Journal of Clinical Epidemiology 170 (2024) 111338

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

Use and reporting of inverse-probability-of-treatment weighting for multicategory treatments in medical research: a systematic review

François Bettega[a], Monique Mendelson[a], Clémence Leyrat[b], Sébastien Bailly[a,*]

[a]*University Grenoble Alpes, Inserm, Grenoble Alpes University Hospital, HP2, 38000 Grenoble, France*

[b]*Department of Medical Statistics, Inequalities in Cancer Outcomes Network, London School of Hygiene and Tropical Medicine, London, UK*

Accepted 25 March 2024; Published online 30 March 2024

Objectives: Causal inference methods for observational data represent an alternative to randomised controlled trials when they are not feasible or when real-world evidence is sought. Inverse-probability-of-treatment weighting (IPTW) is one of the most popular approaches to account for confounding in observational studies. In medical research, IPTW is mainly applied to estimate the causal effect of a binary treatment, even when the treatment has in fact multiple categories, despite the availability of IPTW estimators for multiple treatment categories. This raises questions about the appropriateness of the use of IPTW in this context. Therefore, we conducted a systematic review of medical publications reporting the use of IPTW in the presence of a multi-category treatment. Our objectives were to investigate the frequency of use and the implementation of these methods in practice, and to assess the quality of their reporting.

Study Design and Setting: Using Pubmed, Embase and Web of Science, we screened 5660 articles and retained 106 articles in the final analysis that were from 17 different medical areas. This systematic review is registered on PROSPERO (CRD42022352669).

Results: The number of treatment groups varied between 3 and 9, with a large majority of articles (90 [84.9%]) including 3 or 4 groups. The most commonly used method for estimating the weights was multinomial regression (51 [48.1%]) and generalized boosted models (48 [45.3%]). The covariates of the weight model were reported in 91 articles (85.9 %). Twenty-six articles (24.5 %) did not discuss the balance of covariates after weighting, and only 16 articles (15.1 %) referred to the assumptions needed to obtain correct inferences.

Conclusion: The results of this systematic review illustrate that medical publications scarcely use IPTW methods for more than two treatment categories. Among the publications that did, the quality of reporting was suboptimal, in particular in regard to the assumptions and model building. IPTW for multi-category treatments could be applied more broadly in medical research, and the application of the proposed guidelines in this context will help researchers to report their results and to ensure reproducibility of their research.

Keywords: Causal inference; Multicategory treatments; Medical research; Inverse probability of treatment weighting; Observational data; Reproducibility

1. Introduction

Randomized controlled trials (RCTs) typically provide the highest level of evidence for causal inference. In RCTs, participants are randomized to treatment groups, which ensures that, on average, observed and unobserved participants' characteristics are balanced across groups [1]. This balance is key to make causal inferences about the treatment(s) being studied. However, RCTs are not always feasible for ethical reasons (eg, when the exposure of interest is harmful) [2,3], are very costly, and sometimes lack generalizability and transportability because of stringent inclusion criteria. Real-world data have been increasingly used as an alternative or complement to RCTs [4]. Real-world observational studies present multiple advantages over RCTs: they usually include a wider diversity of patients (eg, older patients with comorbidities are often excluded from trials). Furthermore, the follow-up period is usually longer in retrospective studies using routinely collected data, and the data collection is already completed at the time of study design, reducing the delays for data analysis. Thus, observational studies are useful for investigating

Funding: Sébastien Bailly and François Bettega are supported by the French National Research Agency in the framework of the "Investissements d'avenir'' program (ANR-15-IDEX-02) and the ''e-health and integrated care and trajectories medicine and MIAI artificial intelligence'' Chairs of excellence from the Grenoble Alpes University Foundation. CL is supported by the UK Medical Research Council (MRC Skills Development Fellowship MR/T032448/1). This work has been partially supported by MIAI University Grenoble Alpes, (ANR-19-P3IA-0003)

* Corresponding author. CHU Grenoble Alpes, Service EFCR, Boulevard de la Chantourne, 38700 La Tronche, France.

E-mail address: sbailly@chu-grenoble.fr (S. Bailly).

https://doi.org/10.1016/j.jclinepi.2024.111338

What is new?

Key findings

- Medical publications scarcely use IPTW methods for more than two treatment categories.
- We found that among the publications which report use IPTW methods for multicategory treatments, the quality of reporting was suboptimal, in particular in regard to the assumptions and model building.

What this adds to what was known?

- We proposed a guideline with different steps for their reporting in the case of multi-category treatments.

What is the implication and what should change now?

- IPTW for multi-category treatments could be applied more broadly in medical research, and the application of the proposed guidelines in this context will help researchers to report their results and to ensure reproducibility of their research.

the long-term intervention effects as well as adverse events. Comparative effectiveness research using observational data has received increasing attention, and methodological work has been conducted to propose innovative designs, statistical tools and strategies for their analysis [5]. In recent years, accreditation bodies have been giving more credit to such study designs, as evidenced by the validation of Covid-19 vaccines by the FDA [6].

Nevertheless, unlike RCTs, observational studies are prone to confounding. When estimating the causal effect of a treatment, estimates from unadjusted analyses are biased if risk factors differ between groups. Several causal inference methods have been proposed to account for observed confounding, and they are split into two categories: those modeling the confounders-outcome relationships (eg, g-computation), and those modeling the confounders-treatment relationships (eg, propensity score methods) [7]. The latter has the advantage of mimicking RCTs, by recovering balance between groups on observed covariates using multiple balancing scores and allows researchers to define population based on conterfactuals [8]. Inverse-probability-of-treatment weighting (IPTW), in which patients are reweighted according to the inverse of their propensity of receiving the treatment actually received, creates a pseudo-population in which covariate distributions are similar between treatment groups. Because of its similarity with the philosophy of RCTs, IPTW is widely used for comparative effectiveness research [9].

However, IPTW is mostly implemented to estimate the causal effect of binary treatments [10], although researchers may be interested in the evaluation of several treatments (or several categories of treatment). This is the case when several treatments exist for the same indication, or when researchers want to compare the effect of two treatments and their combination [11−13]. IPTW estimators have been proposed in this context [14]. A review presented a methodological description of causal inference for multiple treatments [15] and suggested how to apply these methods, but it is unclear how often and how well they are used in practice. In particular, models for categorical outcomes are not always used for the estimation of the weights, but instead a series of models for binary outcomes are used. In addition, several types of modeling strategies can be used to estimate the weights, including parametric and nonparametric approaches, but it is unclear which approaches are commonly implemented.

Therefore, we conducted a systematic review of the medical literature to describe current practice in the use of IPTW for a multicategory treatment, and the quality of reporting of these studies. Based on the findings, we propose recommendations that we hope may contribute to a better transparency in the use and reporting of IPTW in this setting.

2. IPTW and causal inference on observational data

Unlike traditional statistics, estimating associations between exposures and outcomes, causal inference refers to specific hypotheses, study designs and statistical methods to draw causal conclusions from the data [16]. A specific framework, based on the concept of potential outcomes, has been developed to propose a causal language allowing the mathematical representation of causal questions. The potential outcome is what would have happened had the patient received a particular treatment [17]. Patients have as many potential outcomes as there are treatment categories. In this framework, the main issue is that only the effect of one treatment on a given patient can be observed, and other potential outcomes must be estimated from the data.

Within this framework, the causal effect can be identified from the data under the assumptions of consistency, no interference, positivity, and conditional exchangeability. Under consistency, the outcome of an individual under their observed exposure is the same as their potential outcome had they received their observed intervention via the hypothetical intervention [5,18]. The no interference assumption states that the treatment received by an individual has no influence on the potential outcomes of the other individuals. The positivity assumption states that, given their own characteristics, every individual has a nonzero probability of receiving any exposure categories [19]. Finally, the conditional exchangeability states that, given the measured variables, the exposure and potential outcomes are independent. The validity of these assumptions is required to be able to identify the causal effect from the data, but the additional assumption of a correct specification of the analysis model(s) is needed to ensure the validity of the causal effect estimate. These assumptions, together with the assumed causal relationships between confounding factors, treatments and outcomes are at the heart of the reasoning necessary for the application of causal inference methods, and their plausibility should always be discussed when reporting results.

IPTW is a weighting propensity score-based method [18]. Regarding Rosenbaum and Rubin's definition, the propensity score is "the conditional probability of assignment to a particular treatment given a vector of observed covariates" [8]. Thus, the propensity score is the probability, given the individuals' characteristics, to receive a specific treatment. When the treatment is binary, the probability of receiving the control treatment (or no treatment) is 1-propensity score. When the treatment has multiple categories, each individual has a propensity score for each treatment category [15,20,21]. The ATE can then be estimated using the IPTW estimator, in which individuals are weighted by the inverse of the probability of the treatment they actually received. Other methods, such as gradient boosting could be used [22]. However, the weights can be modified to target other estimands, such as the average treatment effect on the treated (ATT) or the average treatment effect in the overlapping population (ATO [23]). The balancing ability of the weights can be checked by comparing covariate distributions between treatment groups, using, for instance, standardized mean differences. IPTW estimates are unbiased if a good balance is achieved between groups, but residual imbalance can be addressed with augmented inverse-probability-of-treatment weighting (AIPTW) [24,25]. AIPTW combines multivariable regression and IPTW in a way that only one of the two models needs to be correctly specified to obtain unbiased estimates of the causal effect. Checking for the absence of extreme weights is also key to ensure the validity of the estimation. Weight truncation or trimming [26] is sometimes used to limit the contribution of large weights to the analysis [19], but this may lead to the estimation of an effect which does not coincide with the targeted estimand. Another important consideration when using IPTW estimators is variance estimation which must account for two aspects of the estimation. The uncertainty in propensity score estimation and the intraindividual correlation introduced via weighting should be captured in the outcome model to avoid misestimating the variance. Estimators based on the delta method [27] and nonparametric bootstrap have been proposed.

In summary, for the validity of IPTW and AIPW estimates we must ensure that: (i) the identification assumptions for causal inference are plausible, (ii) the estimator targets the correct estimands, (iii) the propensity score model is correctly specified, and (iv) appropriate variance estimators are used. It is, therefore, very important for these elements to be reported when publishing the findings of a study analyzed using IPTW.

Methods using IPTW for more than two treatment categories face additional technical challenges. Because of data scarcity or strong indication bias, the plausibility of the positivity assumption may be less likely when the number of treatment categories increases. With multiple treatment categories, it is necessary to question the choice of treatment reference for the estimand. Therefore, our systematic review focused on this setting, where a correct implementation and a clear reporting are required to ensure validity and reproducibility.

3. Methods

3.1. Inclusion and exclusion criteria

We performed literature searches on PubMed, Web of Science and Embase from January 01, 2011 to June 27, 2021, for peer-reviewed articles published in English. The systematic review included all publications in medical research involving human participants using an IPTW estimator with multiple treatment categories for the primary analysis. The review was limited to applied research and did not focus on methodological papers. The study is registered on PROSPERO (CRD42022352669).

There was no restriction in terms of research area, study design, type of intervention or outcome. Exclusion criteria were: nonmedical research, methodological studies (eg, simulation study, reviews.), nonoriginal research articles (eg, letters), articles using IPTW for subgroup or sensitivity analyses.

3.2. Search strategy

The search strategy screened articles whose abstracts, title or keywords contained the followings: inverse probability weight, inverse probability of treatment weight, augmented inverse propensity weight, as well as the associated acronyms (IPW, IPTW, AIPW, AIPTW, AIPWE). The generic term "propensity score" was not considered to improve the specificity of the algorithm, as previously done [9]. We also conducted a reverse search for articles citing McCaffrey et al. (2013) [14], Yoshida et al. (2018) [28], or Li and Li (2019) [29]. The search strategy is given in Appendix 1. The abstracts were then manually and independently screened for eligibility by FB and SB.

3.3. Extracted information

The extracted information was divided into nine fields: 1) description of the studies, 2) estimand and measure of association, 3) assumptions, 4) covariate selection, 5) propensity score estimation, 6) covariate balance, 7) analysis model, 8) software and statistical packages, and 9) good research practice (Table 1).

3.4. Data extraction procedure

A standardized, prepiloted form was used to extract data from the included studies and tested on 10 randomly

Table 1. Presentation of extracted information

Fields	Extracted information
Description of the included studies	Area of research
	Study registration number
	Study design
	Wording used to refer to IPTW
	Justification of the method
	Presence and appropriateness of sample size calculation
	Type of analysis model
	Sample size in each treatment groups
	Number of treatment groups
	Nature of the comparator
	Nature of the outcome
Estimand and measure of association	Estimand (ATE, ATT, ATO)
	Measure of association (HR, OR, RR, Other)
Assumptions	Mention of assumptions
	Mention of use of STROBE checklist
Covariate selection	Presence of DAG
	Variable include in weight model
	Method used for variable selection
	Method used for missing values
Propensity score estimation	Model used for propensity scores
	Type of weight
	Summary of weights
	Weight stabilization
	Weight trimming or truncation
Assessing covariate balance	Covariate balance
	Methods used for assessing balance
Analysis model	Method used (IPTW, AIPTW)
	Variance estimation method
Software and statistical packages	Name of software and packages used
Good research practice	Protocol
	Open data
	Open code

selected studies. The full text of the eligible studies identified after screening was retrieved and the data was extracted by FB and SB. Any disagreement over the eligibility or extracted items was resolved through discussion with CL if an agreement could not be reached.

4. Results

4.1. Screening and inclusion

The search yielded a total of 5299 articles (after the removal of duplicates), which were screened based on abstracts. From these, 303 were identified for full-text screening and 106 articles fulfilled the inclusion criteria and were included (complete list given in Appendix 2). The selection process is summarized in Figure 1.

4.2. Description of the included studies

Multicategory treatments were observed in 17 different medical fields, but three medical specialties accounted for half of included studies: 35 studies (47.3%) were either in cardiology (32 studies: 30.2%), 9 in nephrology (8.5%) and 7 in Gastroenterology (6.6%). Almost all the included articles (103, 97.2%) were cohort studies.

A variety of wording was used to refer to the method applied. IPTW and weighted regression were the two most common wording encountered ($n = 73$ (68.9%) of and $n = 15$ studies (14.2%), respectively). The majority of studies ($n = 100$ (94.3%)) justified the use of IPTW, the main reason being confounding adjustment.

The number of treatment groups ranged between 3 and 9 with a majority of studies comparing three groups ($n = 59$, 56.7%). Forty-Four articles (41.5%) had between 4 and 6 groups, and 1 (1%) article included 9 groups.

The total sample sizes ranged from 65 to 12,700,000 with a median of 161,583 participants. The minimum total sample size ranged from 12 to 638,905 with a median of 480. A summary of the results is presented in Figure 2 and Table 2.

4.3. Estimand and measure of association

In two-thirds of the articles (71 (67%)), the estimand was not clearly stated and had to be determined from the calculation of the weights, when this was available. The majority of the studies (90 (85%)) focused on estimating the ATE, (5 (4.7%)) focused on estimating the ATT and 1 (0.9%) article estimated the ATO. It was impossible to identify the estimand in 10 (9.4%) articles.

For the measure of association, 58 studies (54.7%) reported hazard ratios, 22 (20.8%) reported odds ratios, 8 (7.6%) reported risk ratios, 7 (6.6%) reported difference and 9 (8.5%) articles used other measures. It was impossible to determine the measure of association in 2 (1.9%) articles.

4.4. Assumptions

Only sixteen articles (15.1%) explicitly mentioned the assumptions underlying the validity of IPTW and discussed their context-specific plausibility.

4.5. Covariate selection and handling of missing data

The variables included in the weight model were mentioned in 91 (85.9%) articles. The method used for variable selection was specified in 28 (26.4%) of the articles: 14 (50%) used evidence from the literature, 11 (40.3%)

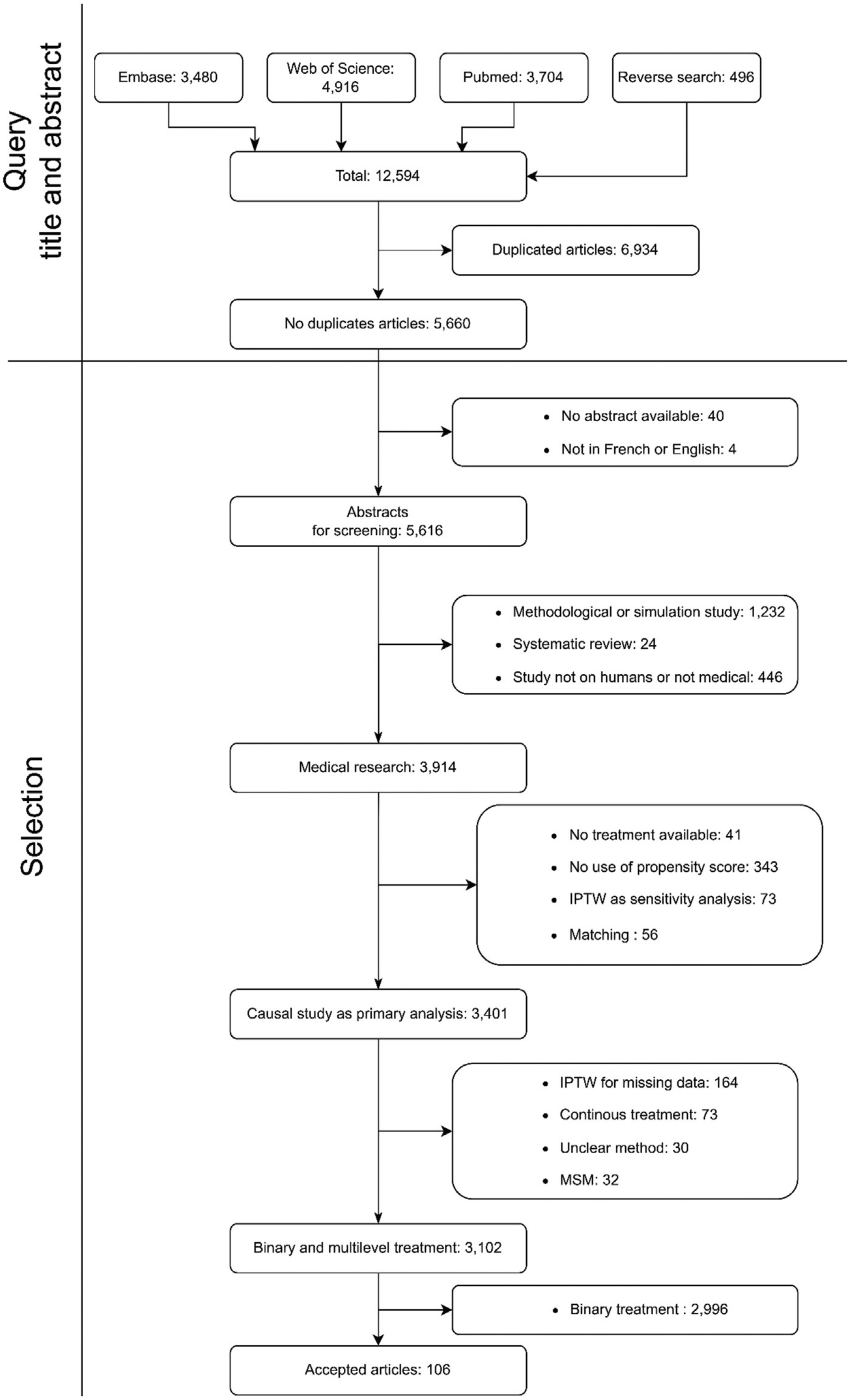
Query
title and abstract
Embase: 3,480
Web of Science: 4,916
Pubmed: 3,704
Reverse search: 496
Total: 12,594
Duplicated articles: 6,934
No duplicates articles: 5,660
Selection
No abstract available: 40
Not in French or English: 4
Abstracts for screening: 5,616
Methodological or simulation study: 1,232
Systematic review: 24
Study not on humans or not medical: 446
Medical research: 3,914
No treatment available: 41
No use of propensity score: 343
IPTW as sensitivity analysis: 73
Matching : 56
Causal study as primary analysis: 3,401
IPTW for missing data: 164
Continous treatment: 73
Unclear method: 30
MSM: 32
Binary and multilevel treatment: 3,102
Binary treatment : 2,996
Accepted articles: 106

Figure 1. Flow chart of the systematic review process. MSM: Marginal structural model, IPTW: inverse probability of treatment weight.

articles used automated selection and 3 (10.7%) used a DAG to inform the selection. In most articles (70 (66%)), the way missing covariate data was handled was not reported. In the articles which did report this, 16 (40%) used a complete case analysis, 8 (20%) used multiple imputation and 12 (30%) used other ad hoc methods.

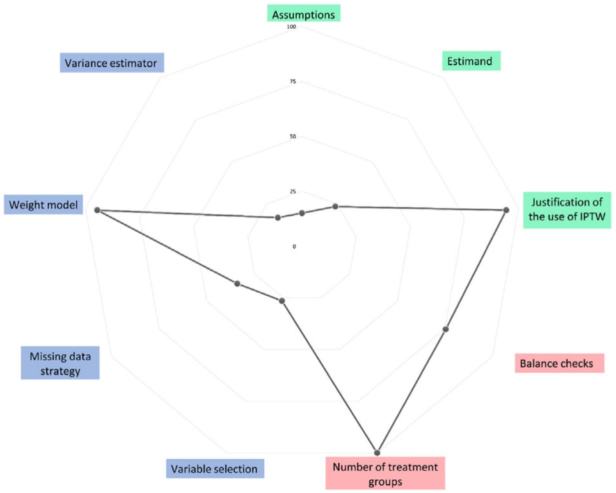

Figure 2. Summary of the main results Results are presented in percentage. In green: points related to the introduction section, in blue: points related to the methods section and in red: points related to the results.

4.6. Propensity score estimation

A majority of the studies (51 (48.1%)) used multinomial regression to estimate the weights, followed by Generalized Boosted Models (GBMs) (48 (45.3%)), and one (0.9%) used covariate balancing propensity score. The remaining 6 studies (5.7%) did not specify the method. One hundred and one (95.3%) of the articles did not present any summary or graphical representation of the weights, 4 (3.8%) articles presented a histogram of the weights and 1 (0.9%) another representation. Among the included articles, 23 (21.7%) articles explicitly stated whether or not they stabilized weights. Among these articles, 16 used (69.6%) stabilized weights. Trimming or weight truncation were mentioned in 24 (22.6%) articles. Of these 24 articles, only 18 studies actually did perform trimming or truncation and the other 6 studies just mentioned trimming or weight truncation without applying it. One potential explanation for not applying weight or trimming in these 6 papers is that the largest weights were below 10.

4.7. Assessing covariate balance

Among the reviewed articles, 26 (24.5%) did not mention whether the covariate balance after weighting was investigated. The most frequent method for estimating equilibrium was the standardized mean difference (ie, 60 articles, 56.6%), 4 (3.8%) the Kolmogorov-Smirnov distance, 12 (11.3%) used *P* values, 3 (2.8%) used graphs and 1 (0.9%) reported the population standard bias.

4.8. Analysis model

IPTW was implemented in 83 (78.3%) articles and AIPTW in 23 (21.7%). These estimators were applied to a wide range of outcomes: time-to-event 64 (60.4%), binary 20 (18.9%), continuous 10 (9.4%), count 6 (5.6%), categorical 5 (4.7%) and ordinal 1 (0.9%). The weighted outcome models used to estimate the causal effect of the treatments were diverse and depended mainly on the type of outcomes. The results are summarized in Table 3 and in Figure 2. Methods for variance estimation were reported in 18 (17%) studies, 13 (72.2%) studies used a robust estimator, 4 (22.2%) used nonparametric bootstrap and 1 used uncorrected variance.

4.9. Software and statistical packages

The three main software packages were: R 39 (37.1%), SAS 35 (33.3%) and STATA 18 (17.1%). Four (3.8%) of the articles used a combination of R, SAS and STATA and in 4 (3.8%) of the articles the statistical software was not clearly identified. Among the articles using a programming language other than R, 7 (6.6%) used R as secondary programming software for the "TWANG" package to estimate the weights using GBMs.

4.10. Good research practice

Only 8 (7.6%) studies had previously registered a protocol. Only 4 (3.8%) study reported that the code was available and 3 (2.8%) proposed an access to all or part of the data. Finally, 9 studies (8.5%) referred to the STROBE statement.

5. Discussion

This systematic review aimed to collect detailed information from published observational studies to assess how IPTW methods with a multicategory treatment are applied in medical research. As we focused the review on practical implementation in applied studies, we excluded methodological papers. From 5660 screened articles, only 106 (3.4%) focused on a multicategory treatment and the reasons for choosing IPTW over other approaches were rarely given. Moreover, the plausibility assumptions underpinning the validity of IPTW were discussed in very few studies. Overall, the quality of the reporting was poor, with key elements missing, thus compromising the interpretability and generalizability of the results.

In the majority of the studies, the estimand was not reported. This is a concern as the estimand determines the way results are interpreted. In addition, estimation of the ATT relies on less stringent assumptions.

The implementation and reporting were also often inadequate. Indeed, one of the most striking results from this review is the low frequency of studies reporting the assumptions for the identification of causal effects, and their plausibility, which is however crucial to make causal claims. This result was already observed in a previous study focusing on binary treatments [9]. Interestingly, a few studies discussed the plausibility of modeling assumptions

Table 2. Summary of the main results

Elements of IPW method		***N* (%)**
Estimand	ATE	90 (85.0%)
	ATT	5 (4.7%)
	ATO	1 (0.9%)
	Unknown	10 (9.4%)
Estimand definition	Guessed from the weights	71 (67%)
	Explicitly written	25 (23.6%)
	Not reported	10 (9.4%)
Measure of association	HR	58 (54.7%)
	OR	22 (20.8%)
	RR	8 (7.6%)
	Other	9 (8.5%)
	Unknown	2 (1.9%)
Assumptions	Mention of assumptions (yes)	16 (15.1%)
	Mention of STROBE (yes)	9 (8.5%)
Covariate selection	Variables included in the weight model	91 (85.9%)
Method used for variable selection	From the literature	14 (13.2%)
	Automated selection	11 (11.4%)
	From the DAG	3 (2.8%)
	Not specified	78 (73.6%)
Method used for missing values	Complete-case	16 (15.1%)
	Multiple imputation	8 (7.6%)
	Adjustment	1 (0.9%)
	Group mean	1 (0.9%)
	Other	9 (8.5%)
	Unknown	70 (66%)
Summary of weights	Histograms	4 (3.8%)
	Other	1 (0.9%)
	Unknown	101 (95.2%)
Weight stabilization	Yes	16 (15.1%)
	No	7 (6.6%)
	Unknown	83 (78.3%)
Model used for propensity scores	Multinomial	51 (48.1%)
	GBM	48 (45.3%)
	Other	1 (0.9%)
	Unclear	6 (5.7%)
Methods used for assessing balance	Standardized mean difference	60 (56.6%)
	KS	4 (3.8%)
	Graph	3 (2.8%)
	P values	12 (11.3%)
	Other	1 (0.9%)
	Unknown	26 (24.5%)
Analysis model	IPTW	83 (78.3%)
	AIPTW	23 (21.7%)
Variance estimation method	Robust	13 (12.3%)
	Bootstrap	4 (3.8%)
	Uncorrected	1 (0.9%)
	Unknown	88 (83%)

Table 3. Type of outcome reported and outcome models

Outcome type	Outcome model	*N* (%)
Continuous	Linear	9 (8.5%)
Binary	Negative binomial	1 (0.9%)
	Logistic	18 (17%)
Time-to-event	Cox	63 (59.4%)
	Linear	1 (0.9%)
Categorical	Multinomial	5 (4.7%)
Count	Poisson	6 (5.7%)
Ordinal	Multinomial	1 (0.9%)

(eg, proportionality of hazards), but failed to report the assumptions for causal inference.

This could be explained by a lack of practical guidelines for the reporting of these studies. Although these assumptions are not empirically verifiable, the plausibility of the assumption of no interference can be determined based on the knowledge of the clinical setting. The plausibility of the consistency assumption may be ensured with a precise definition of the exposure of interest. While the assumption of conditional exchangeability is often questionable in observational studies, the elaboration of a DAG from expert knowledge followed by an application of d-separation rules may dramatically reduce the risk of confounding. Finally, the plausibility of the positivity assumption can be explored from the distribution of the propensity score and the absence of extreme weights [24], although the absence of extreme weights does not guarantee that the positivity assumption holds."

Empirical positivity may be a challenge when estimating causal effects for more than two treatment groups as indication guidelines may be more specific when multiple treatments are available for the same condition and the sample size may be smaller in each group increasing the chance of violation of the positivity assumption in the sample. In this systematic review, one study analyzed up to 9 groups, and did not investigate violations of the positivity assumption.

The propensity score model was most often a multinomial regression model. This can probably be explained by the fact that this method is a direct extension of the logistic regression model used in IPTW for binary treatments, is simple to implement in standard statistical software, and relatively inexpensive in terms of computational power. GBMs, a machine learning method based on regression trees, are featured prominently in this literature review. Although these methods are more computationally expensive, there is an easy-to-use implementation in the "TWANG" package [30] with a tutorial for causal effect estimation in the case of multicategory treatments [14]. An advantage of GBMs is that they are nonparametric and therefore do not require the specification of a functional form. Furthermore, in the TWANG implementation, the stopping rule for the GBM algorithm is based on a balance

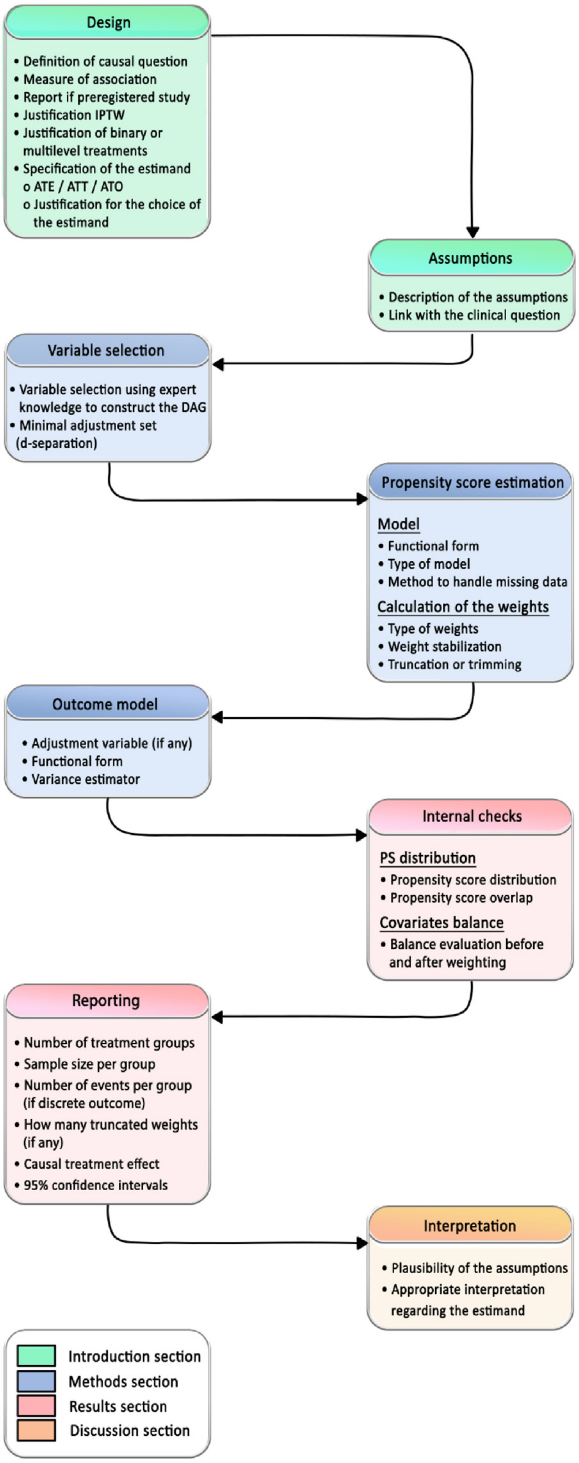
Design
• Definition of causal question
• Measure of association
• Report if preregistered study
• Justification IPTW
• Justification of binary or multilevel treatments
• Specification of the estimand
o ATE / ATT / ATO
o Justification for the choice of the estimand
Assumptions
• Description of the assumptions
• Link with the clinical question
Variable selection
• Variable selection using expert knowledge to construct the DAG
• Minimal adjustment set (d-separation)
Propensity score estimation
Model
• Functional form
• Type of model
• Method to handle missing data
Calculation of the weights
• Type of weights
• Weight stabilization
• Truncation or trimming
Outcome model
• Adjustment variable (if any)
• Functional form
• Variance estimator
Internal checks
PS distribution
• Propensity score distribution
• Propensity score overlap
Covariates balance
• Balance evaluation before and after weighting
Reporting
• Number of treatment groups
• Sample size per group
• Number of events per group (if discrete outcome)
• How many truncated weights (if any)
• Causal treatment effect
• 95% confidence intervals
Interpretation
• Plausibility of the assumptions
• Appropriate interpretation regarding the estimand
🟩 Introduction section
🟦 Methods section
🟥 Results section
🟧 Discussion section

Figure 3. Guideline for causal inference approaches This guideline proposes a list of the main points to follow regarding the introduction (green windows), methods (blue windows), results (red windows) and discussion (red windows) sections to report a study based on weighted approaches in general and using multilevel treatment in particular. IPTW: inverse probability of treatment weight, ATE: Average treatment effect, ATT: Average treatment effects on the treated, ATO: average treatment effect in the overlapping population, DAG: directed acyclic graph, PS: propensity score.

metrics for the covariates, thus maximizing the balance across treatment groups. However, the method to compute the weights was not always reported, which compromises the transparency and reproducibility of the results.

The validity of propensity score methods depends on the ability of the scores to balance treatment groups with respect to the covariates [24]. In our review, balance was assessed in most studies, but a few used *P* values, that are not recommended because they strongly depend on the sample size. Balance should be assessed before and after weighting for instance by presenting standardized mean differences for each covariate and for each pair of treatments or by presenting the mean or maximum standardized mean difference per variable across all treatment comparisons. However, there is currently no consensus on the way to assess balance for multiple treatments and further work is needed to provide practical guidelines.

In terms of analysis model, the type of model was generally well reported, but not the variance estimator. This is very important because the estimated variance must account for (i) the correlation introduced via weighting (ii) the uncertainty around the propensity score estimates. In practice, many authors used sandwich estimators for (i) but issue (ii) is often overlooked, despite available estimators [27] including for multicategory treatments [29] and the validity of nonparametric bootstrap.

Guidelines for the application of IPTW for binary treatments exist [9,31], and we would like to propose steps for their reporting in the case of multicategory treatments. These recommendations are summarized in Figure 3.

6. Conclusion

Causal inference approaches using IPTW are largely applied in medical research but multicategory treatment remains scarcely used. This systematic review highlighted the suboptimal reporting quality of studies in this context, in particular for assumptions and model building. The application of practical guidelines, as proposed here, is needed to help researchers improve the presentation of their results to ensure a better understanding of their methods and the reproducibility of their results.

CRediT authorship contribution statement

François Bettega: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Investigation, Conceptualization, Data curation, Formal analysis. **Monique Mendelson:** Writing – review & editing, Writing – original draft, Investigation, Conceptualization. **Clémence Leyrat:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Sébastien Bailly:** Conceptualization, Data curation, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Data availability

Data will be made available on request.

Declaration of competing interest

None.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111338.

References

[1] Senn S. Seven myths of randomisation in clinical trials. Stat Med 2012;32:1439–50.

[2] Ware JH, Hamel MB. Pragmatic trials — guides to better patient care? N Engl J Med 2011;364:1685–7.

[3] Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878–86.

[4] Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials 2015;16(1).

[5] Hernan MA. A definition of causal effect for epidemiological research. J Epidemiol Ampmathsemicolon Community Health 2004;58(4):265–71.

[6] Pawlowski C, Lenehan P, Puranik A, Agarwal V, Venkatakrishnan AJ, Niesen MJM, et al. FDA-authorized mRNA COVID-19 vaccines are effective per real-world evidence synthesized across a multi-state health system. Med 2021;2(8):979–992.e8.

[7] Smith MJ, Mansournia MA, Maringe C, Zivich PN, Cole SR, Leyrat C, et al. Introduction to computational causal inference using reproducible Stata, R, and Python code: a tutorial. Stat Med 2021;41:407–32.

[8] ROSENBAUM PR, RUBIN DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41–55.

[9] 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;34:3661–79.

[10] Ali MS, Groenwold RHH, Belitser SV, Pestman WR, Hoes AW, Roes KCB, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol 2015;68:122–31.

[11] Bettega F, Leyrat C, Tamisier R, Mendelson M, Grillet Y, Sapène M, et al. Application of inverse-probability-of-treatment weighting to estimate the effect of daytime sleepiness in patients with obstructive sleep apnea. Ann Am Thorac Soc 2022;19(9):1570–80.

[12] Carr DC, Willis R, Kail BL, Carstensen LL. Alternative retirement paths and cognitive performance: exploring the role of preretirement job complexity. Gerontologist 2019;60(3):460–71.

[13] Rannanheimo PK, Tiittanen P, Hartikainen J, Helin-Salmivaara A, Huupponen R, Vahtera J, et al. Impact of statin adherence on cardiovascular morbidity and all-cause mortality in the primary prevention of cardiovascular disease: a population-based cohort study in Finland. Value Health 2015;18(6):896–905.

[14] McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med 2013;32:3388−414.

[15] Lopez MJ, Gutman R. Estimation of causal effects with multiple treatments: a review and new ideas. Stat Sci 2017;32(3):432−54.

[16] Pearl J. An introduction to causal inference. Int J Biostat 2010;6(2):7.

[17] Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. Int J Epidemiol 2016;45:1776−86.

[18] Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550−60.

[19] Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol 2008;168:656−64.

[20] Imbens G. The role of the propensity score in estimating dose-response functions. Biometrika 2000;87(3):706−10.

[21] Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction: Data Mining, Inference, and Prediction. New York: Springer-Verlag; 2009:745.

[22] Hastie T, Tibshirani R, Friedman J. Boosting and additive trees. In: The Elements of Statistical Learning. New York, NY: Springer New York; 2009:337−87. [Springer Series in Statistics)].

[23] Greifer N, Stuart EA. Choosing the causal estimand for propensity score analysis of observational studies 2021: arXiv preprint arXiv: 2106.10577.

[24] Hernán MA, Robins JM. Causal inference: what if. Boca Raton, FL: Chapman & Hall/CRC; 2019:352.

[25] Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. Am J Epidemiol 2011;173:761−7.

[26] Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika 2009; 96(1):187−99.

[27] Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. Stat Med 2013;33:721−37.

[28] Yoshida K, Solomon DH, Haneuse S, Kim SC, Patorno E, Tedeschi SK, et al. Multinomial extension of propensity score trimming methods: a simulation study. Am J Epidemiol 2018;188:609−16.

[29] Li F, Li F. Propensity score weighting for causal inference with multiple treatments. Ann Appl Stat 2019;13(4):2389−415.

[30] Griffin BA, Ridgeway G, Morral AR, Burgette LF, Martin C, Almirall D, et al. Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG). Santa Monica, CA: RAND; 2014. Available at: http://www.rand.org/statistics/twang. Accessed April 12, 2024.

[31] Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naïve enthusiasm to intuitive understanding. Stat Methods Med Res 2011;21(3):273−93.