Propensity score (distance measure)			1.3755			0.0007	99.95
Sex, male (%)	86893 (46.0)	3029 (48.0)	0.040	2875 (47.7)	2875 (47.7)	<0.001	99.75
Age, mean (SD)	46.66 (18.99)	69.41 (12.26)	1.424	69.47 (12.61)	69.34 (12.26)	0.011	99.23
Socioeconomic status, no compensation (%)	160856 (85.2)	5125 (81.2)	0.106	4922 (81.7)	4922 (81.7)	<0.001	99.91
Risk status			1.071			<0.001	99.99
High risk (%)	4336 (2.3)	920 (14.6)		735 (12.2)	735 (12.2)		
Low risk (%)	139375 (73.8)	1725 (27.3)		1723 (28.6)	1723 (28.6)		
Intermediate risk, yes (%)	45104 (23.9)	3665 (58.1)		3564 (59.2)	3564 (59.2)		
Smoking status			0.317			<0.001	99.97
Ex-smoker (%)	50625 (26.8)	2585 (41.0)		2455 (40.8)	2455 (40.8)		
Smoker (%)	48780 (25.8)	1111 (17.6)		1023 (17.0)	1023 (17.0)		
Never-smoker (%)	89410 (47.4)	2614 (41.4)		2544 (42.2)	2544 (42.2)		
Body mass index, obese (%)	32792 (17.4)	1317 (20.9)	0.089	1323 (22.0)	1259 (20.9)	0.026	70.79
Systolic blood pressure, mean (SD)	124.65 (14.88)	128.84 (14.54)	0.285	129.55 (14.88)	129.02 (14.43)	0.037	87.02
Diastolic blood pressure, mean (SD)	76.69 (9.19)	75.49 (8.49)	0.135	76.10 (8.66)	75.58 (8.47)	0.060	55.56
Liver disease, yes (%)	4049 (2.1)	386 (6.1)	0.201	348 (5.8)	370 (6.1)	0.015	92.54
Heart failure, yes (%)	1771 (0.9)	298 (4.7)	0.230	244 (4.1)	276 (4.6)	0.026	88.7
Atrial fibrillation, yes (%)	4349 (2.3)	711 (11.3)	0.362	606 (10.1)	678 (11.3)	0.039	89.23
Heart valve, yes (%)	2206 (1.2)	360 (5.7)	0.251	298 (4.9)	334 (5.5)	0.027	89.24
Atherosclerosis, yes (%)	2502 (1.3)	327 (5.2)	0.219	296 (4.9)	304 (5.0)	0.006	97.26
Chronic obstructive pulmonary disease, yes (%)	3976 (2.1)	1012 (16.0)	0.500	830 (13.8)	820 (13.6)	0.005	99
Asthma, yes (%)	16310 (8.6)	1226 (19.4)	0.315	1035 (17.2)	1093 (18.2)	0.025	92.06
Diabetes, yes (%)	10674 (5.7)	1155 (18.3)	0.397	1148 (19.1)	1100 (18.3)	0.020	94.96
Hypertension, yes (%)	29216 (15.5)	2746 (43.5)	0.646	2612 (43.4)	2596 (43.1)	0.005	99.23
Ischemic disease, yes (%)	5347 (2.8)	746 (11.8)	0.350	672 (11.2)	708 (11.8)	0.019	94.57
Stroke, yes (%)	4152 (2.2)	500 (7.9)	0.263	415 (6.9)	469 (7.8)	0.034	87.07
Cancer, yes (%)	30740 (16.3)	2402 (38.1)	0.505	2216 (36.8)	2223 (36.9)	0.002	99.6
Estimated glomerular filtration rate category			0.370			0.021	94.32
Stage 1 (%)	46271 (24.5)	845 (13.4)		829 (13.8)	812 (13.5)		
Stage 2 (%)	98308 (52.1)	3118 (49.4)		3039 (50.5)	3005 (49.9)		
Stage 3 (%)	40743 (21.6)	2094 (33.2)		1929 (32.0)	1963 (32.6)		
Stage 4 (%)	2948 (1.6)	201 (3.2)		176 (2.9)	191 (3.2)		
Stage 5 (%)	545 (0.3)	52 (0.8)		49 (0.8)	51 (0.8)		
Flu vaccine in 2018, yes (%)	33034 (17.5)	5203 (82.5)	1.709	4923 (81.8)	4932 (81.9)	0.004	99.77

*For continuous variables standardised mean differences (SMD) are used, whereas for categorical covariates standardised proportion differences (SPD)

Before matching, we observed that the vaccinated group was older with higher percentages belonging to the intermediate- and high-risk groups and a large difference in influenza vaccination in 2018 (82.5% versus 17.5%). After matching, 6,022 patients remained in each intervention group. The standardized differences dropped substantially and were less than the threshold of 0.1. In addition to 1:1 PS matching, we investigated the covariate balance on all methods, and the results are shown in Supplementary Tables S3, S4, and S5.

Fig. 2 presents an intuitive and straightforward comparison of all statistical methods in terms of covariate balance utilizing the absolute standardized difference. We observe that the overlap method, followed by the ML techniques, produced the best balance. Nevertheless, for all methods except the unmatched (crude regression), each covariate was well below 0.1, demonstrating that all methods adequately balanced the data.

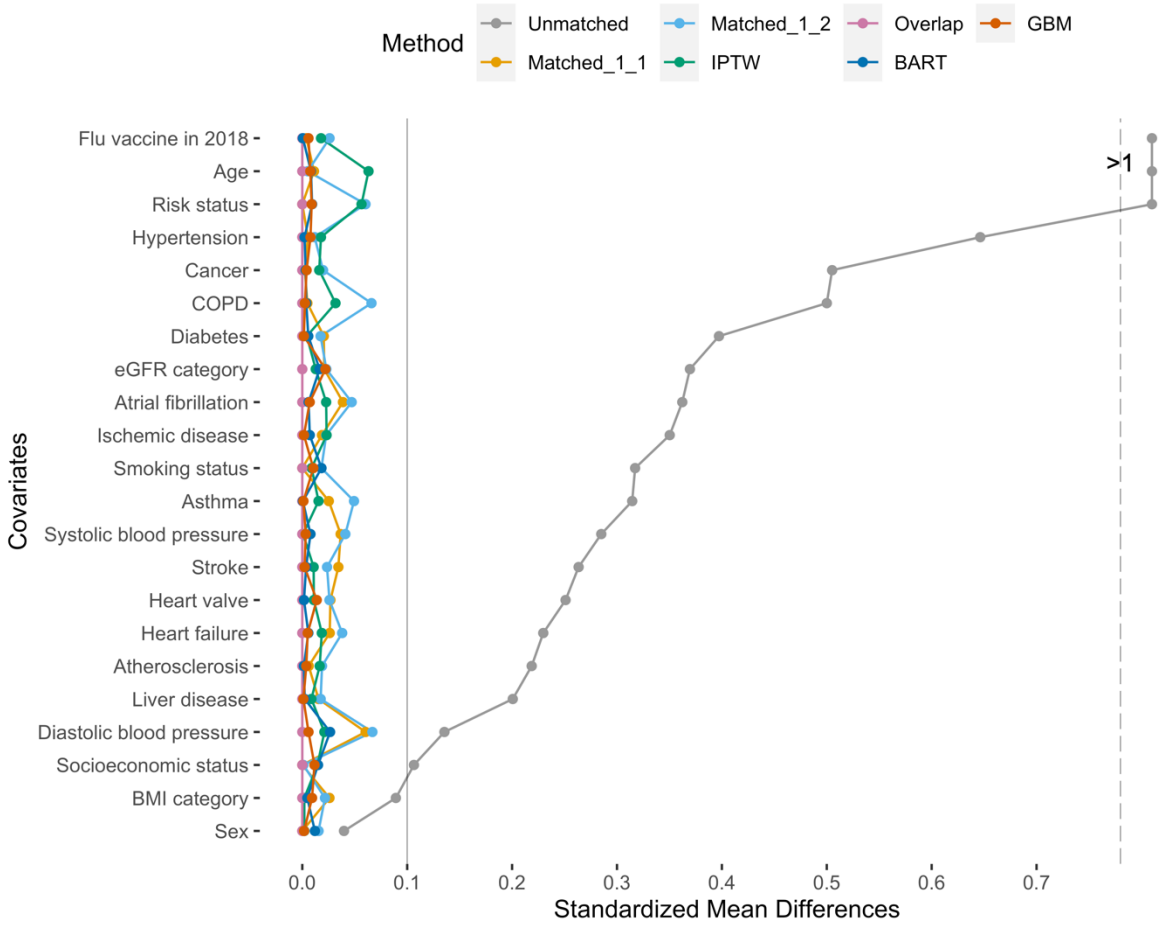

Fig. 2: Comparison of the different methods for subjects who received PCV13 and PPV23 based on the standardized mean difference

3.2 Effect of pneumococcal vaccination on LRTIs

Fig. 3 indicates a protective effect of PPV23 & PCV13 vaccination for the prevention of LRTIs with antibiotics using all methods. However, the effect of PPV23 or PCV13 was not significant using the large majority of models.

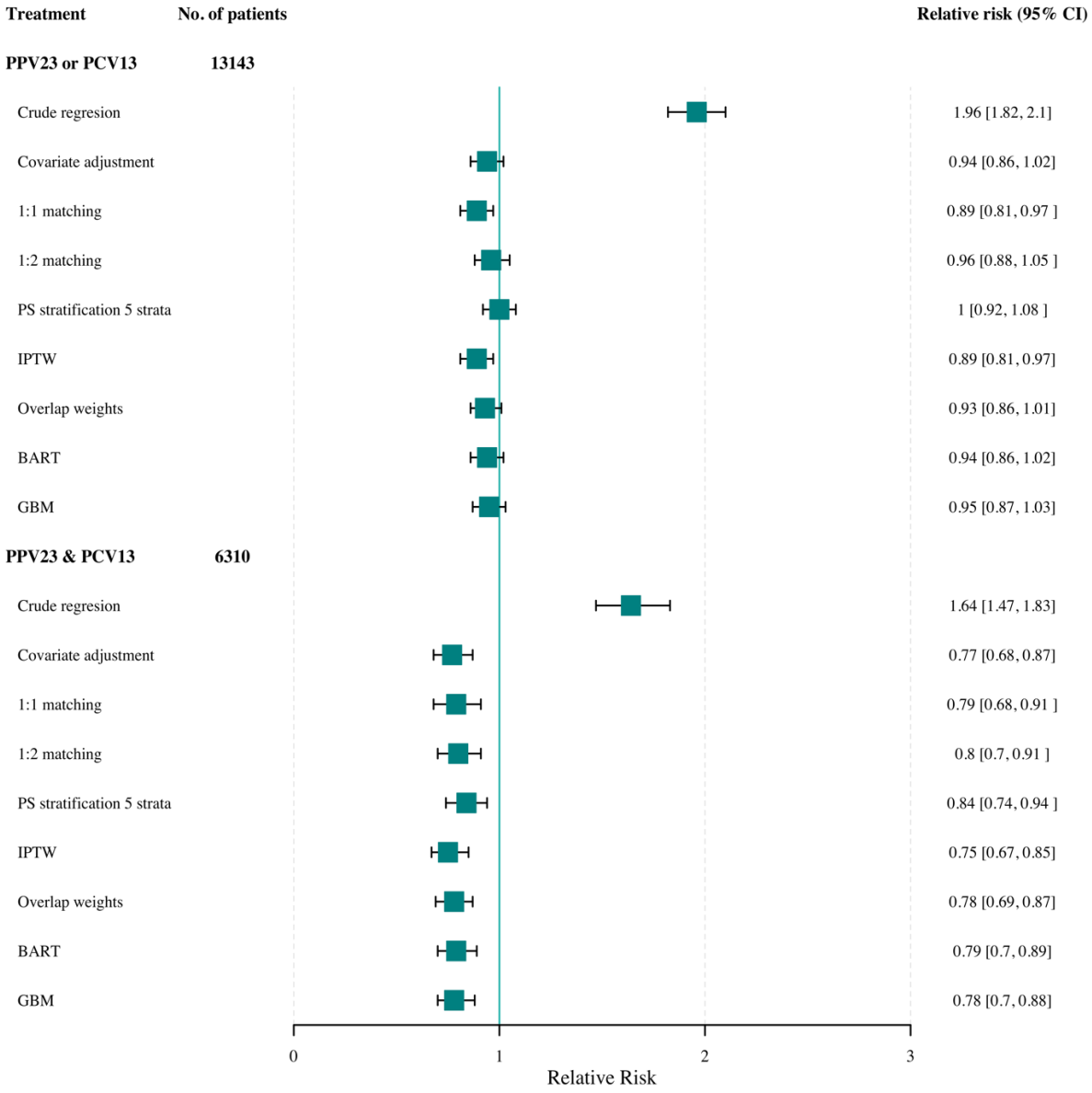

Fig. 3: Forest plot of the RR of PPV23 or PCV13 and PPV23 & PCV13 on LRTIs treated with antibiotics (vaccine and prescription registration) according to each statistical method used.

Furthermore, we investigated the effect of PPV23 & PCV13 vaccination in different age groups for high-, intermediate-, and low-risk statuses. Fig. 4 displays the RR for all patients aged from 65-84 and further stratified by risk status. We observed that the treatment was protective for the entire age group and different risk categories. However, there were few patients in the high-risk group; thus, uncertainty remained, as expressed by the large CI. A protective trend was observed in high-risk people aged 16-84 years as depicted in Supplementary Fig. S1, although the CI was not significant. Supplementary Fig. S2 shows that

in the intermediate-risk people aged 50-84 years, a significant protective effect was found. Furthermore, a protective effect was seen in the 50-plus, 65-plus and 65-74 age categories as described in Supplementary Table S6.

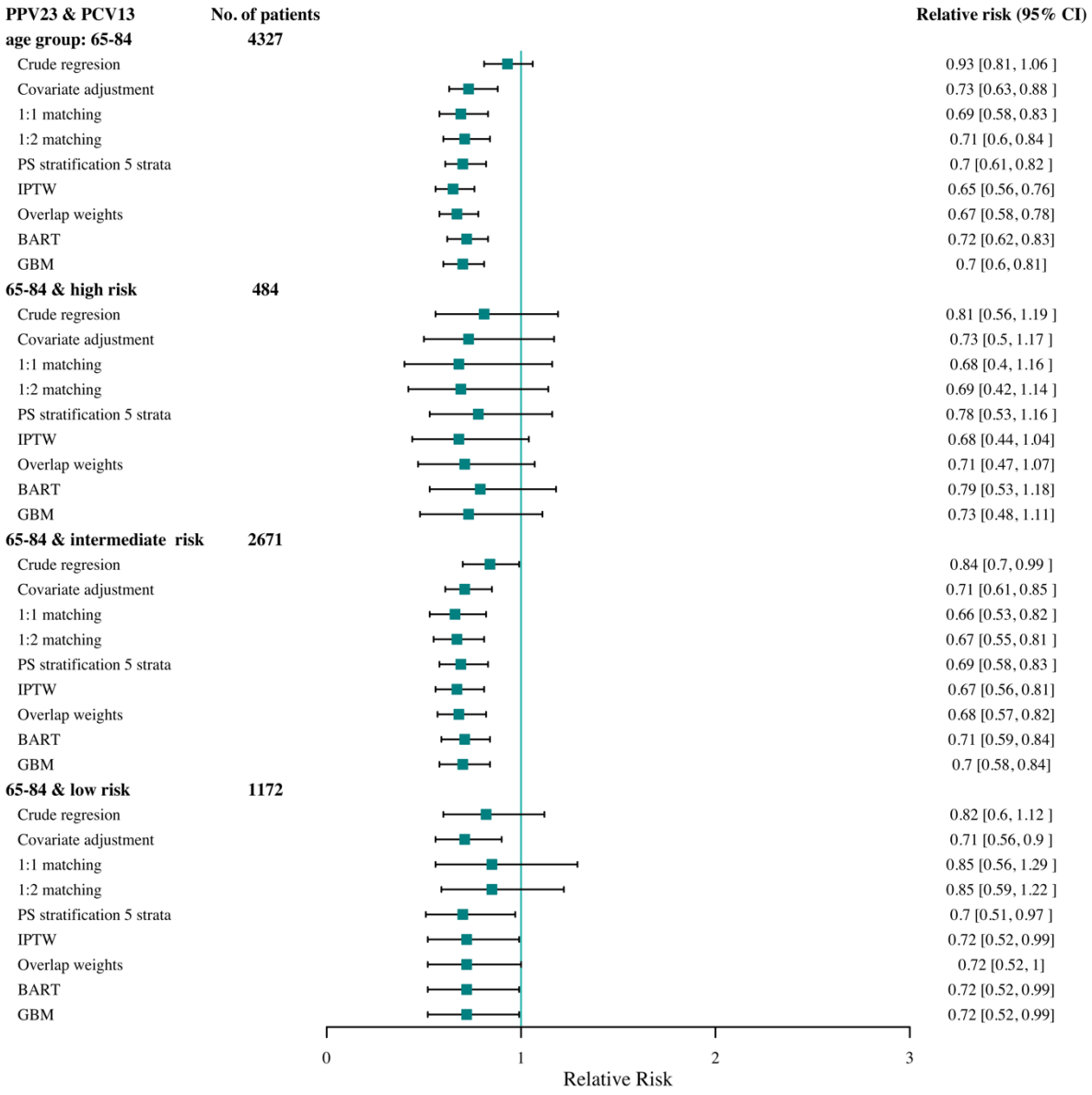

Fig. 4: Forest plot of the RR of PPV23 & PCV13 on LRTIs treated with antibiotics (vaccine and prescription registration) in the 65-84 age group stratified by risk status

3.3 Sensitivity analyses

Starting from PPV23 or PCV13 (registered or prescribed), we observed a harmful trend for LRTIs without antibiotics and a nonsignificant effect for LRTIs with antibiotics. The same

299 effect was observed when we only used the registered vaccinations as depicted in
300 Supplementary Fig. S3. For PPV23 & PCV13 (registered or prescribed), we observed a
301 nonsignificant effect for LRTIs without antibiotics. However, the effect was protective for
302 LRTIs with antibiotics (primary analysis). Supplementary Fig. S4 demonstrates that the same
303 trend was observed when we only used registered vaccinations. Furthermore, a protective effect
304 of vaccination was observed in the 16-84, 50-84, 65-plus, 65-74 and 65-84 age groups for all
305 analyses (primary and sensitivity), as described in Supplementary Table S7. Only for sensitivity
306 analyses 6 and 8, i.e., PPV23 or PCV13 when the outcome was LRTIs without antibiotics, did
307 we observe a nonsignificant effect.

308 ## Discussion
309

310 In this large registry-based study, a causal inference methodological framework was
311 used to estimate the effect of PV to prevent LRTIs in adults. Several methods, including RBMs,
312 PS and ML were utilized to balance the intervention and control groups and estimate an
313 unbiased effect. The overlap method produced the best balance; however, all methods were
314 below the threshold of 0.1. Therefore, no method was found to be superior to the others, which
315 underscores the robustness of the results. Vaccinating adults with PPV23 or PCV13 did not
316 have a protective effect against LRTIs. However, a combination of PPV23 and PCV13 was
317 found to prevent severe LRTIs in the global adult population and in different age and risk
318 groups, confirming the recommendations of the Belgian Superior Health Council.

319 In earlier literature, controversy arose over the preferred or most suitable methodology
320 to balance the intervention and control groups. When differences are large between intervention
321 and control subjects and the true relationship between the covariate and outcome is even
322 moderately nonlinear, RBMs can increase the bias in the treatment effect [37–39]. However,
323 RBMs and PS were compared; and in 43 observational studies, both methods yielded similar

results [40]. Additionally, in several cardiovascular studies, PS methods were not superior to RBMs and were worse in some scenarios [41]. Nevertheless, PS methods are superior to RBMs when modelling rare events [42]. Furthermore, ML techniques are of increasing interest since they automatically detect the best model for balancing the covariates and capture nonlinearities, polynomial terms and interactions. Our conclusion is that no method is superior to another. As Stuart stated, matching techniques should not conflict with RBMs but should be considered to be complementary [25]. However, we would further extend this statement by suggesting that ML techniques should be an extra tool in the methodological framework, because deploying several methods serves as a thorough and informative sensitivity analysis that highlights the robustness of the results. In future research, we suggest (i) carefully choosing the estimand of interest, (ii) utilizing an array of methodologies to build confidence in the robustness of the conclusions, (iii) incorporating missing data to include all covariates, (iv) balancing diagnostics to help determine which method might be preferable, and (v) performing sensitivity analyses with EHR data when registrations might be incomplete.

PPV23 & PCV13 showed a protective effect against severe LRTIs in the overall adult population and in specific age and risk groups, which confirms the recommendations of the Belgian Superior Health Council [5]. No benefit was found for people aged 85-plus. In addition, our conclusions are similar to those of a literature review when both vaccinations were administered [43]. However, we did not investigate the sequentiality of the different vaccines. This will be a topic for further research.

Importantly, the proportion of pneumococcal infection and circulating types of S. pneumoniae among people with LRTIs can differ from year to year, although the change in capsular types is a slow process [44]. This means that the results of our study might change depending on the year used in the analyses. However, using registry data has the advantage that

the analyses can easily be repeated each year in order to continuously monitor the effect of vaccination.

The current study has several strengths. First, having a large sample size allowed more controls to be available; and especially in 1:1 PS matching, we lost very few treated patients. Notably, PS matching differs from weighting since it discards many units, thus in settings where few controls are available, weighting techniques might be preferable. Second, MI was performed for missing covariates, which allowed many covariates to be incorporated in the models. Third, by using several models and performing multiple sensitivity analyses, we were able to show the robustness of our results. Finally, our study is the first to calculate the effect of PV vaccination not only in different age groups but also stratified by risk categories, which targets patients more in need of PV vaccination.

Some limitations of working with registry data should be noted. First, since data on hospitalization and severity of the LRTIs are missing in Intego, we used antibiotic treatment as a 'proxy' for more severe LRTIs. In total, 67% of LRTIs episodes in 2019 were treated with antibiotics. However, in Belgium, the proportion of LRTIs treated with antibiotics is high compared to that in other countries [45], [46]. In this respect, our results should be interpreted with caution. Second, not all vaccinations might be registered. Therefore, we used registered and prescribed vaccinations as the intervention. Our reasoning for including the vaccination prescriptions was that 82% of people with a prescription also had a vaccination registered, indicating that this population is more prone to get vaccinated. Furthermore, sensitivity analyses were able to show the robustness of our results. Third, misclassification of the outcome might be present. Finally, the run-time when deploying ML techniques was significantly high. With big data, many covariates and 20 imputations, the run-time was approximately 20 hours (30 min for Bayesian additive regression trees and 30 min for generalized boosted modelling using

20 imputed datasets) for a single analysis. Since we conducted 7 additional sensitivity analyses, the run time increased to 160 hours.

Conclusion

In this large registry-based study, several methods were utilized to balance the intervention and control groups and estimate an unbiased effect of PV for LRTIs. The overlap method followed by ML techniques produced the best balance. However, all methods sufficiently balanced the covariates, which enhanced the robustness of the results. A combination of PPV23 and PCV13 was found to prevent severe LRTIs in the global adult population and in different age and risk groups, confirming the recommendations of the Belgian Superior Health Council. These findings may assist clinicians in making more informed decisions in vaccinating patients with PV to prevent severe LRTIs. Epidemiologists, statisticians, and biomedical researchers can utilize the unified methodological framework for estimating unbiased effects and derive robust conclusions.

Author contributions: **Pavlos M.**: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – Review & Editing **Severine H.**: Conceptualization, Methodology **Geert M.:** Conceptualization, Methodology **Jan V.:** Conceptualization, Methodology **Guohao L.:** Conceptualization, Software, Methodology **Bert V.:** Conceptualization, Project administration, Supervision. All authors read, reviewed, edited, and approved the manuscript.

Funding: This research received no external funding.

Conflict of interest: The author reports no conflicts of interest in this work.

References

[1] C. Troeger *et al.*, "Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Infect. Dis.*, vol. 18, no. 11, pp. 1191–1210, Nov. 2018, doi: 10.1016/S1473-3099(18)30310-4.

[2] M. Naghavi *et al.*, "Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016," *The Lancet*, vol. 390, no. 10100, pp. 1151–1210, Sep. 2017, doi: 10.1016/S0140-6736(17)32152-9.

[3] T. Vos *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, Oct. 2020, doi: 10.1016/S0140-6736(20)30925-9.

[4] J. Flamaing, W. De Backer, Y. Van Laethem, S. Heijmans, and A. Mignon, "Pneumococcal lower respiratory tract infections in adults: an observational case-control study in primary care in Belgium," *BMC Fam. Pract.*, vol. 16, May 2015, doi: 10.1186/s12875-015-0282-1.

[5] "Gezondheidsraad H. Vaccinatie tegen pneumokokken;" 2014. [Online]. Available: https://www.zorg-en-gezondheid.be/sites/default/files/atoms/files/hgr_9562_vaccinatie_tegen_pneumokokken_vweb%20%281%29.pdf

[6] G. G. Pitsiou and I. P. Kioumis, "Pneumococcal vaccination in adults: Does it really work?," *Respir. Med.*, vol. 105, no. 12, pp. 1776–1783, Dec. 2011, doi: 10.1016/j.rmed.2011.07.008.

[7] R. J. José and J. S. Brown, "Adult pneumococcal vaccination: advances, impact, and unmet needs," *Curr. Opin. Pulm. Med.*, vol. 23, no. 3, pp. 225–230, May 2017, doi: 10.1097/MCP.0000000000000369.

[8] B. A. Winje *et al.*, "Efficacy and effectiveness of pneumococcal vaccination in elderly–an update of the literature," 2019.

[9] S. Greenland, J. Pearl, and J. M. Robins, "Causal diagrams for epidemiologic research," *Epidemiology*, pp. 37–48, 1999.

[10] M. A. Hernán and J. M. Robins, "Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available," *Am. J. Epidemiol.*, vol. 183, no. 8, pp. 758–764, Apr. 2016, doi: 10.1093/aje/kwv254.

[11] L. P. Garrison, P. J. Neumann, P. Erickson, D. Marshall, and C. D. Mullins, "Using Real-World Data for Coverage and Payment Decisions: The ISPOR Real-World Data Task Force Report," *Value Health*, vol. 10, no. 5, pp. 326–335, 2007, doi: https://doi.org/10.1111/j.1524-4733.2007.00186.x.

[12] P. C. Austin, "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivar. Behav. Res.*, vol. 46, no. 3, pp. 399–424, May 2011, doi: 10.1080/00273171.2011.568786.

[13] D. Westreich, J. Lessler, and M. J. Funk, "Propensity score estimation: machine learning and classification methods as alternatives to logistic regression," *J. Clin. Epidemiol.*, vol. 63, no. 8, pp. 826–833, Aug. 2010, doi: 10.1016/j.jclinepi.2009.11.020.

[14] P. Mamouris, V. Nassiri, G. Molenberghs, M. van den Akker, J. van der Meer, and B. Vaes, "Fast and optimal algorithm for case-control matching using registry data: application on the antibiotics use of colorectal cancer patients," *BMC Med. Res. Methodol.*, vol. 21, no. 1, p. 62, Apr. 2021, doi: 10.1186/s12874-021-01256-3.

[15] A. Torres, W. E. Peetermans, G. Viegi, and F. Blasi, “Risk factors for community-acquired pneumonia in adults in Europe: a literature review,” *Thorax*, vol. 68, no. 11, pp. 1057–1065, Nov. 2013, doi: 10.1136/thoraxjnl-2013-204282.
[16] A. Blommaert, “Use of pneumococcal vaccines in the elderly: an economic evaluation,” p. 164, 2016.
[17] G. W. Imbens, “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *Rev. Econ. Stat.*, vol. 86, no. 1, pp. 4–29, Feb. 2004, doi: 10.1162/003465304323023651.
[18] P. C. Austin and E. A. Stuart, “Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies,” *Stat. Med.*, vol. 34, no. 28, pp. 3661–3679, 2015, doi: https://doi.org/10.1002/sim.6607.
[19] U. Benedetto, S. J. Head, G. D. Angelini, and E. H. Blackstone, “Statistical primer: propensity score matching and its alternatives,” *Eur. J. Cardiothorac. Surg.*, vol. 53, no. 6, pp. 1112–1117, 2018.
[20] L.-A. McNutt, C. Wu, X. Xue, and J. P. Hafner, “Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes,” *Am. J. Epidemiol.*, vol. 157, no. 10, pp. 940–943, May 2003, doi: 10.1093/aje/kwg074.
[21] E. A. Stuart, “Matching methods for causal inference: A review and a look forward,” *Stat. Sci. Rev. J. Inst. Math. Stat.*, vol. 25, no. 1, pp. 1–21, Feb. 2010, doi: 10.1214/09-STS313.
[22] F. Li, L. E. Thomas, and F. Li, “Addressing Extreme Propensity Scores via the Overlap Weights,” *Am. J. Epidemiol.*, vol. 188, no. 1, pp. 250–257, Jan. 2019, doi: 10.1093/aje/kwy201.
[23] B. Griffin, D. McCaffrey, D. Almirall, C. Setodji, and L. Burgette, “Chasing balance and other recommendations for improving nonparametric propensity score models,” *J. Causal Inference*, vol. 5, no. 2, 2017, doi: 10.1515/jci-2015-0026.
[24] J. L. Hill, “Bayesian Nonparametric Modeling for Causal Inference,” *J. Comput. Graph. Stat.*, vol. 20, no. 1, pp. 217–240, Jan. 2011, doi: 10.1198/jcgs.2010.08162.
[25] B. K. Lee, J. Lessler, and E. A. Stuart, “Improving propensity score weighting using machine learning,” *Stat. Med.*, vol. 29, no. 3, pp. 337–346, 2010, doi: https://doi.org/10.1002/sim.3782.
[26] R. Ferri-García and M. del M. Rueda, “Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys,” *PloS One*, vol. 15, no. 4, p. e0231500, 2020.
[27] H. A. Chipman, E. I. George, and R. E. McCulloch, “BART: Bayesian additive regression trees,” *Ann. Appl. Stat.*, vol. 4, no. 1, pp. 266–298, Mar. 2010, doi: 10.1214/09-AOAS285.
[28] C. M. Setodji, D. F. McCaffrey, L. F. Burgette, D. Almirall, and B. Ann Griffin, “The right tool for the job: choosing between covariate balancing and generalized boosted model propensity scores,” *Epidemiol. Camb. Mass*, vol. 28, no. 6, pp. 802–811, Nov. 2017, doi: 10.1097/EDE.0000000000000734.
[29] J. A. C. Sterne *et al.*, “Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls,” *BMJ*, vol. 338, p. b2393, Jun. 2009, doi: 10.1136/bmj.b2393.
[30] C. Leyrat *et al.*, “Propensity score analysis with partially observed covariates: how should multiple imputation be used?,” *Stat. Methods Med. Res.*, vol. 28, no. 1, pp. 3–19, 2019.
[31] R. Mitra and J. P. Reiter, “A comparison of two methods of estimating propensity scores after multiple imputation,” *Stat. Methods Med. Res.*, vol. 25, no. 1, pp. 188–204, 2016.

[32] B. B. L. Penning de Vries and R. H. H. Groenwold, *Comments on propensity score matching following multiple imputation*. SAGE Publications Sage UK: London, England, 2016.
[33] P. C. Austin, "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples," *Stat. Med.*, vol. 28, no. 25, pp. 3083–3107, 2009, doi: https://doi.org/10.1002/sim.3697.
[34] P. C. Austin, P. Grootendorst, and G. M. Anderson, "A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study," *Stat. Med.*, vol. 26, no. 4, pp. 734–753, 2007, doi: https://doi.org/10.1002/sim.2580.
[35] S.-L. T. Normand *et al.*, "Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores," *J. Clin. Epidemiol.*, vol. 54, no. 4, pp. 387–398, Apr. 2001, doi: 10.1016/S0895-4356(00)00321-8.
[36] R. C. Team, "R: A language and environment for statistical computing," 2013.
[37] "Matching As An Econometric Evaluation Estimator | The Review of Economic Studies | Oxford Academic." https://academic.oup.com/restud/article-abstract/65/2/261/1580756 (accessed Jun. 03, 2021).
[38] D. B. Rubin and N. Thomas, "Combining propensity score matching with additional adjustments for prognostic covariates," *J. Am. Stat. Assoc.*, vol. 95, no. 450, pp. 573–585, 2000.
[39] D. B. Rubin, "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation," *Health Serv. Outcomes Res. Methodol.*, vol. 2, no. 3, pp. 169–188, Dec. 2001, doi: 10.1023/A:1020363010465.
[40] B. R. Shah, A. Laupacis, J. E. Hux, and P. C. Austin, "Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review," *J. Clin. Epidemiol.*, vol. 58, no. 6, pp. 550–559, Jun. 2005, doi: 10.1016/j.jclinepi.2004.10.016.
[41] M. C. Elze *et al.*, "Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies," *J. Am. Coll. Cardiol.*, vol. 69, no. 3, pp. 345–357, 2017.
[42] M. S. Cepeda, R. Boston, J. T. Farrar, and B. L. Strom, "Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders," *Am. J. Epidemiol.*, vol. 158, no. 3, pp. 280–287, Aug. 2003, doi: 10.1093/aje/kwg115.
[43] A. W. Cripps, T. Folaranmi, K. D. Johnson, L. Musey, M. S. Niederman, and U. K. Buchwald, "Immunogenicity following revaccination or sequential vaccination with 23-valent pneumococcal polysaccharide vaccine (PPSV23) in older adults and those at increased risk of pneumococcal disease: a review of the literature," *Expert Rev. Vaccines*, vol. 20, no. 3, pp. 257–267, Mar. 2021, doi: 10.1080/14760584.2021.1889374.
[44] "Streptococcus pneumoniae 2019.pdf." Accessed: Jul. 28, 2021. [Online]. Available: https://nrchm.wiv-isp.be/nl/ref_centra_labo/streptococcus_pneumoniae_invasive/Rapporten/Streptococcus%20pneumoniae%202019.pdf
[45] D. Tell, S. Engström, and S. Mölstad, "Adherence to guidelines on antibiotic treatment for respiratory tract infections in various categories of physicians: a retrospective cross-sectional study of data from electronic patient records," *BMJ Open*, vol. 5, no. 7, p. e008096, Jul. 2015, doi: 10.1136/bmjopen-2015-008096.

[46] C. C. Butler *et al.*, “Variation in antibiotic prescribing and its impact on recovery in patients with acute cough in primary care: prospective study in 13 countries,” *BMJ*, vol. 338, p. b2242, Jun. 2009, doi: 10.1136/bmj.b2242.