Journal of Clinical Epidemiology 169 (2024) 111305

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

A methodological review of the high-dimensional propensity score in comparative-effectiveness and safety-of-interventions research finds incomplete reporting relative to algorithm development and robustness

Guillaume Louis Martin[a,b,*], Camille Petri[c,d], Julian Rozenberg[e], Noémie Simon[a], David Hajage[a], Julien Kirchgesner[f], Florence Tubach[a], Louis Létinier[b], Agnès Dechartres[a]

[a]*Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, AP-HP, Hôpital Pitié Salpêtrière, Département de Santé Publique, Paris, France*

[b]*Synapse Medicine, Bordeaux, France*

[c]*UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London, UK*

[d]*National Heart and Lung Institute, Imperial College London, London, UK*

[e]*Sorbonne Université, AP-HP, Paris, France*

[f]*Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, AP-HP, Hôpital Saint-Antoine, Département de Gastroentérologie et Nutrition, Paris, France*

Accepted 20 February 2024; Published online 28 February 2024

Abstract

Objectives: The use of secondary databases has become popular for evaluating the effectiveness and safety of interventions in real-life settings. However, the absence of important confounders in these databases is challenging. To address this issue, the high-dimensional propensity score (hdPS) algorithm was developed in 2009. This algorithm uses proxy variables for mitigating confounding by combining information available across several healthcare dimensions. This study assessed the methodology and reporting of the hdPS in comparative effectiveness and safety research.

Study Design and Setting: In this methodological review, we searched PubMed and Google Scholar from July 2009 to May 2022 for studies that used the hdPS for evaluating the effectiveness or safety of healthcare interventions. Two reviewers independently extracted study characteristics and assessed how the hdPS was applied and reported. Risk of bias was evaluated with the Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I) tool.

Results: In total, 136 studies met the inclusion criteria; the median publication year was 2018 (Q1−Q3 2016−2020). The studies included 192 datasets, mostly North American databases (n = 132, 69%). The hdPS was used in primary analysis in 120 studies (88%). Dimensions were defined in 101 studies (74%), with a median of 5 (Q1−Q3 4−6) dimensions included. A median of 500 (Q1−Q3 200−500) empirically identified covariates were selected. Regarding hdPS reporting, only 11 studies (8%) reported all recommended items. Most studies (n = 81, 60%) had a moderate overall risk of bias.

Conclusion: There is room for improvement in the reporting of hdPS studies, especially regarding the transparency of methodological choices that underpin the construction of the hdPS.

Keywords: hdPS; Confounding; Bias; Real-world evidence; Reporting; Methodology

Role of the funding source: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author. Department of Public Health, Hôpital Pitié-Salpêtrière, 47-83, Boulevard de l'Hôpital, 75013 Paris, France.
E-mail address: guillaume.martin.md@gmail.com (G.L. Martin).

https://doi.org/10.1016/j.jclinepi.2024.111305

Plain language summary

The high-dimensional propensity score (hdPS) is a recently developed method to minimize bias in studies using medico-administrative databases, which lack important confounders as they were not built for research. Its algorithm works by automatically selecting variables to balance patients' characteristics in nonrandomized comparative groups. Despite potential benefits and widespread use, the hdPS has faced criticism for its perceived ''black-box'' nature and inconsistent implementation or performance. Our study aimed to evaluate the practical application of hdPS, specifically examining the methodology and reporting practices in studies using hdPS for comparative effectiveness or safety research. Our findings indicate that hdPS has been applied across diverse databases and research topics. However, we identified areas for improvement in reporting, particularly concerning the transparency of methodological choices guiding algorithm construction. We advocate for increased transparency in reporting methodological decisions to enhance the credibility and reproducibility of research using hdPS. Our results are valuable for researchers considering the implementation of hdPS in their studies, as well as for scientific and regulatory stakeholders that critically evaluate studies using this method.

What is new?

Key findings

- The high-dimensional propensity score (hdPS) has been proposed in 2009 to mitigate unmeasured confounding in real-world studies comparing the effectiveness or safety of interventions.
- The hdPS has been used in various contexts, but details regarding its development and robustness are often inadequately reported, which limits transparency and confidence in results.

What this adds to what was known

- The hdPS can be an effective tool to mitigate unmeasured confounding in real-world studies comparing the effectiveness or safety of interventions, but there are areas for improvement in its reporting, particularly concerning the transparency and robustness of methodological choices guiding algorithm construction.

What is the implication and what should change now

- Established guidelines for performing and reporting hdPS analyses should be endorsed. Future research could continue to explore whether parameters specific to databases or topics enhance its performance.

1. Introduction

The use of secondary healthcare databases has become increasingly common in the field of pharmacoepidemiology [1−3], notably for assessing the effectiveness and safety of healthcare interventions in real-world settings [4−7]. However, although secondary databases offer advantages in terms of population representativeness, low attrition rate, long-term follow-up, and statistical power, their use also presents some important limitations, especially regarding the lack of important confounders or the difficulty in identifying them [3].

To address this challenge, new statistical methods have been developed to handle confounding in such databases, including the high-dimensional propensity score (hdPS) [8], proposed in 2009. This approach operates on the premise that secondary healthcare data actually contain valuable information regarding unmeasured confounding via the concept of surrogate covariates or ''proxy'' [9,10], which are typically organized across multiple dimensions of different natures, such as diagnoses, medications, and procedures. Within these diverse dimensions, the algorithm aims to automatically rank and select hundreds of variables according to their prevalence, recurrence, and potential for confounding. Then in a propensity score (PS), it combines these selected data-driven covariates, called ''empirically identified,'' with ''investigator-specified'' covariates (ie, pertinent confounding variables that are available in the database, predefined by the investigators). This composite ''high-dimensional'' PS is considered an augmented PS that enhances the mitigation of confounding as compared with a PS composed exclusively of investigator-specified covariates.

The hdPS has been found an effective alternative for reducing bias as compared with classical PS approaches in secondary databases, according to the results of individual simulation or comparative studies [11,12], and has been used in many studies over the last decade. It has also been promoted by various Health Technology Assessment authorities, such as the US Food and Drug Administration, which included it in its routine drug safety Sentinel Initiative program [13,14], or by the French *Haute Autorité de Santé* in real-world evidence guidelines for industries [4]. Despite its potential benefits, the hdPS has been criticized, notably for its ''black-box'' approach, with heterogenous implementation or performance [15,16]. To date, its use

by researchers in practice has not been comprehensively assessed.

This study aimed to evaluate the methodology and reporting of studies that used the hdPS to assess the comparative effectiveness or safety of healthcare interventions in real-world settings.

2. Methods

This is a methodological review, adhering to the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement [17] (checklist provided in Appendix A).

2.1. Search strategy, eligibility, and selection

We searched MEDLINE via PubMed and Google Scholar from July 1, 2009 (date of initial publication of the method) to May 1, 2022 for studies by using the terms ''hdPS,'' ''hd-PS,'' or ''high-dimensional propensity scor*.'' Studies citing the hdPS initial paper [8] were also retrieved. All references were imported to Rayyan [18], and two reviewers (G.L.M. and C.P./J.R.) independently screened titles and abstracts to include studies assessing the comparative effectiveness or safety of healthcare interventions in real-world settings and using the hdPS. We did not include methodological papers or studies evaluating other questions (eg, prognostic or medico-economic studies). Disagreements were resolved by consensus, with the help of a third reviewer (A.D.) whenever necessary.

2.2. Data extraction

Two reviewers (G.L.M. and C.P.) independently extracted data from selected studies using a standardized data extraction form. Disagreements were resolved by consensus, with the help of a third reviewer (A.D.) whenever necessary. All data were extracted directly from the original publication including supplements, or referenced publications when the information had previously been reported elsewhere.

The following data were extracted for each study:

2.2.1. General characteristics

Publication date, involvement of epidemiologists or statisticians as defined elsewhere [19] (at least one of the authors belonging to a department/unit of epidemiology, clinical epidemiology, and/or biostatistics), country of first author, funding source, main study topic (specialty), study objective (effectiveness, safety, or both), and intervention type (pharmacological or not).

2.2.2. Database characteristics

Name(s) and number of dataset(s), type (ie, Payer Claims Database, Electronic Medical Records [EMRs]), and country(ies) of origin.

2.2.3. Methodological characteristics

Study design, causal contrast (''as-started'' [observational analog of intent-to-treat] or ''as-treated'' [analog of per-protocol]), sample size, or power calculation.

2.2.4. Reporting characteristics

Availability of registration and/or protocol, use of reporting guidelines, flow chart provided, availability of statistical code and/or data, and reported statistical software used.

2.2.5. Characteristics of the hdPS

First, whether the hdPS was used as primary analysis, then, for the first reported comparison using the hdPS, the following information regarding each of the 7 steps of the algorithm's development as explained and illustrated in Box 1:

1. Specify data dimensions and baseline period: number of data dimensions used, dimension categories (eg, inpatient diagnoses, outpatient drugs), and length of the main baseline covariate assessment period (ie, temporal window in which covariates will be identified before the index date, eg, 12 months).
2. Specify code terminology and granularity: The term ''code'' refers to standardized identifiers or labels used to categorize and classify medical information. We extracted whether code terminology (eg, International Classification of Diseases, 10th revision) and code granularity (eg, three digits) were reported.
3. Apply prevalence filtering: whether a prevalence filter for available variables was used and which type (ie, a specific number of most prevalent codes, eg, 200).
4. Assess code recurrence: whether variable recurrence was assessed across patients and how (eg, ''once, sporadic, frequent'' indicators, dividing each code in three binary variables: the code was recorded at least once for the patient, more than the median number of times, or more than the 75th percentile number of times, respectively).
5. Prioritize covariates: prioritization algorithm used to rank covariates (eg, Bross formula).
6. Select covariates: whether specific empirically identified covariates were excluded, whether authors reported the total number of empirically identified covariates to select and their list, and the number/list of investigator-specified variables added (ie, already known confounders available in the database).
7. Build and apply the hdPS: model used to build the hdPS, whether the hdPS was trimmed (eg, discarding observations with PSs below or above a threshold),

Box 1 High-dimensional propensity score (hdPS) principles and recommendations

hdPS principles and recommendations

Step 1: Specify data dimensions and baseline period

Choose *p* data dimensions that capture coherent, various aspects of care, based on the database characteristics

Choose a baseline covariate assessment period that should best capture unmeasured confounding in each dimension

Step 2: Specify code terminologies and granularity, aiming to capture the best level of information related to confounding

Step 3: Apply prevalence filtering
Use a prevalence filter of top *n* variables, based on the variety of codes in the dimension aiming to omit infrequent/erroneously used variables or for practical reasons

Step 4: Assess code reccurence
Across patients, assess the recurrence of codes such as using pre-defined cut-offs based on their frequency in the population, aiming to provide an indicator of a patient's underlying health.

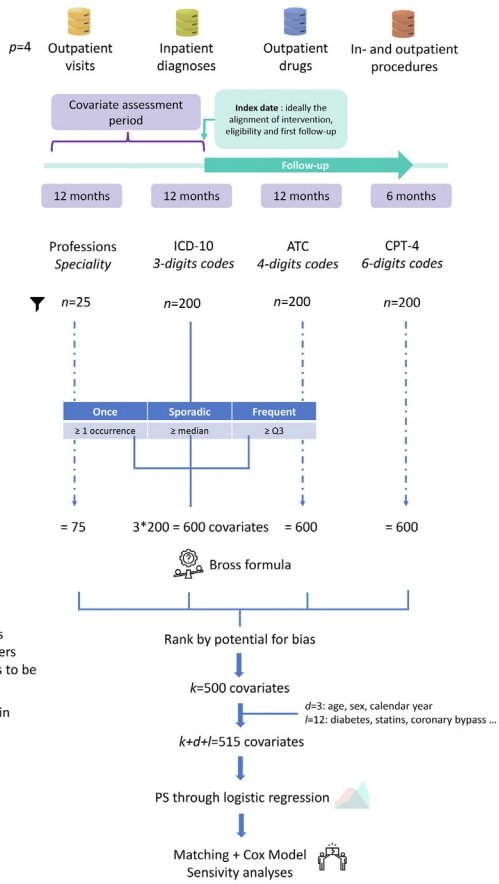
Step 5: Prioritize covariates
Apply a covariate selection algorithm within dimensions

Step 6: Select covariates
Rank variables according to their potential for bias across dimensions
Manually inspect and remove potential instrumental variables/colliders
Specify and report a final number of *k* empirically identified variables to be retained, taking into account practicality for model regression

Add investigator specified covariates (*d* = demographics, *l* = clinical), in particular those not represented in the previously used data dimensions

Step 7: Build and apply the hdPS
Estimate the propensity score, using an appropriate model
Report exchangeability diagnostics (i.e standardized differences)
Apply the propensity score to estimate intervention effect
Compare with non-hdPS method (e.g., PS with only investigator specified variables) and sensitivity analyses

Illustrated example



- Caption: This box shows in the left column the main principles for developing a high-dimensional propensity score, and relevant recommendations for each step. The right column shows an illustrated example, aligned with the aforementioned steps.

how the hdPS was applied on the population (ie, matching, stratification, weighting, adjustment), and model used to estimate the intervention effect (eg, Cox model).

2.2.6. *Inspection of the hdPS and assessment of confounding mitigation*

Whether the authors reported the use of diagnostic methods for exchangeability (standardized differences or c-statistic), the PS distribution plot, screened for instrumental variables (ie, using Z-bias [20]) and whether they reported crude and/or other non-hdPS analyses, methods assessing residual bias (eg, negative controls), or sensitivity analyses evaluating the hdPS robustness relative to methodological choices (eg, varying number of selected empirically identified covariates).

2.2.7. *Key reporting items regarding hdPS*

We evaluated the summary key items to report for hdPS analysis proposed by Tazare et al. [15]: specify data dimensions (corresponding to *Step 1*), describe parameters for generating pre-exposure features (*Step 2*), describe feature recurrence assessment (*Step 4*), specify covariate prioritization method (*Step 5*), specify total number of covariates to select (*Step 6*), and specify software used describe the results of diagnostics and sensitivity analyses (*Step 6 and 7*).

2.2.8. *Risk of bias*

Two reviewers independently evaluated the risk of bias for each study for their main result from hdPS analysis, using the ROBINS-I tool [21].

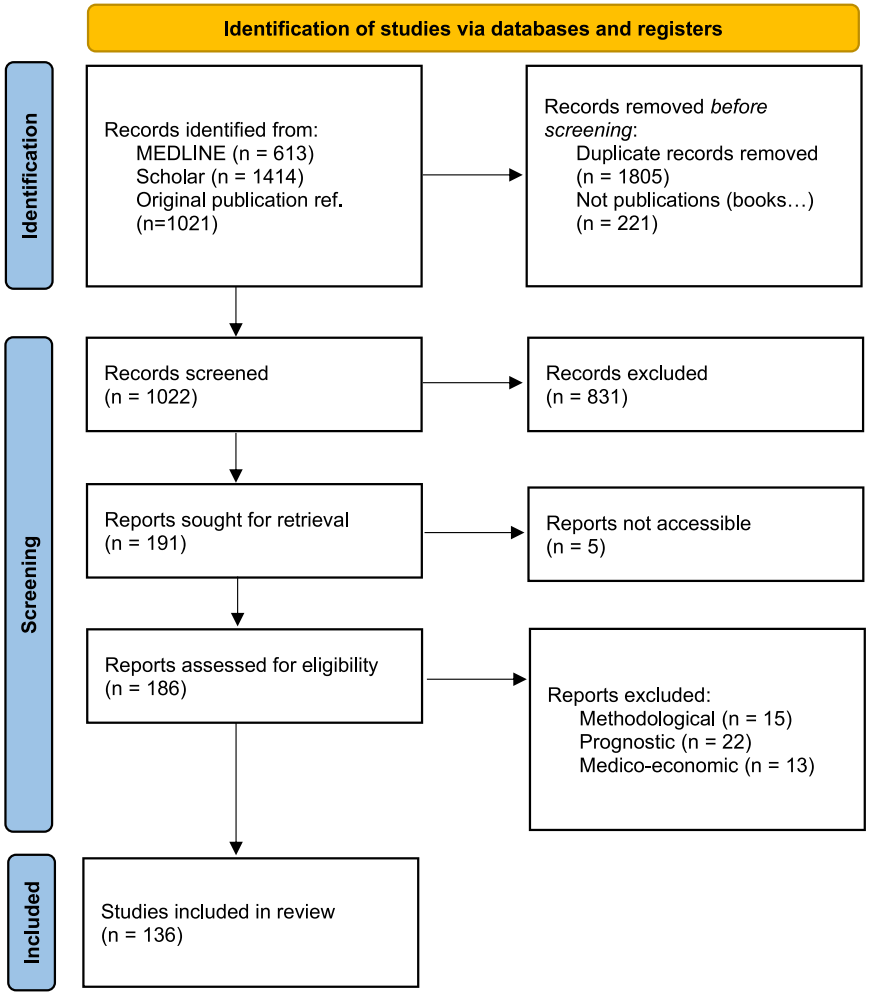


Figure 1. Flow diagram of selection of articles. Caption: The flow diagram depicts the flow of information through the different selection phases of the review. It maps out the number of records identified, included, and excluded, and the reasons for exclusions. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

2.3. Analysis

Our analysis was descriptive, reporting categorical variables as number (%) and quantitative variables as median (interquartile range [Q1−Q3]). We conducted all analyses using R 4.2.2 (R Foundation for Statistical Computing, Vienna, Austria).

3. Results

Our search retrieved 1022 unique studies; 136 were selected (PRISMA flow diagram in Fig 1).

3.1. General characteristics

General characteristics of the studies are in Table 1. The median publication year was 2018 (Q1−Q3 2016−2020). Most studies involved an epidemiologist or a statistician (n = 117, 86%) and the first author was mainly affiliated with institutions in North America, mainly Canada (n = 56, 41%) or the United States (n = 50, 37%). Public funding was reported for 109 studies (80%). The main topics were related to cardiology (n = 34, 25%), diabetology (n = 23, 17%), and psychiatry (n = 16, 12%). Studies evaluated the safety of interventions (n = 75, 55%), effectiveness (n = 29, 21%), or both (n = 32, 24%). Most interventions assessed were pharmacological (n = 121, 89%).

3.2. Database(s) characteristics

Most studies (n = 118, 87%) reported results based on a single dataset of patients (Supplemental Table 1). The studies analyzed 192 datasets overall, mostly based on Payer Claims Database (n = 134, 70%), EMRs/registries (n = 27, 14%) or chained from both types (n = 28, 15%). Datasets were from 15 different countries, mostly consistent with the authors' affiliations, with a majority from Canada (n = 68, 35%), the United States

Table 1. General characteristics of studies using the high-dimensional propensity score (hdPS) (N = 136)

Publication year, *median (Q1–Q3)*	2018 (2016–2020)
Participation of epidemiologists or statisticians[a]	
Yes	117 (86%)
No	7 (5%)
Unclear	12 (9%)
Country of first author	
Canada	56 (41%)
USA	50 (37%)
European country	20 (15%)
Other	10 (7%)
Funding source	
Public	109 (80%)
Private	13 (10%)
Both	7 (5%)
Not reported	7 (5%)
Medical topic (specialty)	
Cardiology	34 (25%)
Diabetology	23 (17%)
Psychiatry	16 (12%)
Rheumatology	10 (7%)
Infectious disease medicine	8 (6%)
Pulmonary medicine	8 (6%)
Others	37 (27%)
Study aim	
Safety	75 (55%)
Effectiveness	29 (21%)
Both	32 (24%)
Intervention type	
Pharmacological	121 (89%)
Nonpharmacological	15 (89%)

[a] An epidemiologist and/or biostatistician was considered as coauthor if at least one of the authors belonged to a department/unit of epidemiology, clinical epidemiology, and/or biostatistics.

Table 2. Methodological characteristics and reporting of studies using the hdPS (N = 136)

Study design	
New-user cohort design	124 (91%)
Prevalent-user cohort design	5 (4%)
Prevalent new-user cohort design	2 (1%)
Nested case-control study	5 (4%)
Comparator type	
Active comparator	100 (74%)
Standard of care	36 (26%)
Causal contrast	
As-started (intent-to-treat observational analog)	59 (43%)
As-treated (per-protocol observational analog)	44 (32%)
Both	31 (23%)
Unclear	2 (2%)
Reported sample size or power calculation	5 (4%)
Study protocol	
Available	7 (5%)
Reported but unavailable	39 (29%)
Not reported	90 (66%)
Use of a reporting guideline[a]	
RECORD, RECORD-PE	8 (6%)
STROBE	11 (8%)
Not reported	118 (87%)
Reporting of a flow chart	68 (50%)
Availability of statistical code	
On demand[b]	1 (1%)
Not reported	133 (99%)
Availability of data	
On demand (following approval)[b]	9 (7%)
Not available	3 (2%)
Not reported	124 (91%)
Reported statistical software used[a]	
SAS	104 (76%)
R	10 (7%)
Stata	8 (6%)
Other	2 (1%)
Not reported	18 (13%)

RECORD, REporting of studies Conducted using Observational Routinely Collected health data; RECORD-PE, REporting of studies Conducted using Observational Routinely Collected health data-PharmacoEpidemiology; STROBE, STrengthening the Reporting of OBservational studies in Epidemiology.

[a] Percentage may exceed 100% in cases of multiple reports within a single study.

[b] Considered ''on demand'' when reported as such in publications. We did not try to reach authors for verification.

(n = 64, 33%), and also the United Kingdom (n = 29, 15%).

3.3. Methodological characteristics

Most studies followed a new-user cohort design (n = 124, 91%), and most (n = 100, 74%) compared the intervention with an active comparator (Table 2). Causal contrast was ''as-started'' in 59 studies (43%), ''as-treated'' in 44 (32%), both in 31 (23%), and unclear in the remaining 2 (2%). Sample size or power calculation was infrequent, reported for only five studies (4%).

3.4. Reporting characteristics

A protocol was provided for seven studies (5%) or mentioned but unavailable for 39 (29%) (Table 2). Reporting guidelines for observational studies were reported for 18 studies (13%), specifically REporting of studies Conducted using Observational Routinely Collected health data (RECORD, including RECORD-PE, ie, for PharmacoEpidemiology) (n = 8, 6%) and STrengthening the Reporting of OBservational studies in Epidemiology

Table 3. hdPS characteristics and reporting for the first-reported comparison in studies using the hdPS (N = 136)

Use in primary analysis	120 (88%)
Step 1	
Definition of dimensions used (=*p*)	101 (74%)
Number of dimensions, *median (Q1–Q3)*	5 (4–6)
Definition of the baseline covariate assessment period	92 (68%)
Length of the period in days, *median (Q1–Q3)*	365 (365–365)
Step 2	
Reporting of codes terminology	42 (33%)
Reporting of codes granularity	37 (27%)
Step 3	
Prevalence filter used	
Cutoff of most prevalent codes	44 (32%)
Number of codes used (=*n*), *median (Q1–Q3)*	200 (200–200)
No prevalence filter used	3 (2%)
Reference to the original publication only[a]	86 (63%)
Not reported	3 (2%)
Step 4	
Feature recurrence assessment method	
''Once, sporadic, frequent'' indicators	32 (25%)
No feature recurrence assessment method used	6 (4%)
Reference to the original publication only[b]	76 (60%)
Not reported	3 (2%)
Step 5	
Covariate prioritization algorithm	
Bross bias formula	43 (32%)
Strength of association with the exposure	4 (3%)
Strength of association with the outcome	6 (4%)
No covariate prioritization algorithm used	1 (1%)
Reference to the original publication only[c]	79 (58%)
Not reported	3 (2%)
Step 6	
Exclusion of some empirically identified covariates	14 (10%)
Report of the number of selected empirically identified covariates (=*k*)	115 (85%)
Number included, *median (Q1–Q3)*	500 (200–500)
Report the list of selected empirically identified covariates	31 (23%)
Report investigator-specified covariates (=*d* + *l*)	117 (86%)
Number included, *median (Q1–Q3)*	15 (8–31)
Step 7	
Model used to build the propensity score	
Logistic regression	62 (46%)
LASSO	1 (1%)
Reference to the original publication only[d]	69 (51%)
Not reported	4 (3%)
Whether the hdPS was trimmed	
Yes	36 (26%)
No	6 (4%)
Not reported	94 (69%)
Primary use of the hdPS	
Adjustment	20 (15%)
Adjustment on deciles (poststratification)	23 (17%)
Matching	75 (55%)
Stratification	4 (3%)
Weighting	14 (10%)
Secondary use (eg, comparing matching and adjustment)	26 (19%)
Model used to estimate intervention effect (first reported outcome)	
Cox regression	98 (72%)
Logistic regression	22 (16%)
Poisson regression	11 (8%)
Fine Gray regression	3 (2%)
Other models	2 (2%)

[a] that is, 200 most prevalent codes.
[b] that is, ''Once, sporadic, frequent'' indicators.
[c] that is, Bross bias formula.
[d] that is, Logistic regression.

(STROBE) (n = 11, 8%). The flow chart for the population selection was reported for 68 studies (45%). The statistical code (on-demand, n = 1, 1%) and data (on-demand, n = 9, 7%) were rarely available. The statistical software used was reported for most studies, with SAS being the most popular (n = 104, 76%).

3.5. Characteristics of the hdPS

Most studies used the hdPS in the primary analysis (n = 120, 88%) (Table 3). For the 101 studies defining the dimensions used (74%), the authors used a median of 5 (Q1-Q3 4-6) dimensions. Supplemental Figure 1 shows the most frequently used dimensions: drugs in 97 studies (95%), diagnostics in 99 (97%), and procedures in 81 (79%). The baseline covariate assessment period was described for 92 studies (68%), with most studies (n = 59, 64%) using a 12-month period (*Step 1*). Code terminologies were described for 42 studies (33%) and code granularity for 37 (27%) (*Step 2*). The prevalence filter was reported for 47 studies (35%), with 44 (32%) using a cutoff for most prevalent codes of a median of 200 (Q1–Q3 200–200), whereas 86 additional studies (63%) only referred to the original publication that used this cutoff as well (*Step 3*). The recurrence assessment method was detailed for 38 studies (29%), most using the ''once, sporadic, frequent'' approach, whereas 76 additional studies (60%) only referred to the original publication that used this approach [8] (*Step 4*). Likewise, one-third of studies reported a specific covariate prioritization method (n = 54, 40%), mostly the Bross formula, whereas 79 additional studies (58%) only referred to the original publication that used this formula [8](*Step 5*). The total number of selected

Table 4. Inspection and assessment of confounding mitigation of included studies using the hdPS (N = 136)

Diagnostics for differences[a]	
Standardized differences[b]	73 (54%)
Within studies using matching (n = 75)	53 (71%)
C-statistic[c]	13 (10%)
None reported	55 (40%)
Inspection of the hdPS[a]	
Screening of instrumental variables	5 (4%)
hdPS distribution plots	26 (19%)
Sequential addition of variables	3 (2%)
None reported	103 (76%)
Sensitivity analysis modifying hdPS parameters (step 1–6)[a]	9 (7%)
Different dimensions used (Step 1)	1 (1%)
Different baseline period (Step 1)	1 (1%)
Different code granularity (Step 2)	1 (1%)
Different prioritization method (Step 5)	1 (1%)
Different total number of selected empirically identified covariates (Step 6)	6 (4%)
Reporting of crude analysis	67 (49%)
Comparison with other confounding mitigation methods (non-hdPS)[a]	
Propensity score with investigator-specified covariates only	30 (22%)
Multivariable analysis with investigator-specified covariates only	16 (12%)
None reported	89 (65%)
Use of negative controls (''falsification test'')	13 (10%)

[a] Percentages may exceed 100% in cases involving multiple approaches within a single study.
[b] Including alternatives (eg, weighted standardized differences).
[c] Either on the exposure model or after applying the hdPS to the population (eg, matching).

empirically identified variables was reported for 115 studies (85%), with a median of 500 (Q1–Q3 200–500) included variables. The list of selected empirically identified covariates was reported for 31 studies (23%). Investigator-specified covariates in the hdPS were reported for 117 studies (86%), with a median of 14 (Q1–Q3 8–31) covariates included (*Step 6*). Logistic regression was the most commonly reported model used for building the hdPS (n = 62, 46%), whereas 69 additional studies (51%) only referred to the original publication that used logistic regression [8]. Trimming was reported for 42 studies (31%). Most authors used matching on the estimated hdPS (n = 79, 62%) and 26 studies (19%) reported at least another use (eg, comparing matching and adjustment). Cox regression was the most frequently used intervention effect model (n = 98, 72%) (*Step 7*). Overall, 78 studies (57%) mimicked or referred to the base case from the original hdPS publication [8] (n = 200 prevalence filter, ''once, sporadic, frequent'' recurrence assessment, Bross formulation prioritization, k = 500 selected empirically identified covariates, logistic regression to build the hdPS).

3.6. Inspection of the hdPS

A total of 73 studies (54%) reported standardized differences and 13 (10%) the c-statistic (Table 4). Screening for instrumental variables and hdPS distribution plots were rarely reported (n = 5, 4% and n = 26, 19%). Sensitivity analyses assessing the hdPS robustness to parameter choices were reported for only nine studies (7%), mostly by varying the total number of selected empirically identified covariates (n = 6, 4%).

3.7. Assessment of confounding mitigation by the hdPS

Half of studies reported the results of a crude analysis (n = 67, 49%); 30 (22%) and 16 (12%) reported a comparison with a PS analysis or multivariable analysis based on investigator-specified covariates only, respectively (Table 4). Residual confounding was assessed with negative controls in 13 studies (10%).

3.8. Key reporting items for hdPS

Reporting of the key items proposed by Tazare et al. [15] is shown in a radar chart (Fig 2). With an evaluation strictly based on the publications, most items were poorly reported, except for data dimensions (74%), the total number of covariates to select (85%), and the software used (87%). Only 11 studies (8%) reported all items. When including the citation to the original publication [8] in the evaluation as appropriate reporting for feature recurrence assessment (''once, sporadic, frequent'' indicators), covariate prioritization method (Bross formula) and total number of covariates to select (k = 500), 16 studies (12%) reported all items. A stratified analyses between studies published $<$ 2018 vs $\geq$2018 (median publication date) did not show notable reporting differences over time (Supplemental Figure 2).

3.9. Risk of bias

The risk-of-bias assessment using the ROBINS-I tool is presented in Supplemental Figure 3. Most studies were considered to have a moderate overall risk of bias (n = 81, 60%) and 54 serious risk of bias (40%). The item regarding confounding was the most prone to bias, with no study considered at low risk. The items regarding deviations from intended interventions was the second most prone to bias, with 71 studies (52%) considered at low risk. Studies reporting all summary key items and those that did not report did not differ notably in overall risk of bias (Supplemental Figure 4).

4. Discussion

In this methodological review of the literature, we systematically evaluated the methodology of studies using the hdPS in comparative effectiveness and safety research and

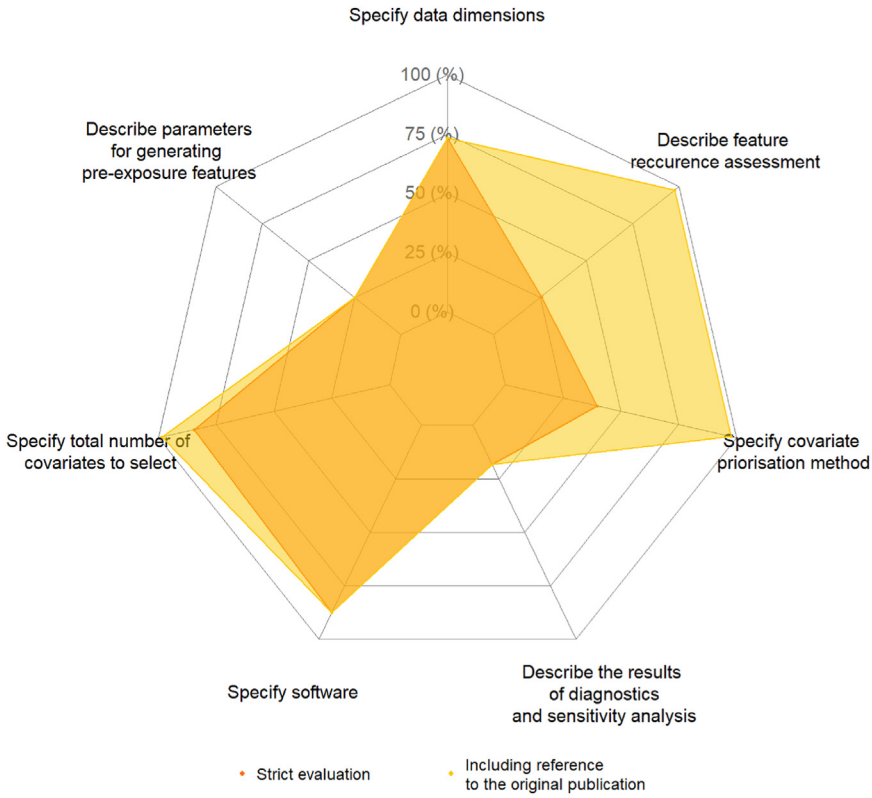


Figure 2. Transparency items recommended by Tazare et al., reported in included studies ($N = 136$). Caption: This radar chart shows the percentage of studies who appropriately reported the summary key items for hdPS (as recommended by Tazare et al.). The dark yellow shows a strict evaluation, only using the information available in studies. The faint yellow shows a looser evaluation, considering the reference to the original justification as appropriate reporting (base case parameters of the original publication). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

their transparency of reporting. The hdPS has been used in a wide variety of topics and across multiple databases from different countries, although with a strong North American predominance. Most studies used the hdPS in a new-user cohort design, used PS matching, and estimated as-started causal effects. Key parameters used to build the hdPS were infrequently reported, and most studies mimicked the base case of the original publication [8]. The assessment of confounding mitigation by the hdPS was seldom challenged in sensitivity analyses or by reporting inspection methods, although many studies adequately reported standardized differences used to assess the baseline characteristics balance between populations. Finally, half of the studies did not compare intervention effect estimates with the hdPS to more classical methods such as using a PS based on investigator-specified variables only.

To the best of our knowledge, this is the first study evaluating this topic. We tried to apply rigorous methodology to increase the validity of our findings, including independent data selection and extraction. We also used recently published tools by experts in the field to assess the transparency of hdPS reporting [15]. In addition, to assess risk of bias, we used the ROBINS-I tool, which is specifically designed for observational studies evaluating interventions [21]. Several limitations should be noted. First, although we used multiple sources to identify studies in our search strategy, studies reporting hdPS analysis outside searchable elements may have been missed, and we did not include the gray literature. Our selection of articles did not aim to be exhaustive but rather to provide a representative sample of the literature on this topic. Second, we did not compare the methodology, reporting, or risk of bias for studies not using the hdPS in similar settings, which limits the comparative interpretability of our findings. Third, some items, like trimming, may be reported selectively based on their applicability, but may be omitted when irrelevant. This is a common challenge in research assessing the transparency of studies that integrate diverse designs or approaches. Finally, our use of recently published tools should be considered in the context of their evolving nature, which may limit their applicability to earlier studies that predate their development.

Our review reveals several limitations in the reporting and methodology of hdPS studies. First, most studies failed to report key parameters used to build the score, in particular steps 1−4 and the list of generated empirically identified covariates. Such information is important for replicability and to understand the data used to reduce confounding by proxy. Reporting it might help alleviate

criticisms of the method, which is frequently considered a ''black-box'' approach. Authors also rarely reported sensitivity analyses testing the robustness of results in relation to methodological choices in constructing the hdPS. As previously noted, despite its standardized nature, the hdPS is a rather opaque process that could be tweaked to ''cherry-pick'' favorable results [22]. Reporting key parameters used to build the score and sensitivity analyses in a preregistered protocol can help mitigate these concerns. Second, a notable proportion of studies did not report the incorporation of investigator-specified covariates in the hdPS despite being recommended. Although including these covariates in a subsequent phase (ie, concurrently with hdPS for matching or adjustment) is theoretically feasible, their a priori integration enhances the clarity and comprehensibility of the hdPS ability to address confounding [23]. This enables the comparison of covariate distributions before and after hdPS application in a table, accompanied by standardized differences. Third, many studies did not compare the results of hdPS analyses with more classical methods (eg, using a PS of only investigator-specified covariates) or with residual bias methods such as negative controls, which raises concerns about the appropriateness of only reporting hdPS results given its previously mentioned ''black box'' approach. Although individual simulation or comparative studies have indicated that using hdPS may produce less biased results than regular approaches [8,24], the use of data-driven approaches for confounding variable selection remains a debate. Recent developments in epidemiology have for example emphasized the use of robust investigator-driven variable selection approaches to reduce confounding, notably with directed acyclic graphs [25]. In this context, we think that hdPS should continue to be compared with results obtained from nondata-driven approaches, such as a classical PS with investigators-specified variables only. Such comparisons are important for some readers to improve their confidence in results, and transparency is essential in establishing hdPS as a potential gold standard in secondary databases studies. Additionally, the target trial emulation framework [26] has also placed greater emphasis on the significance of study design relative to confounding mitigation. Our review revealed that hdPS was mostly used in new-user cohort design studies with an active comparator, which are optimal for the implementation of the algorithm. Nonetheless, hdPS benefits must not mask other methodological issues such as selection or measurement biases that can frequently occur in observational studies, as shown in our ROBINS-I assessment. Of note, regardless of the potential performance of hdPS in mitigating confounding bias, none of our studies was considered at low risk of bias related to confounding. This observation raises the question of the suitability of the ROBINS-I tool for assessing risk of bias in studies using innovative methods to address this bias, in particular those that rigorously incorporate techniques such as negative controls.

Overall, our findings indicate significant potential for improvement, especially concerning the transparency and reporting of methodological choices that underpin the construction of the hdPS. As is often the case with data-driven approaches, the lack of transparency is a common issue. In this context, checklists for the development, diagnosis, and reporting of hdPS analyses were recently published by two research groups led by Tazare et al. [15] and Rassen et al. [23]. The checklist from Tazare et al. was used in our study protocol development given its availability at that time. There are variations in key reporting concepts between the two groups. We have included a comparative table in Appendix B for researchers to evaluate these differences, which suggests the checklist from Rassen et al. tends to offer more details. The utilization of either tool should be encouraged, as only a limited number of included studies appropriately reported all items outlined by Tazare et al. Studies that effectively reported all key items typically presented hdPS specifications in a generic table. We provide a template for such tables in Appendix C to encourage better reporting.

In addition, only a few of studies compared or justified the parameters used to build the hdPS, which can affect the confidence in results. In particular, most studies only referred to the original publication of the algorithm, which reported parameters tested on US Medicare claims data [8]. These parameters were the most frequently reported in included studies, but they might not be optimal across other databases or clinical topics. This issue has been well assessed in a study using UK EMRs [27], showing that modifications in the feature recurrence assessment as well as code terminology used allowed for results closer to those obtained by randomized controlled trials. Rassen et al. began addressing this issue by developing in their publication guidelines for hdPS implementation applied to a specific study question and in specifically selected fit-for-purpose data sources [23]. As a next step, identifying topic-specific hdPS parameters could be useful for future researchers, especially in frequently used national secondary databases. In the presence of shared statistical codes, researchers could better reproduce published studies and challenge parameter choices in ancillary or sensitivity analyses. In the meantime, we provide in a GitHub repository the data analyzed in this study, which researchers may use to gain insights into how hdPS studies were implemented in specific contexts or databases.

5. Conclusion

This methodological review showed that the hdPS has been used across numerous databases and settings in comparative effectiveness and safety research. However, key parameters related to the construction of the hdPS score and sensitivity analyses were frequently under-reported, thus limiting the transparency and confidence in the results.

Addressing this concern could involve adopting recently published methodological tools [15,23] that offer recommendations on the implementation and reporting of hdPS analyses. We advocate for the inclusion of these tools in real-world evidence guidance issued by health authorities or in the evaluation of study reporting by journals. Additionally, future research could identify or compare topic-specific parameters used in specific databases and help inform a more targeted and effective use of the hdPS.

CRediT authorship contribution statement

Guillaume Louis Martin: Writing − original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Camille Petri:** Writing − review & editing, Conceptualization. **Julian Rozenberg:** Writing − review & editing, Data curation. **Noémie Simon:** Writing − review & editing. **David Hajage:** Writing − review & editing. **Julien Kirchgesner:** Writing − review & editing. **Florence Tubach:** Writing − review & editing. **Louis Létinier:** Writing − review & editing, Conceptualization. **Agnès Dechartres:** Writing − review & editing, Supervision, Methodology, Conceptualization.

Data availability

The dataset is available from the corresponding author on request and on the following link: https://github.com/GuillaumeMartinMD/hdPS-methodological-review.

Declaration of competing interest

G.L.M. reports a relationship with Synapse Medicine that includes employment. C.P. reports a relationship with Parexel International that includes employment. L.L. reports a relationship with Synapse Medicine that includes board membership and employment. There are no competing interests for any other author.

Acknowledgments

Laura Smales, BioMedEditing, for English editing of the manuscript.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111305.

References

[1] Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, et al. The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2017;26(8):954−62.

[2] Wang SV, Verpillat P, Rassen JA, Patrick A, Garry EM, Bartels DB. Transparency and reproducibility of observational cohort studies using large healthcare databases. Clin Pharmacol Ther 2016;99:325−32.

[3] Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323−37.

[4] Études en vie réelle pour l'évaluation des médicaments et dispositifs médicaux [Internet]. Haute Autorité de Santé. 2021. Report No.: 978-2-11-162649−2. Available at: https://www.has-sante.fr/upload/docs/application/pdf/2021-06/guide_etude_en_vie_reelle_medicaments__dm.pdf. Accessed January 12, 2023.

[5] Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. Clin Pharmacol Ther 2019;105:867−77.

[6] Jarow JP, LaVange L, Woodcock J. Multidimensional Evidence Generation and FDA Regulatory Decision Making [Internet]. JAMA 2017;318:703.

[7] McDonald L, Lambrelli D, Wasiak R, Ramagopalan SV. Real-world data in the United Kingdom: opportunities and challenges. BMC Med 2016;14(1):97.

[8] Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 2009;20:512−22.

[9] Greenland S, Robins JM. Confounding and misclassification. Am J Epidemiol 1985;122:495−506.

[10] VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat 2013;41(1):196−220.

[11] Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Anal 2014;72:219−26.

[12] Li L, Vollmer WM, Butler MG, Wu P, Kharbanda EO, Wu AC. A comparison of confounding adjustment methods for assessment of asthma controller medication effectiveness. Am J Epidemiol 2014;179:648−59.

[13] Connolly JG, Wang SV, Fuller CC, Toh S, Panozzo CA, Cocoros N, et al. Development and application of two semi-automated tools for targeted medical product surveillance in a distributed data network. Curr Epidemiol Rep 2017;4(4):298−306.

[14] Gagne JJ, Wang SV, Rassen JA, Schneeweiss S. A modular, prospective, semi-automated drug safety monitoring system for use in a distributed data environment. Pharmacoepidemiol Drug Saf 2014;23(6):619−27.

[15] Tazare J, Wyss R, Franklin JM, Smeeth L, Evans SJW, Wang SV, et al. Transparency of high-dimensional propensity score analyses: guidance for diagnostics and reporting. Pharmacoepidemiol Drug Saf 2022;31(4):411−23.

[16] Austin PC, Wu CF, Lee DS, Tu JV. Comparing the high-dimensional propensity score for use with administrative data with propensity scores derived from high-quality clinical data. Stat Methods Med Res 2020;29(2):568−88.

[17] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.

[18] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210.

[19] Delgado-Rodriguez M, Ruiz-Canela M, De Irala-Estevez J, Llorca J, Martinez-Gonzalez A. Participation of epidemiologists and/or biostatisticians and methodological quality of published controlled clinical trials. J Epidemiol Community Health 2001;55:569−72.

[20] Sauer BC, Brookhart MA, Roy J, VanderWeele T. A review of covariate selection for non-experimental comparative effectiveness research. Pharmacoepidemiol Drug Saf 2013;22(11):1139−45.

[21] Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.
[22] Hill J, Weiss C, Zhai F. Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. Multivariate Behav Res 2011;46(3):477−513.
[23] Rassen JA, Blin P, Kloss S, Neugebauer RS, Platt RW, Pottegård A, et al. High-dimensional propensity scores for empirical covariate selection in secondary database studies: planning, implementation, and reporting. Pharmacoepidemiol Drug Saf 2023;32(2):93−106.
[24] Schneeweiss S. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. Clin Epidemiol 2018;10:771−88.
[25] Hernán MA, Robins J. Causal inference: what if. Boca Raton: Chapman & Hall/CRC; 2023.
[26] Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016;183:758−64.
[27] Tazare J, Smeeth L, Evans SJW, Williamson E, Douglas IJ. Implementing high-dimensional propensity score principles to improve confounder adjustment in UK electronic health records. Pharmacoepidemiol Drug Saf 2020;29(11):1373−81.