Pragmatic and Observational Research

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

An evaluation of exact matching and propensity score methods as applied in a comparative effectiveness study of inhaled corticosteroids in asthma

This article was published in the following Dove Press journal:
Pragmatic and Observational Research
22 March 2017
Number of times this article has been viewed

Anne Burden[1]
Nicolas Roche[2]
Cristiana Miglio[1]
Elizabeth V Hillyer[1]
Dirkje S Postma[3]
Ron MC Herings[4]
Jetty A Overbeek[4]
Javaria Mona Khalid[5]
Daniela van Eickels[6]
David B Price[1,7]

[1]Observational and Pragmatic Research Institute Pte Ltd, Singapore; [2]University Paris Descartes (EA2511), Cochin Hospital Group (AP-HP), Paris, France; [3]Department of Pulmonology, University Medical Center Groningen, University of Groningen, Groningen, [4]PHARMO Institute for Drug Outcomes Research, Utrech, the Netherlands; [5]Takeda Development Centre Europe Ltd, London, UK; [6]Takeda Pharmaceuticals International GmbH, Zurich, Switzerland; [7]Academic Primary Care, University of Aberdeen, Aberdeen, UK

Background: Cohort matching and regression modeling are used in observational studies to control for confounding factors when estimating treatment effects. Our objective was to evaluate exact matching and propensity score methods by applying them in a 1-year pre–post historical database study to investigate asthma-related outcomes by treatment.

Methods: We drew on longitudinal medical record data in the PHARMO database for asthma patients prescribed the treatments to be compared (ciclesonide and fine-particle inhaled corticosteroid [ICS]). Propensity score methods that we evaluated were propensity score matching (PSM) using two different algorithms, the inverse probability of treatment weighting (IPTW), covariate adjustment using the propensity score, and propensity score stratification. We defined balance, using standardized differences, as differences of <10% between cohorts.

Results: Of 4064 eligible patients, 1382 (34%) were prescribed ciclesonide and 2682 (66%) fine-particle ICS. The IPTW and propensity score-based methods retained more patients (96%–100%) than exact matching (90%); exact matching selected less severe patients. Standardized differences were >10% for four variables in the exact-matched dataset and <10% for both PSM algorithms and the weighted pseudo-dataset used in the IPTW method. With all methods, ciclesonide was associated with better 1-year asthma-related outcomes, at one-third the prescribed dose, than fine-particle ICS; results varied slightly by method, but direction and statistical significance remained the same.

Conclusion: We found that each method has its particular strengths, and we recommend at least two methods be applied for each matched cohort study to evaluate the robustness of the findings. Balance diagnostics should be applied with all methods to check the balance of confounders between treatment cohorts. If exact matching is used, the calculation of a propensity score could be useful to identify variables that require balancing, thereby informing the choice of matching criteria together with clinical considerations.

Keywords: asthma, exact matching, propensity score, observational

Background

Observational studies provide important information about the effectiveness and safety of therapies in real-life clinical settings. Indeed, many have argued that the results of observational studies are an essential complement to the findings of randomized controlled trials.[1–5] A fundamental limitation of observational studies, however, is that treatment assignment is not random. Therefore, demographic and clinical patient characteristics that influence doctors' prescribing choices or that affect treatment outcomes may systematically differ between patient cohorts being compared, resulting in a biased estimation of treatment effects.

Correspondence: David B Price
Academic Primary Care, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen AB25 2ZD, UK
Tel +44 12 2455 4588
Fax +44 12 2455 0683
Email dprice@opri.sg

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/POR.S122563

Cohort matching and regression modeling are methods used to reduce biases and confounding factors to enable comparison between treatment options in observational studies. Two commonly used matching methods are exact matching and propensity score matching (PSM).[6–8] Exact matching has the advantage of ensuring that patients are paired on key variables of interest; however, increasing the number of matching variables to improve the precision of matching increases the chance of excluding patients who do not match, reducing study sample size and variability of the patient population.[7] In addition, patients excluded due to unavailability of some variables may represent a specific population, which would induce a selection bias and limit the representativeness of the sample. With PSM, patients are matched on a single propensity score representing the probability of receiving the exposure of interest given the observed baseline characteristics. This method can be especially useful when treatment cohorts are dissimilar and the number of potential confounding factors is large; however, an important drawback of PSM is the risk of matching dissimilar patients who have similar scores but important differences in key variables of interest, especially those that may interact with treatment effectiveness.

With both exact matching and PSM, patients who do not match are excluded from analysis, which has implications for the power of the subsequent comparisons and on representativeness of the matched cohorts with regard to the true population. Other methods of causal analysis retain the full dataset (so no biases are introduced through patient selection) but use the propensity score in other ways to achieve balance between treatment cohorts (ie, not just for matching patients). These include the inverse probability of treatment weighting (IPTW), covariate adjustment using the propensity score, and propensity score stratification.[6,8–10]

Previous research has evaluated the performance of various matching methods. Austin's research[11] compares the balance obtained in matched cohorts when using several types of propensity score methods, in both real and simulated data sets. He does not include exact matching, however, and does not investigate the impact of the method choice on the primary endpoint of interest. Studies by Wells et al[12] and Fullerton et al[13] compare matching methods in real data sets, including exact matching, and discuss the strengths and weaknesses of each. Despite these useful studies, it is important to gain more evidence in this area, to strengthen the conclusions that are drawn from the analyses of one data set and to allow investigators to make more informed decisions on the design of observational studies.

Our current study contributes to the evidence base by comparing matching methods in a real-life data set in which we previously investigated asthma-related outcomes in two treatment groups. We compared extrafine-particle inhaled corticosteroid (ICS) to larger fine-particle ICS – a comparison that has been investigated previously.[14] We found that extrafine-particle ICS was associated with similar or better asthma-related outcomes than a larger fine-particle ICS at significantly lower prescribed doses.[15] In this study, we aim to compare the performance of exact matching with that of PSM by applying these methods in this historical cohort study, including both balance diagnostics and the impact on the primary endpoint of the original study. In addition, we examine the performance of the propensity score-based causal analysis techniques (IPTW, covariate adjustment, and stratification).

Methods

We compared analytic methods by applying them to a real-life observational study previously reported elsewhere.[15]

Data source and study design

The previous study used anonymized pharmacy dispensing and hospital discharge data drawn from the Dutch PHARMO Database Network (September 2005 through December 2012).[16] These data were used to identify patients with asthma prescribed extrafine-particle ICS (Alvesco [ciclesonide]) or one of two fine-particle ICS (Flixotide [fluticasone] and non-extrafine-particle beclomethasone). The aim of the study was to investigate the role of particle size in the long-term effectiveness of ICS therapy. A 1-year pre–post historical cohort analysis of asthma-related outcomes was conducted, for patients 12–60 years of age prescribed their first ICS therapy as either extrafine-particle or fine-particle ICS. Additional criteria required that patients had received two or more prescriptions for asthma at any time in addition to the first ICS prescription: at least one of these prescriptions had to be for ICS during the outcome period (1-year period following first ICS prescription), but there had to be no ICS prescribed in the baseline period (1-year period preceding first ICS prescription). Patients were excluded if they had evidence of any other chronic respiratory disease or if they were prescribed long-acting muscarinic antagonists or maintenance oral corticosteroids during the baseline period.

The three coprimary endpoints evaluated over 1 outcome year were the severe exacerbation rate and the dichotomous variables of risk-domain asthma control and overall asthma control. We defined severe exacerbations as an asthma-related hospitalization or acute course of oral corticosteroids.[17] Risk-domain asthma control was defined as the absence of severe exacerbations, and overall asthma control was defined as achieving risk-domain asthma control in addition to receiving a prescribed mean daily dose of salbutamol ≤200 µg/day. Change in therapy was the secondary endpoint.

We conducted the study according to recommended standards for observational research, including an a priori research plan, study registration, an independent steering committee, and commitment to publish.[5,18,19]

Methods of matching and causal analysis

To compare outcomes between the two treatment cohorts, we evaluated exact matching and four approaches using the propensity score, namely, PSM, IPTW, covariate adjustment using the propensity score, and propensity score stratification.[6,10]

Exact matching with statistical adjustment for residual confounders

Exact matching with statistical adjustment for residual confounders (exact matching) has been described in previous publications from our research team.[7,20–23] In brief, we first compiled a list of potential matching criteria informed by expert clinical advice and previous research experience, including those predictive of outcomes and the key baseline clinical characteristics differing between unmatched cohorts, identified using χ^2 and Mann–Whitney *U* tests, as appropriate. Our matching criteria for this study were sex, age, baseline risk-domain asthma control (controlled/not controlled), baseline long-acting β-agonist (LABA) prescription (yes/no), baseline short-acting β2-agonist (SABA) daily dose, baseline leukotriene receptor antagonist prescription (yes/no), baseline prescription of antifungals to treat oral candidiasis (yes/no), and year of ICS therapy initiation.

Matching criteria were then applied sequentially to produce two matched cohorts containing all possible pairings; bespoke software was used to randomly select final matched pairs by eliminating double matches. Endpoints were compared via conditional regression models and adjusted for any residual noncollinear baseline confounders and for those demographic and baseline variables predictive of the outcome through full multivariable analysis.

Propensity score matching

By definition, the propensity score ranges from 0 to 1 and is the probability of treatment assignment (in our study, the probability of being prescribed ciclesonide), conditional on baseline characteristics.[6] For PSM, patients are matched on one variable, namely, the estimated propensity score or logit of the propensity score within a predefined caliper, usually employing a 1:1 matching ratio although other ratios can be considered, as appropriate to the size and characteristics of the available sample.

The list of covariates included in the propensity score should include all potential confounders. We selected appropriate confounding factors from predictors of outcomes identified using multivariable analysis, previous research evidence, and differences in demographic and key baseline clinical characteristics. The propensity score was estimated using a logistic regression model whereby the treatment was the dependent variable and the identified covariates were the independent variables. The model was stepwise reduced to construct a more parsimonious final model to avoid overfitting, which has the potential to inflate variability in the model estimates and to increase bias in the presence of unmeasured confounders.[9,24]

We used two different algorithms to match patients in the two cohorts in a 1:1 ratio using the propensity score. The first algorithm, developed by our research team at Research in Real-Life (RiRL; RiRL algorithm), matched patients on the logit of the propensity score, initially considering all possible matches within 0.1 times the pooled standard deviation of the logit and then randomly selecting unique matched pairings. The second algorithm, developed by Parsons,[25] was the so-called greedy algorithm, which ordered patients in the ciclesonide cohort and sequentially matched them on the propensity score to the nearest unmatched patient in the fine-particle ICS cohort. If >1 unmatched patients in the fine-particle ICS cohort were a match, then the matching patient was selected at random. Matches were made sequentially with a decreasing level of accuracy (initially matching exactly on the propensity score to 5 decimal places reducing to 1 decimal place).

After matching on the propensity score, we checked balance of the matched cohorts via standardized differences to compare mean values and prevalences, respecifying the propensity score model until balance was achieved.[26] When a satisfactory propensity score was identified based on the balance assessment of the matched cohorts using the two matching methods, the score was used to carry out the remaining methods.

The inverse probability of treatment weighting

For the IPTW, propensity scores are used directly as inverse weights to estimate average treatment effect (ATE).[7,10] This method weights individual patients based on the inverse of the probability of their treatment allocation, conditional on baseline characteristics, to create a pseudo-dataset in which the distribution of potentially confounding variables is balanced between the treatment and control groups.[8,27] We used stabilized weights, which multiply the IPTW by the unconditional probability of treatment allocation in order to stabilize the variance estimates so that treatment effects and their variance can be estimated directly using conventional regression methods. Using stabilized weights also preserves the original sample size when creating the pseudo-dataset.

Covariate adjustment using the propensity score

Covariate adjustment using the propensity score applies the propensity score as a covariate in the regression models to adjust the treatment effect.[6] Models include the treatment cohort and the estimated propensity score as explanatory variables, with the estimated propensity score treated as a continuous variable. Endpoints are then compared across unmatched cohorts. In addition, we adjusted for residual confounders to evaluate any potential residual influence of baseline predictors.

Propensity score stratification

Rosenbaum and Rubin[28] showed that creating 5 propensity score subclasses removes at least 90% of the bias in the estimated treatment effect of the covariates included in the propensity score. Stratification involves the creation of a predefined number of strata and then estimation of the comparative effects of exposures in the two cohorts within each stratum. The stratum-specific estimates of the effects, weighted by the proportion of patients within the stratum, are then pooled together to obtain the overall treatment effect by using the mean of each estimate across the strata. As noted by Austin,[6] stratification on the propensity score can be conceptualized as a meta-analysis of a set of quasi-randomized controlled trials, the latter being the strata.

To apply this method to our dataset, we stratified the unmatched treatment cohorts into quintiles by propensity score before outcome evaluation.

More details about the above methods can be seen in the "Additional methods" section of the Supplementary materials.

Comparison between exact matching and propensity score methods

We compared the performance of exact matching and propensity score methods by evaluating the following three criteria: 1) the balance obtained using standardized differences to compare mean values and prevalences of baseline variables (for exact matching, PSM, and IPTW); 2) modeled outcome results (for all methods); and 3) the number of patients lost during matching (with exact matching and PSM).

Standardized differences were calculated using a macro written in SAS statistical software, developed by Yang and Dalton and available via the website of the Lerner Research Institute.[29] Using standardized differences, we considered balance as being achieved for differences lying within a 10% window, which has been used in the literature as the definition of a negligible difference.[6]

Conventional regression models were used to estimate and compare the outcomes between the unmatched treatment cohorts and those constructed using IPTW methods. Conditional regression models were used to estimate and compare the outcomes between matched cohorts. Results included rate ratios (RRs) and odds ratios (ORs) with 95% confidence intervals (CIs), first calculated using only the matching or propensity score method and then additionally adjusted for any residual confounders.

For the exact-matched analyses, the results for the dichotomous outcomes of risk-domain and overall asthma control differed slightly from previously published results because this study used PROC LOGISTIC (rather than PROC GENMOD with a binomial distribution and logit link) for these analyses. This was because PROC GENMOD cannot be used for a stratified analysis (by propensity score) and so, for consistency and to allow comparison, PROC LOGISTIC was used throughout.

Analyses were conducted with SAS v9.3 (SAS Institute, Marlow, Buckinghamshire, UK) and SPSS v22 (IBM Corporation, Armonk, NY, USA). Statistical significance was set at $P<0.05$ and trends at $P<0.10$.

Ethical approval

The study was approved by the PHARMO compliance and governance board – the independent Compliance Committee STIZON/PHARMO Institute. This committee is approved by the Dutch Data Protection Authority to control the provision of PHARMO data for scientific research. Due to the anonymization of the data, formal patient consent was not required, upon approval of the research question and the methods planned to analyze the data.

Results

Sample sizes and power

Of 4064 eligible patients identified in the database during the study period, 1382 (34%) were prescribed extrafine-particle ciclesonide and 2682 (66%) fine-particle ICS. Hence, this was the size of the original, unmatched data set, to which the matching methods were applied (creating data subsets). The mean (standard deviation) age was 43 (13) years in the ciclesonide cohort and 38 (15) in the fine-particle ICS cohort; 36% of patients in each cohort were male (Table 1).

Of the 1382 unmatched patients initiating ICS therapy as ciclesonide, 1244 (90%) were retained using exact matching, and 1321 and 1323 (both 96%) were retained using PSM (RiRL and greedy algorithms, respectively). According to a posteriori power calculations, these sample sizes all provided adequate power: using the unmatched proportions achieving risk-domain asthma control (0.897 and 0.850; OR 1.537) and a two-cohort χ^2 test with two-sided significance with α of 0.05, unmatched analyses were powered at 99%; the exact matching comparison was powered at 93%; and the PSM comparisons were both powered at 94% to detect a difference between cohorts in risk-domain asthma control.

Cohort matching and representation of the full population

A list of 12 covariates to use for the propensity score estimation was identified after excluding seven collinear variables and three variables not contributing to the final model (Table 2). Baseline daily SABA dose and evidence of gastroesophageal reflux disease (GERD) both strongly influenced the propensity score (Table S1 for correlation coefficients).

Table 1 Baseline demographic and clinical characteristics of patients

Patient characteristics	Unmatched		Exact matching		Propensity score matching				Stabilized IPTW pseudo-dataset	
					RiRL algorithm		Greedy algorithm			
	Ciclesonide (n=1382)	FP ICS (n=2682)	Ciclesonide (n=1244)	FP ICS (n=1244)	Ciclesonide (n=1321)	FP ICS (n=1321)	Ciclesonide (n=1323)	FP ICS (n=1323)	Ciclesonide (n=1380)	FP ICS (n=2683)
Sex, male	492 (36)	969 (36)	436 (35)	436 (35)	470 (36)	493 (37)	478 (36)	461 (35)	487 (35)	961 (36)
Age, mean (SD)	43 (13)	38 (15)[a]	43 (13)	43 (13)[b]	42 (13)	43 (13)	43 (13)	43 (13)	40 (14)	39 (14)
Comorbidity[c]										
Rhinitis	612 (44)	1021 (38)[d]	539 (43)	469 (38)[b]	567 (43)	568 (43)	569 (43)	560 (42)	554 (40)	1076 (40)
Eczema	427 (31)	744 (28)[d]	381 (31)	358 (29)	407 (31)	386 (29)	412 (31)	400 (30)	406 (29)	785 (29)
GERD	572 (41)	771 (29)[d]	504 (41)	420 (34)[b]	529 (40)	521 (39)	535 (40)	493 (37)[b]	463 (34)	889 (33)
Thrush	20 (1.4)	21 (0.8)[d]	2 (0.2)	2 (0.2)	16 (1.2)	14 (1.1)	16 (1.2)	15 (1.1)	13 (1.0)	26 (1.0)
Acetaminophen script[a]	24 (1.7)	65 (2.4)	23 (1.8)	33 (2.7)	23 (1.7)	23 (1.7)	23 (1.7)	22 (1.7)	33 (2.4)	58 (2.2)
Year of ICS initiation, median (IQR)	2009 (2007–2010)	2008[a] (2007–2009)	2009 (2007–2010)	2009 (2007–2010)	2009 (2007–2009)	2009 (2008–2010)	2009 (2007–2009)	2009 (2007–2010)	2008 (2007–2009)	2008 (2007–2009)
≥1 acute OCS prescription	136 (10)	332 (12)[a]	99 (8)	112 (9)	129 (10)	128 (10)	130 (10)	127 (10)	155 (11)	309 (12)
Mean daily SABA dose (μg/d)										
0	989 (72)	1519 (57)[d]	902 (73)	902 (73)	934 (71)	930 (70)	938 (71)	945 (71)	847 (61)	1653 (62)
1–100	294 (21)	759 (28)	269 (22)	269 (22)	289 (22)	274 (21)	287 (22)	286 (22)	362 (26)	695 (26)
101–200	65 (5)	234 (9)	50 (4)	50 (4)	64 (5)	79 (6)	64 (5)	59 (5)	107 (8)	200 (8)
>200	34 (3)	170 (6)	23 (2)	23 (2)	34 (3)	38 (3)	34 (3)	33 (3)	63 (5)	134 (5)
LABA	44 (3.2)	68 (2.5)	8 (0.6)	8 (0.6)	38 (2.9)	33 (2.5)	40 (3.0)	36 (2.7)	34 (2.5)	71 (2.7)
LTRA	40 (2.9)	21 (0.8)[d]	3 (0.2)	3 (0.2)	22 (1.7)	18 (1.4)	19 (1.4)	18 (1.4)	20 (1.5)	39 (1.4)
≥1 hospital admission	30 (2.2)	20 (0.7)[d]	24 (1.9)	6 (0.5)[b]	16 (1.2)	20 (1.5)	13 (1.0)	18 (1.4)	18 (1.3)	35 (1.3)
≥1 severe exacerbations	159 (12)	348 (13)	117 (9)	117 (9)	139 (11)	144 (11)	138 (10)	141 (11)	167 (12)	339 (13)
Risk-domain asthma control	1223 (89)	2334 (87)	1127 (91)	1127 (91)	1182 (90)	1177 (89)	1185 (90)	1182 (89)	1213 (88)	2344 (87)
Overall control	1195 (87)	2194 (82)[a]	1105 (89)	1105 (89)	1154 (87)	1145 (87)	1157 (88)	1155 (87)	1159 (84)	2233 (83)

Notes: Data are n (%) unless otherwise noted. Smoking status and body mass index are not reported as data were available for only 1.5% and 7% of patients, respectively. [a]*P*<0.001 Mann–Whitney for comparison between cohorts. [b]*P*<0.05 conditional logistic regression for comparison between cohorts. [c]Evidence of comorbidities defined as recorded ICD-9 or ICD-10 code (International Classification of Disease) or via appropriate prescriptions during baseline and/or outcome year: nasal corticosteroids for rhinitis, proton pump inhibitors for GERD, topical corticosteroids for eczema, and topical oral antifungal medication for thrush. [d]*P*<0.05 χ^2 for comparison between cohorts.
Abbreviations: FP ICS, fine-particle inhaled corticosteroid; GERD, gastroesophageal reflux disease; ICS, inhaled corticosteroid; IPTW, inverse probability of treatment weighting; IQR, interquartile range; LABA, long-acting beta-agonist; LTRA, leukotriene receptor antagonist; OCS, oral corticosteroid; RiRL, Research in Real-Life; SABA, short-acting β2-agonist; SD, standard deviation.

Baseline patient characteristics for unmatched cohorts and the cohorts selected by exact matching and propensity score methods are depicted in Table 1. In the unmatched population, patients in the ciclesonide cohort received fewer baseline prescriptions for SABA but more for proton pump inhibitors (for treating GERD) than patients in the fine-particle ICS cohort. All matched samples tended toward the characteristics of the unmatched ciclesonide cohort.

Table 2 Demographic and baseline covariates included in the propensity score estimation

Covariates included		
Initial list of covariates examined (22)	**Non-collinear covariates included (15)**	**Variables contributing to the model (12)**
Age[a]	X	X
Sex	X	
Year of ICS initiation[a]	X	X
Time from first asthma prescription[a]		
Evidence of rhinitis (Y/N)[a,b]	X	X
Evidence of eczema (Y/N)[a,b]	X	X
Evidence of GERD (Y/N)[a,b]	X	X
Evidence of cardiac disease or hypertension (Y/N)[a,b]		
Prescriptions for beta blockers (Y/N)[a,c]		
Prescriptions for NSAIDs (Y/N)[c]		
Prescriptions for paracetamol (Y/N)[c]	X	X
Prescriptions for tricyclic agents (Y/N)[c]	X	
Prescriptions for statins (Y/N)[c]	X	
Number of prescriptions for allergies (categorized)[c]		
Number of prescriptions for acute oral corticosteroids (0/≥1)[a]	X	X
Number of prescriptions for SABA (categorized)[a]		
Number of SABA inhalers (categorized)[a]		
Average daily SABA dose (categorized)[a,d]	X	X
LABA prescription (Y/N)	X	X
LTRA prescription (Y/N)[a]	X	X
Hospital admissions for asthma (Y/N)[a]	X	X
Evidence of thrush (Y/N)[a,b]	X	X

Notes: [a]$P<0.05$ for comparison between cohorts (for beta blockers $0.05<P<0.10$). [b]Evidence of comorbidities defined as recorded ICD-9 or ICD-10 code or via appropriate prescriptions during baseline and/or outcome year: nasal corticosteroids for rhinitis, topical corticosteroids for eczema, proton pump inhibitors for GERD, topical oral antifungal medication for thrush, and cardiac glycosides, antihypertensive agents, diuretics, beta blocking agents, calcium channel blockers, and ACE (angiotensin-converting enzyme) inhibitors for cardiac disease/hypertension. [c]One or more prescription(s) received during the baseline year or at the initiation date of ICS therapy. [d]Calculated as (count of inhalers × doses in pack/365) × µg strength.
Abbreviations: GERD, gastroesophageal reflux disease; ICS, inhaled corticosteroid; NSAIDs, nonsteroidal anti-inflammatory drugs; LABA, long-acting β2-agonist; LTRA, leukotriene receptor antagonist; SABA, short-acting β2-agonist; Y/N, yes/no.

In the exact-matched dataset, standardized differences were outside the 10% corridor for prescriptions for allergies (both measured on the interval scale and categorized), hospital admissions for asthma, evidence of rhinitis, and evidence of GERD (Figure 1). Exact matching selected the sample with least severe asthma: 91% recorded no exacerbations at baseline, compared with 87%–88% in the unmatched dataset and 89%–90% in the datasets matched on propensity score.

For the PSM datasets produced using the RiRL and greedy algorithms (Table 1), all standardized differences were within the range of −0.1 to 0.1 (ie, absolute values within 10%) for both matching algorithms (Figure 1).

Using the IPTW method, a pseudo-dataset was created with sample size of 4063 (1380 and 2683 patients in ciclesonide and fine-particle ICS cohorts, respectively). The two cohorts were well balanced, and overall characteristics of the full unmatched population were retained (Table 1). All standardized differences were within the −0.1 to 0.1 range for the weighted pseudo-dataset, including those for the two variables where there remained a statistically significant difference at baseline (Figure 1).

For the unmatched and all matched populations, and the IPTW pseudo-dataset, the median (interquartile range) prescribed dose of ciclesonide at initiation was 160 µg/day (160–160) whereas that of fine-particle ICS (fluticasone-equivalent dose) was 500 µg/day (250–500; $P<0.001$).

Evaluation of treatment effects by study endpoint

Unadjusted results for study endpoints are presented in Table S2; unadjusted and adjusted RRs and ORs with each method are presented in Figure 2A–D. Details of the variables used to adjust the models are listed in the footnotes of Figure 2A–D.

Results for severe exacerbations showed a reduction in the RR for the treatment effect relative to the unmatched, unadjusted results (RR: 0.73; 95% CI: 0.58–0.90) using all analysis methods except for stratification by propensity score, which could not be used for this endpoint (see the "Additional results" section in Supplementary materials for explanation) and PSM with RiRL algorithm, which did not require adjusting for evidence of GERD (Table 3; Figure 2A). In the other matched datasets, the reduction in the RR was a result of adjustment for evidence of GERD, as the proportion

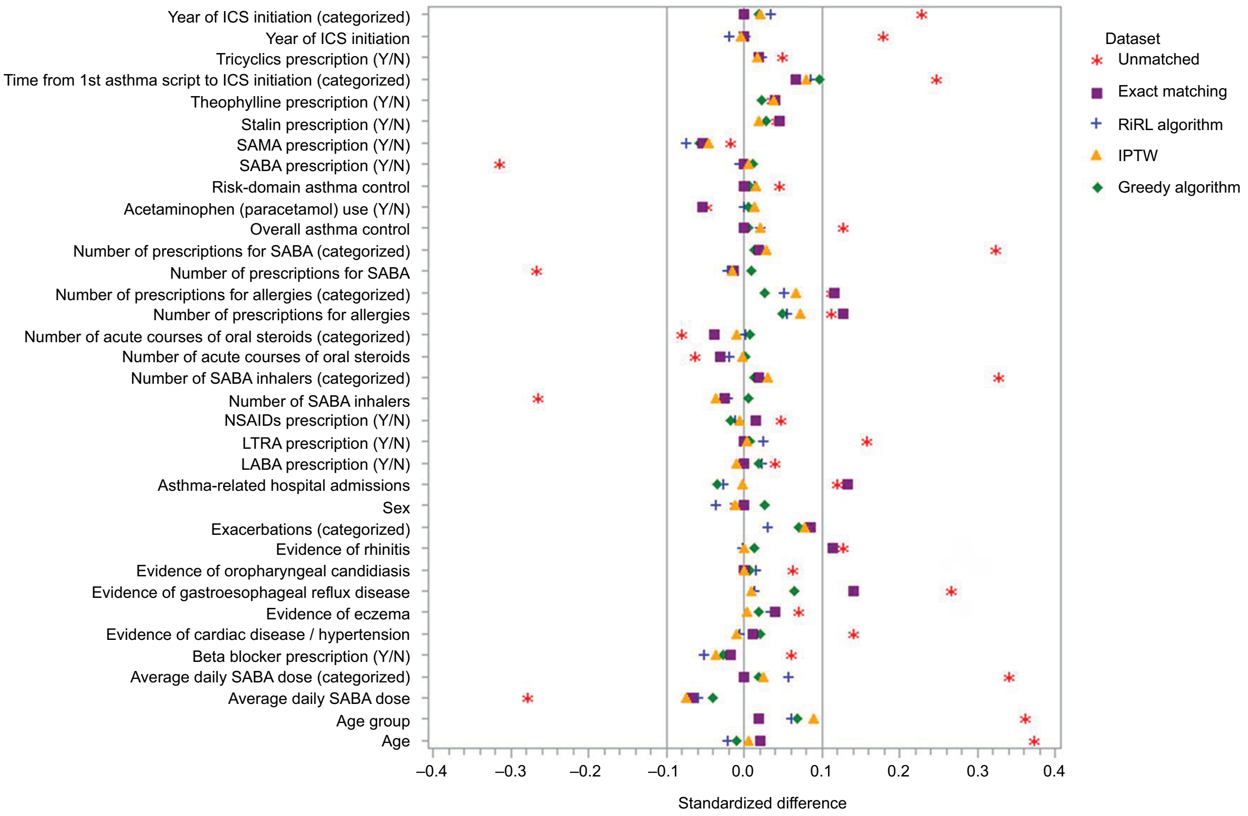

Figure 1 Standardized differences between cohorts in key baseline characteristics for the unmatched dataset, exact matching, propensity score matching, and the pseudo-dataset weighted by the stabilized IPTW. Absolute standardized differences in the unmatched dataset extended to 0.375, and for the exact-matched dataset, standardized differences were outside of the ±0.1 interval defining balance for allergy prescriptions, asthma-related hospital admissions, evidence of rhinitis, and evidence of GERD. All standardized differences were within ±0.1 for the datasets matched on propensity score and the pseudo-dataset weighted by IPTW.

Abbreviations: ICS, inhaled corticosteroid; GERD, gastroesophageal reflux disease; IPTW, inverse probability of treatment weighting; LABA, long-acting β2-agonist; LTRA, leukotriene receptor antagonist; NSAIDs, nonsteroidal anti-inflammatory drugs; RiRL, Research in Real-Life; SABA, short-acting β2-agonist; SAMA, short-acting muscarinic antagonist; Y/N, yes/no.

of patients with evidence of GERD remained significantly higher in the ciclesonide cohort than the fine-particle ICS cohort in those datasets.

For risk-domain asthma control, the adjusted ORs varied from 1.46 (PSM with RiRL algorithm) to 1.66 (exact matching and stratification). Both estimates using PSM lowered the OR from the unmatched, unadjusted whereas all other methods increased the OR fairly consistently to 1.63–1.66, when adjusted (Table 3; Figure 2B). CI widths for adjusted estimates varied from 0.68 (IPTW) to 0.88 (exact matching).

For overall asthma control, adjusted ORs varied between analysis methods from 1.80 (PSM with RiRL algorithm) to 2.21 (IPTW), all lower than the unmatched, unadjusted OR (2.29; 1.93–2.71; Table 3; Figure 2C). For PSM with RiRL algorithm, the adjusted result was lower than the unadjusted, driven by an adjustment for SABA use and negligible difference in evidence of GERD between cohorts, which drove the increase in adjusted ORs when using other methods.

Results of analyses of change in therapy were quite consistent across analysis methods, ranging from 0.69 (adjusted/unadjusted OR for PSM with RiRL algorithm) to 0.74 (weighted OR for IPTW). CIs were marginally greater using the matched datasets (Table 3; Figure 2D).

Table 3 summarizes our findings with regard to the use of each method of analysis both generally and specific to this study.

Discussion

We compared cohort matching and other methods of causal analysis and found that all methods – exact matching, PSM, IPTW, covariate adjustment, and stratification – produced similar results, namely, that ciclesonide, at much lower prescribed doses, was associated with better asthma-related outcomes than fine-particle ICS. The results varied slightly by method, depending on the patient subgroup selected, absolute and relative asthma severity, and residual differences between cohorts. However, the direction and statistical significance of the results remained comparable with all methods. Standardized differences lay outside of the 10% corridor in the exact-matched dataset for several variables,

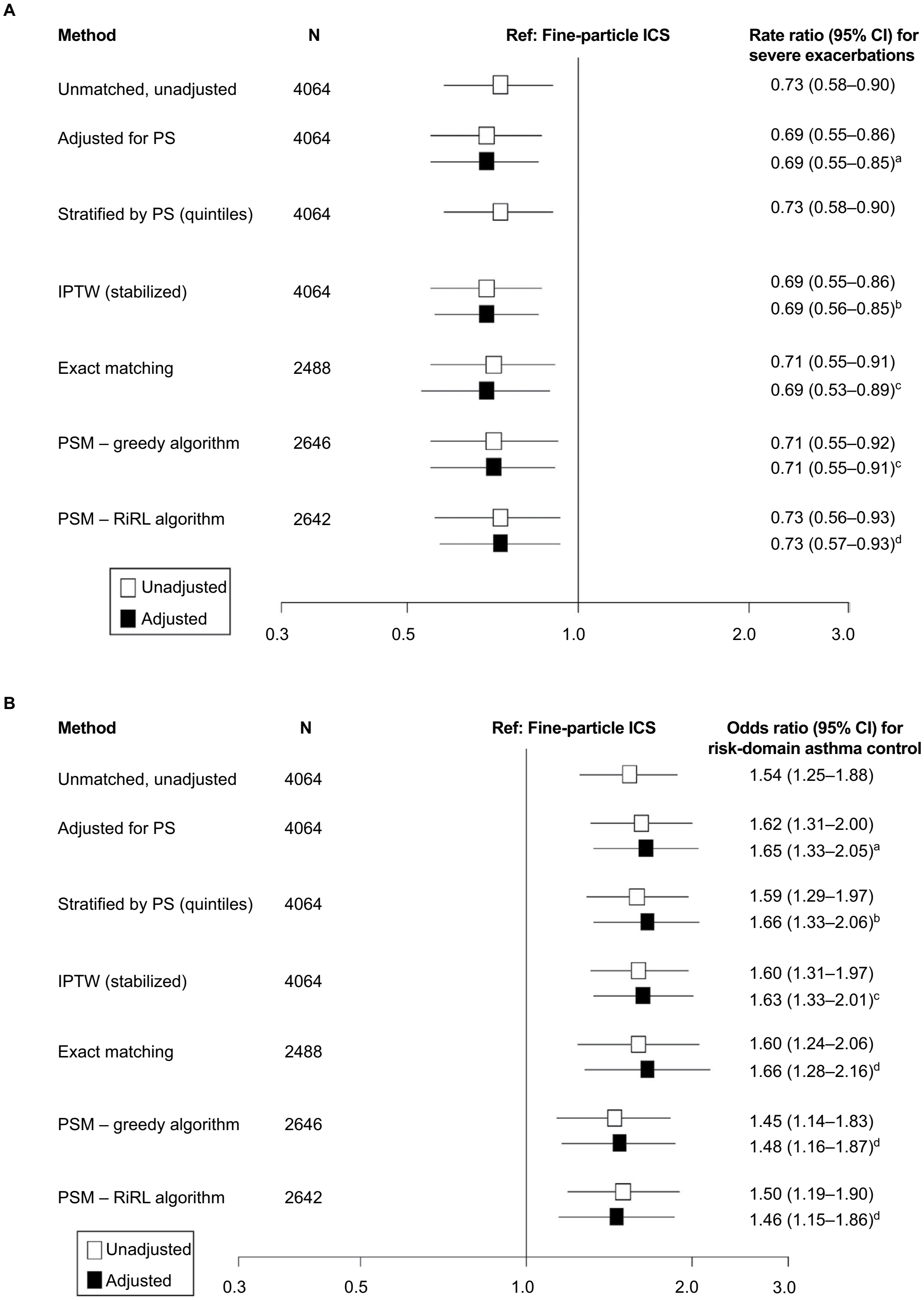

Figure 2 *(Continued)*

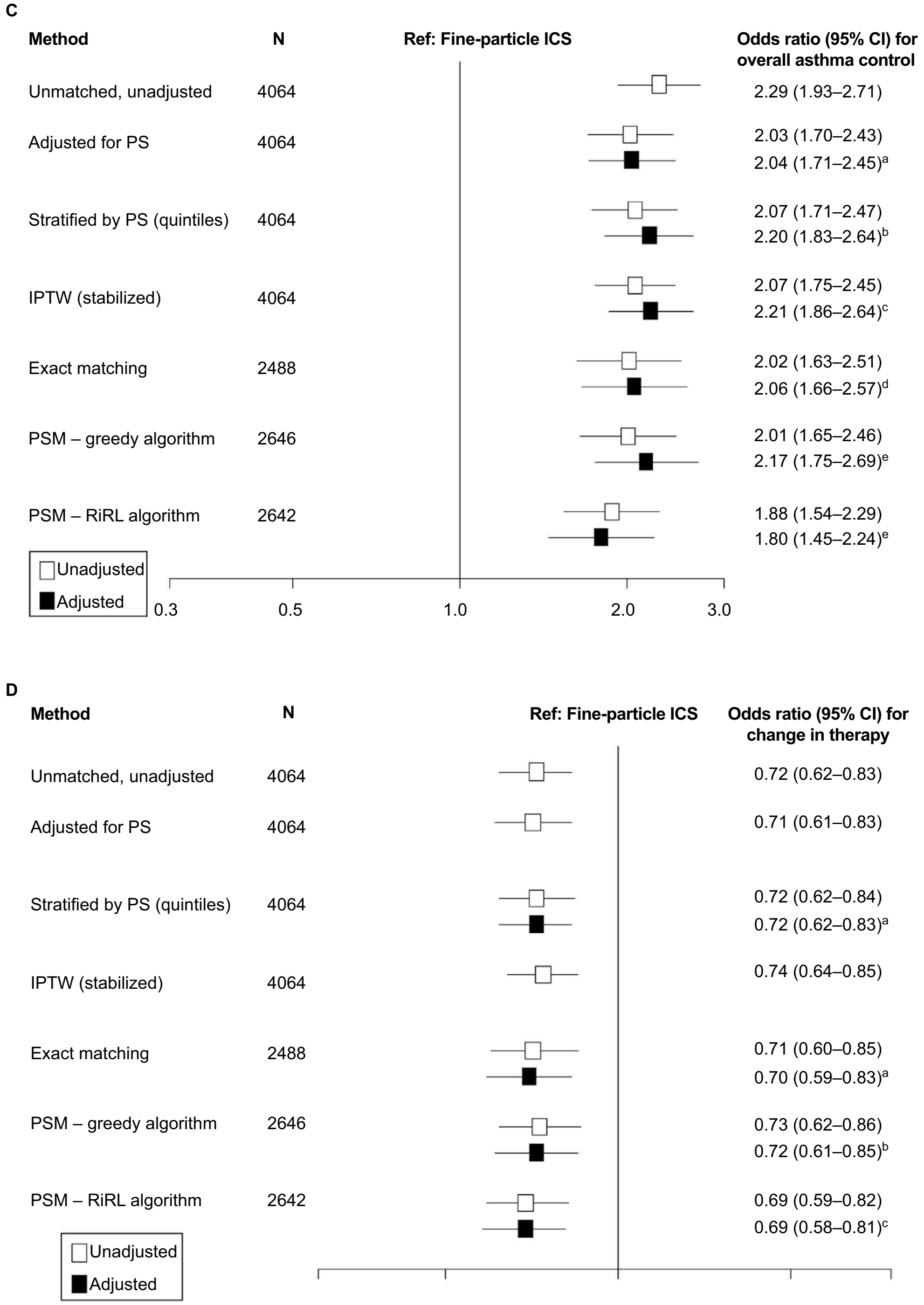

Figure 2 Comparison of outcomes using exact matching and propensity score methods.

Notes: (A) Results for comparison of exacerbation rates using exact matching and propensity score methods. [a]Adjusted for propensity score and baseline exacerbations (0/≥1). [b]Adjusted for age group and baseline exacerbations (0/≥1). [c]Adjusted for evidence of GERD and baseline exacerbations (0/≥1). [d]Adjusted for baseline exacerbations (0/≥1). Comparison of rate ratios (95% CIs) for severe exacerbation rates estimated using a Poisson regression model. **(B)** Results for comparison of risk-domain asthma control using exact matching and propensity score methods. [a]Adjusted for propensity score and baseline RDAC status. [b]Adjusted for the evidence of GERD and baseline RDAC status. [c]Adjusted for age group, evidence of GERD, and time from first asthma prescription. [d]Adjusted for evidence of GERD. Odds ratios compare ciclesonide versus the fine-particle ICS cohort (the latter set at odds=1.0). Odds ratios (95% CIs) for risk-domain asthma control estimated using a logistic regression model. (**C**) Results for comparison of overall asthma control using exact matching and propensity score methods. [a]Adjusted for propensity score, baseline RDAC status, and time from first asthma prescription. [b]Adjusted for evidence of GERD, leukotriene receptor antagonist use, baseline average daily SABA dose (categorized) and baseline RDAC status. [c]Adjusted for age group, evidence of GERD, baseline average daily SABA dose (categorized) and baseline RDAC status. [d]Adjusted for evidence of GERD and baseline overall asthma control. [e]Adjusted for evidence of GERD, baseline average daily SABA dose (categorized as 0/1–100/101–200/>200 μg) and baseline RDAC status. Odds ratios compare ciclesonide versus the fine-particle ICS cohort (the latter set at odds =1.0) and were estimated using a logistic regression model. (**D**) Results for comparison of change in therapy using exact matching and propensity score methods. [a]Adjusted for evidence of rhinitis and evidence of GERD. [b]Adjusted for evidence of GERD. [c]Adjusted for evidence of rhinitis. Odds ratios compare ciclesonide versus the fine-particle ICS cohort (the latter set at odds=1.0). Odds ratios (95% CIs) for change in therapy estimated using a logistic regression model.

Abbreviations: CI, confidence interval; GERD, gastroesophageal reflux disease; ICS, inhaled corticosteroid; IPTW, inverse probability of treatment weighting; PS, propensity score; PSM, propensity score matching; RDAC, risk-domain asthma control; RiRL, Research in Real-Life.

Table 3 Comparative characteristics of causal analysis methods tested for comparison between extrafine ciclesonide and larger fine-particle ICS in real-life patients with asthma from the PHARMO database

Methods	Advantages	Limitations	Measured effect
Exact matching	Patients are paired on defined key variables of interest	Some variables may remain unbalanced between cohorts Fewer remaining patients May select a sample not representative of the true population (in this study selected patients with slightly less severe asthma)	Average treatment effect for a typical treated patient
Propensity score matching	All variables of interest are well balanced (appropriate for situations with high numbers of confounders) In this study preserved close to full sample size (almost no excluded patients)		Average treatment effect for a typical treated patient
Inverse probability of treatment weighting	Preserves sample size (no excluded patients)		Average treatment effect at the population level
Covariate adjustment using propensity score	Preserves sample size (no excluded patients)		Average treatment effect at the population level
Propensity score stratification	Preserves sample size (no excluded patients)	PSS: inappropriate for count data outcomes modeled with Poisson	Average treatment effect at the population level

Notes: The term balance refers to standardized differences >10%. All methods provided similar results in terms of direction and statistical significance, in favor of the extrafine ciclesonide treatment. All results remained largely unchanged after adjustment for residual confounders.
Abbreviations: ICS, inhaled corticosteroid; PSS, propensity score stratification.

whereas all standardized differences were <10% for both PSM algorithms and the weighted pseudo-dataset.

Exact matching retained the lowest number of patients, hence had the lowest power and was potentially the least likely to be representative of the full population. However, adjusting for residual confounders after matching made only modest differences, particularly in the analysis of overall asthma control for which the adjustments in some other methods made quite large differences. This suggests that exact matching was effective in reducing confounding.

With PSM, both algorithms used to match on the propensity score (RiRL matching and greedy algorithms) retained similar numbers of patients. The pseudo-dataset generated by IPTW preserved almost all of the original sample size. For both PSM methods, and the IPTW, adjusting for residual confounders after matching again made only modest differences for most of the outcomes. However, there were larger differences after adjustment for residual confounders in the analysis of overall asthma control. This suggests that such confounding is important to investigate, rather than relying on the matching alone.

Covariate adjustment using the propensity score gave results consistent with other methods for all endpoints, and further adjustment had limited effects. Stratification by propensity score was not a suitable method for analyzing exacerbation rates as a primary endpoint but was suitable for the dichotomous endpoints.

In a prior case study comparing propensity score methods, Austin[30] reported that systematic differences between treatment cohorts were reduced more by PSM and IPTW than by covariate adjustment using the propensity score or stratification by propensity score. Recent studies have compared PSM and coarsened exact matching, a newer method that uses stratification followed by exact matching of cohorts for key variables influencing study endpoints with strata-based weighting according to the proportion of patients in each stratum.[12,31–33] The matching methods produced similar results in these studies; however, Wells et al[12] reported that coarsened exact matching retained more patients and achieved better balance between cohorts than PSM.

Our findings suggest that exact matching criteria could be informed by a propensity score calculation in addition to the usual clinical considerations. An alternative is to match on the propensity score following exact matching on key clinical characteristics.[34–36] For example, Kozma et al[36] in their study of health care resource use and costs for chronic obstructive pulmonary disease applied exact matching on four variables (sex, south region, pneumonia, and ischemic heart disease) followed by nearest available Mahalanobis distance matching within calipers defined by propensity scores. With any matching method, we recommend that standardized differences should be used, in conjunction with statistical testing, to assess the balance of treatment groups at baseline.[6,9,11,26] Other proposed methods of assessing balance may also be appropriate, including the z-difference or a weighted summary balance measure accounting for the strength of association of each covariate with the outcome.[37,38]

Another consideration when choosing matching and analytic methods is whether the ATE on the treated (ATT) or the ATE is of greater interest. The ATT, calculable using exact matching or PSM, is defined as the average response to treatment for a sample of individuals who are assigned treatment (in our study, "typical" patients prescribed ciclesonide). Instead, the ATE, which is calculable using IPTW, covariate adjustment for the propensity score, or stratification by the propensity score, is the average response to treatment for a random sample from a population.

Our findings suggest that the most appropriate matching method for a particular study should be selected according to study objectives, endpoints, and the available dataset. For example, if a treatment is prescribed primarily to a certain demographic group, the ATT may be more relevant than considering the treatment effect (by extrapolation) across other demographic groups (ATE). However, if the treatment effect across all demographic groups is of interest but limited data are available, the ATE would be more relevant. As noted above, our analyses using the matched datasets estimated the ATT, whereas the analysis methods that used the full unmatched dataset (IPTW, covariate adjustment, stratification) estimated the ATE. The proximity of the ATT to the ATE depends on the amount of overlap between the two cohorts in the unmatched dataset (in this case, ciclesonide and fine-particle ICS cohorts).

Limitations

Our methodological exercise has some limitations. The data in PHARMO reflect the real-life prescribing practices of Dutch physicians. As such, the individual decision to prescribe ciclesonide or a fine-particle ICS such as fluticasone, or no ICS, to a patient with asthma at any given time is likely variable. While we matched on measured baseline variables, the possibility of differences in unmeasured variables remains, and we cannot rule out residual confounding.

Another limitation of the present methodology study is inherent to the design of the original study on which the analyses were based: namely, it has to be assumed that a prescription identified in pharmacy data reflects the medications actually taken by the patient. However, a difference in adherence between treatment cohorts cannot be excluded and may have introduced bias into the comparison between cohorts.

Finally, the extrapolation of prescribing habits of Dutch physicians to other settings should be applied with caution.

Conclusion

The results of this study suggest that stratification by propensity score is not a suitable method where exacerbation rates are a primary endpoint. Otherwise, our findings suggest that all other methods (exact matching, PSM, IPTW, and covariate adjustment using the propensity score) have their particular strengths; and the most suitable method to fulfill study aims with regard to the dataset should be selected while factoring in study endpoints, relevance of ATT versus ATE, the overlap of treatment cohorts in the available data, and the estimated power of each method. Balance diagnostics should be applied with all methods to check the balance of confounders between treatment cohorts. Moreover, we recommend that at least two methods be applied for each matched cohort study to evaluate the robustness of the findings. If exact matching is used, the calculation of a propensity score could be useful to identify variables that require balancing, thereby to inform the choice of matching criteria together with clinical considerations.

Acknowledgments

We thank R Brett McQueen and Joan B Soriano for reviewing the manuscript and offering useful feedback.

This study was funded in equal parts by Takeda Pharmaceuticals International GmbH, Zurich, Switzerland; and by Research in Real-Life Ltd, UK, under a subcontract by Observational and Pragmatic Research Institute Pte Ltd, Singapore.

The dataset supporting the conclusions of this article is not available because the data were derived from a proprietary database provided by the PHARMO Database Network.

Author contributions

AB, JMK, DvE, and DBP developed the protocol for the study. RMCH and JAO provided expertise regarding use of the PHARMO database. AB and CM conducted the analyses, and EVH developed the first draft of the manuscript. All authors were involved in the interpretation of the data and the critical review and revision of the manuscript. All authors read and approved the final manuscript.

Disclosure

AB and CM were employees of Research in Real-Life (RiRL), Cambridge, UK. Research in Real-Life was subcontracted by Observational and Pragmatic Research Institute Pte Ltd, Singapore, to conduct this study and has conducted paid research in respiratory disease on behalf of the following other organizations in the past 5 years: Aerocrine, AKL Ltd, Almirall, AstraZeneca, Boehringer Ingelheim, Chiesi, GlaxoSmithKline, Meda, Mundipharma, Napp, Novartis, Orion, Takeda, Teva, and Zentiva, a Sanofi company.

NR has received over the past 3 years: 1) fees for speaking, organizing education, participation in advisory boards or consulting from 3M, Aerocrine, Almirall, AstraZeneca, Boehringer Ingelheim, Chiesi, Cipla, GlaxoSmithKline, MSD-Chibret, Mundipharma, Novartis, Pfizer, Sanofi, Sandoz, Teva; 2) research grants from Novartis, Boehringer Ingelheim and Pfizer.

EVH is a consultant to RiRL and has received payment for writing and editorial support to Merck.

The University of Groningen has received money for DSP regarding an unrestricted educational grant for research from AstraZeneca, Chiesi. Travel to conferences for the European Respiratory Society (ERS) and/or the American Thoracic Society (ATS) has been partially funded by AstraZeneca, Chiesi, GSK, Takeda. Fees for consultancies were given to the University of Groningen by AstraZeneca, Boehringer Ingelheim, Chiesi, GSK, Takeda, and TEVA. Travel and lectures in China were paid by Chiesi.

RMCH and JAO are employees of the PHARMO Institute. This independent research institute performs financially supported studies for government and related health care authorities and several pharmaceutical companies.

DvE and JMK are employees of Takeda.

DBP has Board Membership with Aerocrine, Almirall, Amgen, AstraZeneca, Boehringer Ingelheim, Chiesi, Meda, Mundipharma, Napp, Novartis, and Teva. Consultancy: Almirall, Amgen, AstraZeneca, Boehringer Ingelheim, Chiesi, GlaxoSmithKline, Meda, Mundipharma, Napp, Novartis, Pfizer, Teva, and Zentiva; Grants/Grants Pending with UK National Health Service, British Lung Foundation, Aerocrine, AstraZeneca, Boehringer Ingelheim, Chiesi, Eli Lilly, GlaxoSmithKline, Meda, Merck, Mundipharma, Novartis, Orion, Pfizer, Respiratory Effectiveness Group, Takeda, Teva, and Zentiva; Payments for lectures/speaking: Almirall, AstraZeneca, Boehringer Ingelheim, Chiesi, Cipla, GlaxoSmithKline, Kyorin, Meda, Merck, Mundipharma, Novartis, Pfizer, SkyePharma, Takeda, and Teva; Payment for manuscript preparation: Mundipharma and Teva; Patents (planned, pending or issued): AKL Ltd.; payment for the development of educational materials: GlaxoSmithKline, Novartis; Stock/Stock options: Shares in AKL Ltd which produces phytopharmaceuticals and owns 80% of Research in Real-Life Ltd, 75% of the social enterprise Optimum Patient Care Ltd and 75% of Observational and Pragmatic Research Institute Pte Ltd; received payment for travel/accommodations/meeting expenses from Aerocrine, Boehringer Ingelheim, Mundipharma, Napp, Novartis, and Teva; funding for patient enrolment or completion of research: Almirral, Chiesi, Teva, and Zentiva; peer reviewer for grant committees: Medical Research Council (2014), Efficacy and Mechanism Evaluation programme (2012), HTA (2014); and received unrestricted funding for investigator-initiated studies from Aerocrine, AKL Ltd, Almirall, Boehringer Ingelheim, Chiesi, Meda, Mundipharma, Napp, Novartis, Orion, Takeda, Teva, and Zentiva. The authors report no other conflicts of interest in this work.

References

1. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med.* 2000;342(25):1887–1892.
2. Krishnan JA, Schatz M, Apter AJ. A call for action: comparative effectiveness research in asthma. *J Allergy Clin Immunol.* 2011;127(1):123–127.
3. Price D, Bateman ED, Chisholm A, et al. Complementing the randomized controlled trial evidence base. Evolution not revolution. *Ann Am Thorac Soc.* 2014;11 (Suppl 2):S92–S98.
4. Rawlins M. De testimonio: on the evidence for decisions about the use of therapeutic interventions. *Lancet.* 2008;372(9656):2152–2161.
5. Roche N, Reddel HK, Agusti A, et al. Integrating real-life studies in the global therapeutic research framework. *Lancet Respir Med.* 2013;1(10):e29–30.
6. Austin PC. An Introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res.* 2011;46(3):399–424.
7. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci.* 2010;25(1):1–21.
8. Williamson EJ, Forbes A. Introduction to propensity scores. *Respirology (Carlton, Vic).* 2014;19(5):625–635.
9. Ali MS, Groenwold RH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *J Clin Epidemiol.* 2015;68(2):112–121.
10. Austin PC, Stuart EA. 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.* 2015;34(28):3661–3679.
11. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Making.* 2009;29(6):661–677.
12. Wells AR, Hamar B, Bradley C, et al. Exploring robust methods for evaluating treatment and comparison groups in chronic care management programs. *Popul Health Manag.* 2013;16(1):35–45.
13. Fullerton B, Pohlmann B, Krohn R, Adams JL, Gerlach FM, Erler A. The comparison of matching methods using different measures of balance: benefits and risks exemplified within a study to evaluate the effects of German disease management programs on long-term outcomes of patients with type 2 diabetes. *Health Serv Res.* 2016;51(5):1960–1980.
14. Dahl R, Engelstatter R, Trebas-Pietras E, Kuna P. A 24-week comparison of low-dose ciclesonide and fluticasone propionate in mild to moderate asthma. *Respir Med.* 2010;104(8):1121–1130.
15. Postma DS, Dekhuijzen R, Van der Molen T, et al. Asthma-related outcomes in patients initiating extrafine ciclesonide or fine-particle inhaled corticosteroids. *Allergy Asthma Clin Immunol.* 2016;8:e45.
16. PHARMO Database network. Available from: http://pharmo.nl/what-we-have/pharmo-database-network/. Accessed November 17, 2016.
17. Reddel HK, Taylor DR, Bateman ED, et al. An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. *Am J Respir Crit Care Med.* 2009;180(1):59–99.
18. Roche N, Reddel H, Martin R, et al. Quality standards for real-world research. Focus on observational database studies of comparative effectiveness. *Ann Am Thorac Soc.* 2014;11 (Suppl 2):S99–104.

19. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol.* 2008;61(4):344–349.
20. Colice G, Martin RJ, Israel E, et al. Asthma outcomes and costs of therapy with extrafine beclomethasone and fluticasone. *J Allergy Clin Immunol.* 2013;132(1):45–54.
21. Israel E, Roche N, Martin RJ, et al. Increased dose of inhaled corticosteroid versus add-on long-acting beta-agonist for step-up therapy in asthma. *Ann Am Thorac Soc.* 2015;12(6):798–806.
22. Martin RJ, Price D, Roche N, et al. Cost-effectiveness of initiating extrafine- or standard size-particle inhaled corticosteroid for asthma in two health-care systems: a retrospective matched cohort study. *NPJ Prim Care Respir Med.* 2014;24:14081.
23. van Aalderen WM, Grigg J, Guilbert TW, et al. Small-particle inhaled corticosteroid as first-line or step-up controller therapy in childhood asthma. *J Allergy Clin Immunol Pract.* 2015;3(5):721 e716–731 e716.
24. Schuster T, Lowe WK, Platt RW. Propensity score model overfitting led to inflated variance of estimated odds ratios. *J Clin Epidemiol.* 2016;80:97–106.
25. Parsons LS. Reducing bias in a propensity score matched-pair sample using greedy matching techniques. In: 26th Annual SAS Users Group International Conference; 2001; Long Beach, California.
26. Groenwold RH, de Vries F, de Boer A, et al. Balance measures for propensity score methods: a clinical example on beta-agonist use and the risk of myocardial infarction. *Pharmacoepidemiol Drug Saf.* 2011;20(11): 1130–1137.
27. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health.* 2006;60(7):578–586.
28. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat.* 1985;39(1):33–38.
29. Yang D, Dalton JE. *A Unified Approach to Measuring the Effect Size Between Two Groups Using SAS®*. Orlando, FL: SAS Global Forum; 2012.
30. Austin PC. A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behav Res.* 2011;46(1):119–151.
31. Htet S, Alam K, Mahal A. Economic burden of chronic conditions among households in Myanmar: the case of angina and asthma. *Health Policy Plan.* 2015;30(9):1173–1183.
32. Iacus SM, King G, Porro G. Multivariate matching methods that are monotonic imbalance bounding. *J Am Stat Assoc.* 2011;106(493): 345–361.
33. Winn AN, Shah GL, Cohen JT, Lin PJ, Parsons SK. The real world effectiveness of hematopoietic transplant among elderly individuals with multiple myeloma. *J Natl Cancer Inst.* 2015;107(8).
34. Deitelzweig S, Amin A, Christian R, Friend K, Lin J, Lowe TJ. Health care utilization, costs, and readmission rates associated with hyponatremia. *Hosp Pract (1995).* 2013;41(1):89–95.
35. Deitelzweig S, Amin A, Christian R, Friend K, Lin J, Lowe TJ. Hyponatremia-associated healthcare burden among US patients hospitalized for cirrhosis. *Adv Ther.* 2013;30(1):71–80.
36. Kozma CM, Paris AL, Plauschinat CA, Slaton T, Mackowiak JI. Comparison of resource use by COPD patients on inhaled therapies with long-acting bronchodilators: a database study. *BMC Pulm Med.* 2011;11:61.
37. Caruana E, Chevret S, Resche-Rigon M, Pirracchio R. A new weighted balance measure helped to select the variables to be included in a propensity score model. *J Clin Epidemiol.* 2015;68(12):1415 e1412–1422 e1412.
38. Kuss O. The z-difference can be used to measure covariate balance in matched propensity score analyses. *J Clin Epidemiol.* 2013;66(11): 1302–1307.

Supplementary materials

Additional methods

Methods of matching and causal analysis

We evaluated exact matching, propensity score matching (PSM), the inverse probability of treatment weighting (IPTW), covariate adjustment using the propensity score, and propensity score stratification. The IPTW, covariate adjustment, and stratification methods differ from PSM in that they retain the full dataset (so no biases are introduced through patient selection) but use the propensity score in other ways to achieve balance (ie, not just for matching patients).

Propensity score matching

For PSM, patients are matched on one variable, the estimated propensity score or logit of the propensity score within a predefined caliper, usually employing a 1:1 matching ratio although other ratios can be considered, as appropriate to the data. Because the precision of the propensity score is based on the inclusion of potential confounders into the statistical regression model used for its estimation, the true propensity score is not known. As a consequence, residual confounders can persist even after the application of the propensity score approaches.

Therefore, after applying PSM, we conducted a balance assessment by repeating the baseline analysis to ensure that the balance between cohorts was obtained and to test whether the propensity score model was adequately specified. We respecified the propensity score model by adding more variables (based on previous research experience), interactions, and non-linear terms until appropriate balance was obtained. Balance between cohorts was evaluated by comparing summary statistics of baseline variables via comparison of *P* values, using conditional logistic regression with significance set at *P*<0.05, and via use of standardized differences to compare mean values and prevalence of baseline variables; balance was considered achieved for differences lying within a 10% window. Standardized differences were calculated using a SAS macro developed by Yang and Dalton and available via the website of the Lerner Research Institute.[1]

Additional results

Exact matching

Exact matching retained the fewest patients (2488) and so was the lowest powered and least likely to be representative of the full population. Indeed, patients in the ciclesonide cohort selected for matching were marginally less severe than the overall unmatched population. Adjustment for residual confounders after matching made only modest differences (particularly in the analysis of overall asthma control for which adjustments in some other methods made quite large differences), suggesting that the matching was effective in reducing confounding. All models were adjusted for evidence of gastroesophageal reflux disease (GERD). This was not a matching variable and significant differences (41% vs. 34% in ciclesonide vs. fine-particle cohorts) remained at baseline after matching; standardized differences were in excess of 10%. Calculation of the propensity score showed this to be a strong predictor of treatment allocation, which maybe could have been improved by using the propensity score to influence choice of exact matching criteria. It would have been interesting to repeat the exact matching process, matching also on evidence of GERD, although the gain in balance across treatment arms would need to be weighed against a further loss in sample size and therefore power.

Propensity score matching

A list of 12 covariates to use for the propensity score estimation was identified after excluding 7 collinear variables and 3 variables not contributing to the final model (Table 2). Baseline daily short-acting β2-agonist (SABA) dose and evidence of GERD both strongly influenced the propensity score (see Table S1 for correlation coefficients).

Both algorithms used to match on the propensity score (Research in Real-Life [RiRL] matching algorithm and greedy algorithm) retained similar numbers of patients (2642 and 2646, respectively). In the PSM dataset produced using the RiRL algorithm, there were no significant differences

Table S1 Correlation coefficients between the propensity score and its components, ranked in order of absolute magnitude

Variable	Correlation coefficient
Average daily SABA dose (categorized)	–0.532
Evidence of GERD (Y/N)	0.446
Year of ICS initiation	0.291
LTRA prescription (Y/N)	0.288
Baseline asthma-related hospital admissions (categorized)	0.215
Evidence of rhinitis (Y/N)	0.210
Number of prescriptions for acute oral corticosteroids (categorized)	–0.132
Evidence of eczema (Y/N)	0.116
Evidence of thrush (Y/N)	0.110
Prescriptions for paracetamol (Y/N)	–0.078
LABA prescription (Y/N)	0.066
Sex	–0.018

Abbreviations: GERD, gastroesophageal reflux disease; ICS, inhaled corticosteroid; LABA, long-acting β2-agonist; LTRA, leukotriene receptor antagonist; SABA, short-acting β2-agonist; Y/N, yes/no.

Table S2 Unadjusted results for study endpoints

	Unmatched		Exact matching		Propensity score matching				Stabilized IPTW pseudo-dataset	
					RiRL algorithm		Greedy algorithm			
	Ciclesonide (n=1382)	FP ICS (n=2682)	Ciclesonide (n=1244)	FP ICS (n=1244)	Ciclesonide (n=1321)	FP ICS (n=1321)	Ciclesonide (n=1323)	FP ICS (n=1323)	Ciclesonide (n=1380)	FP ICS (n=2683)
Severe exacerbations										
0	1240 (90)	2281 (85)	1123 (90)	1065 (86)	1187 (90)	1128 (85)	1189 (90)	1136 (86)	1242 (90)	2277 (85)
≥1	142 (10)	401 (15)	121 (10)	179 (14)	134 (10)	193 (15)	134 (10)	187 (14)	138 (10)	406 (15)
Risk-domain asthma control	1240 (90)	2281 (85)	1123 (90)	1065 (86)	1187 (90)	1128 (85)	1189 (90)	1136 (86)	1242 (90)	2277 (85)
Overall control	1180 (85)	1928 (72)	1075 (86)	947 (76)	1127 (85)	996 (75)	1129 (85)	983 (74)	1169 (85)	1951 (73)
Change in therapy	360 (26)	882 (33)	329 (26)	416 (33)	341 (26)	444 (34)	343 (26)	432 (33)	372 (27)	894 (33)

Note: Data are n (%).
Abbreviations: FP ICS, fine-particle inhaled corticosteroid; IPTW, inverse probability of treatment weighting; RiRL, Research in Real-Life.

between cohorts in baseline variables at the 5% level. A trend (*P*<0.10) was recorded for shorter duration of asthma (*P*=0.083), lower mean daily SABA doses (*P*=0.096 on the ratio scale), and less short-acting muscarinic antagonist use (*P*=0.056) in the ciclesonide cohort as compared with the fine-particle inhaled corticosteroid cohort. Using the greedy algorithm, there was one significant difference between cohorts at the 5% level (higher incidence of GERD in the ciclesonide cohort; *P*=0.022) and a trend (*P*=0.094) for shorter duration of asthma in the ciclesonide cohort.

Adjustment for residual confounders after matching made only modest differences except in the analysis of overall asthma control, suggesting that the matching was, generally, effective in reducing confounding. Unadjusted and adjusted odds ratios (ORs) for risk-domain asthma control were lower than the unmatched and unadjusted, whereas ORs were higher using all other methods, which likely reflects the sample of patients selected. Furthermore, the adjusted ORs for risk-domain and overall asthma control were lower than the unadjusted ORs when using PSM with RiRL algorithm, whereas adjustments to the model in all other analysis methods increased the ORs. This apparent anomaly was driven by the magnitude of the residual baseline difference in evidence of GERD between cohorts (negligible difference using the RiRL algorithm, significant differences in other datasets), further confirming that models and results were sensitive to the sample selected, including both the absolute and relative severity of the patients and residual baseline and standardized differences between cohorts in key variables.

Inverse probability of treatment weighting

Stabilized weights – which multiply the IPTW by the unconditional probability of treatment allocation – were used to create a pseudo-dataset with sample size of 4063, so near-preserving the sample size of the original data. There were statistically significant differences between cohorts in mean daily SABA doses and prescriptions for allergies when measured on the ratio/interval scale, but cohorts were balanced when these variables were categorized; there were no other significant differences between treatment arms at the 5% level (Table 1 in the main paper). There was a trend (*P*=0.062) for a different distribution of severe exacerbations across treatment arms with greater proportions of patients in the ciclesonide cohort in the 0 and ≥2 categories. The largest differences were seen in baseline exacerbations (categorized) (0.08) and age group (0.09). Standardized differences were within the −0.1 to 0.1 range for the weighted pseudo-dataset, including the two variables where there remained a statistically significant difference at baseline; however, as Austin[2] notes, statistical significance is not the recommended method to assess balance and the standardized differences confirmed that an acceptable balance was achieved.

Adjustment for residual confounders made only minimal differences to the exacerbation and risk-domain asthma control endpoints, and no difference to the change in therapy endpoint, suggesting that the weighting was effective in reducing confounding. Adjusting for residual confounders increased the OR for overall asthma control, driven mainly by the residual difference in SABA use between cohorts. Overall, this method seemed effective in estimating the average treatment effect using the full power of the original dataset without selection bias. By using the stabilized weightings, treatment effects and their variances could be estimated simply using conventional modeling methods, with adjustments for any residual confounding.

Covariate adjustment using the propensity score

Covariate adjustment using the propensity score gave results consistent with other methods for all endpoints. Further adjustment was limited (outcome exacerbation rate was additionally adjusted for baseline exacerbations; outcome risk-domain asthma control and overall asthma control were additionally adjusted for baseline risk-domain asthma control status), but other adjustments (baseline SABA use, evidence of GERD) correlated strongly with the propensity score leading to collinearity in the model. As an exercise, we adjusted endpoint models for component baseline confounders rather than the propensity score and compared the results. Certainly for the exacerbation and risk-domain asthma control endpoints, there was very little difference in results between adjusting for the propensity score plus baseline exacerbation count and risk-domain asthma control status, respectively, than a full covariate list, but the propensity score adjustment made the models more parsimonious and simpler to reduce. There was more variation in results for the overall asthma control endpoint, but the fully adjusted model was very sensitive to adjustments and again, propensity score adjustment provided a simple and parsimonious option. When many potential confounders are involved, many of which may be collinear, the final model choice can be subjective. Thus, provided the propensity score is correctly specified, adjusting for the propensity score provides a simple, effective, and repeatable method to account for differences between treatment arms.

Interestingly, the propensity score was not a significant covariate in the model for therapy change; adjustments to the treatment effect could be made by adjusting for evidence of GERD and rhinitis and, for this endpoint, seemed the preferable option. The nonsignificance of the propensity score in this model, and the general stability of the therapy change results across all analysis methods, suggests that this endpoint was quite robust and not greatly influenced by treatment bias.

Stratification by propensity score

This method would not be an appropriate choice in a study where the primary endpoint is exacerbation rate. A negative binomial model cannot be stratified (using PROC GENMOD), and, whereas a Poisson model can be stratified, the unadjusted, stratified model took several hours to run, and there was insufficient memory to stratify and additionally adjust, even for one additional variable. Furthermore, the unadjusted stratified model gave identical results to the unmatched, unadjusted model. Stratification was possible and practical for the dichotomous endpoints.

Steering committee

The following independent steering committee agreed the study design and methods of the current study, before seeking approval from the governance board of the PHARMO database (detailed in the “Ethical approval” section of the main article).

Steering Committee

1. Professor Dirkje Postma, Groningen University, The Netherlands;
2. Thys van der Molen, Department of General Practice, University Medical Center Groningen, University of Groningen, The Netherlands;
3. Dr Elliot Israel: Brigham and Women’s Hospital and Harvard Medical School, USA;
4. Dr Gene Colice: George Washington University School of Medicine, Washington, DC, USA.

References

1. Yang D, Dalton JE. A unified approach to measuring the effect size between two groups using SAS®. In: SAS Global Forum 2012. *Statistics and Data Analysis*. Orlando, Florida. Paper 335–2012. Available from: http://www.lerner.ccf.org/qhs/software/lib/stddiff.pdf. Accessed April 28, 2016.
2. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399–424.