Research Article

Gabriel Danelian, Yohann Foucher, Maxime Léger, Florent Le Borgne, and Arthur Chatton*

Identification of in-sample positivity violations using regression trees: The PoRT algorithm

https://doi.org/10.1515/jci-2022-0032
received April 29, 2022; accepted May 20, 2023

Abstract

Background – The positivity assumption is crucial when drawing causal inferences from observational studies, but it is often overlooked in practice. A violation of positivity occurs when the sample contains a subgroup of individuals with an extreme relative frequency of experiencing one of the levels of exposure. To correctly estimate the causal effect, we must identify such individuals. For this purpose, we suggest a regression tree-based algorithm.

Development – Based on a succession of regression trees, the algorithm searches for combinations of covariate levels that result in subgroups of individuals with a low (un)exposed relative frequency.

Application – We applied the algorithm by reanalyzing four recently published medical studies. We identified the two violations of the positivity reported by the authors. In addition, we identified ten subgroups with a suspicion of violation.

Conclusions – The PoRT algorithm helps to detect in-sample positivity violations in causal studies. We implemented the algorithm in the R package RISCA to facilitate its use.

Keywords: causal inference, common support, decision tree, experimental treatment assignment, identifiability, target population

MSC2020: 62-04, 68W40, 92C60

1 Introduction

The positivity assumption, also known as experimental treatment assignment or common support, is a cornerstone for drawing causal inferences. The positivity assumption means that, theoretically, each individual has a nonzero probability of being exposed and unexposed [1,2]. In practice, however, small but nonzero exposure frequencies can be observed in the sample. For instance, individuals with a contraindication to a specific treatment should theoretically never receive it but might, in practice, be prescribed it as a last resort. Thereby, we define the positivity assumption, at a sample level, as a nonextreme probability to experience

*** Corresponding author: Arthur Chatton,** IDBC/A2COM, Pacé, France; UMR INSERM 1246 – SPHERE, Université de Nantes, Université de Tours, Nantes, France; Faculté de pharmacie, Université de Montréal, Montréal, QC, Canada, e-mail: arthur.chatton@umontreal.ca
Gabriel Danelian: Université de Lille, Lille, France; IDBC/A2COM, Pacé, France
Yohann Foucher: UMR INSERM 1246 – SPHERE, Université de Nantes, Université de Tours, Nantes, France; Centre Hospitalier Universitaire de Nantes, Nantes, France; Centre d'Investigation Clinique CIC 1402, INSERM, Université de Poitiers, CHU Poitiers, Poitiers, France
Maxime Léger: UMR INSERM 1246 – SPHERE, Université de Nantes, Université de Tours, Nantes, France; Département d'anesthésie-réanimation, Centre Hospitalier Universitaire d'Angers, Angers, France
Florent Le Borgne: IDBC/A2COM, Pacé, France; UMR INSERM 1246 – SPHERE, Université de Nantes, Université de Tours, Nantes, France

each exposure's value (i.e., bounded at a study-specific threshold), formally $\beta < P(A \mid C) < 1\text{-}\beta$, where A is the exposure, C is the adjustment set, and β is the specified threshold [3].

In-sample violations of positivity can be addressed via statistical procedures, for instance, by trimming or truncating propensity scores (PS) [4]. Alternatively, specific weighting systems or decision trees enable us to estimate a causal effect in a subpopulation respecting positivity [5–8]. However, these approaches change the characteristics of the sample and therefore shift the targeted population [9,10], compromising the external validity of the results [11]. Alternatively, one can shift toward an estimand that requires a weaker positivity assumption, such as the average causal effect on the exposed [12]. But, again, this changes the target population and thus answers a different question. Platt et al. argued that the exclusion of subjects unlikely to be (un) exposed, by redefining the eligibility criteria, leads to clear statements about the generalizability of the results [13]. This approach is closely related to a well-defined target trial [14,15]. Traskin and Small proposed a tree-based algorithm to redefine the study population [16]. More recently, Karavani et al. suggested, in an unpublished article, a similar approach based on random forest [17]. Both propositions learn the trees with all the predictors at once. However, the predictors have varying degrees of influence on the tree's construction, the most important being those the algorithm considers most apt at dividing the population in an informative manner (i.e., optimizing a loss/cost function). These variables do not always correspond to the most important for the stakeholders (from the perspective of the interpretation of the positivity violations). Therefore, using all covariates to build a tree might not be the most effective manner to detect such violations.

This article proposes an explanatory algorithm based on a succession of regression trees (RTs) to identify observed subgroups of individuals potentially affected by such violations. Since an RT consists of a set of nodes (i.e., binary decision rules), themselves subdivided into other nodes (and so on, see Figure S1), using the exposure as the outcome and one or several covariates as predictors enables one to obtain an estimate of the exposure probability (i.e., the relative frequency) in each subgroup represented by the nodes [18]. In the presence of a node with an exposure's extreme relative frequency, we would therefore suspect a violation of the positivity assumption in the corresponding subgroup. To facilitate the use of this method, we implemented it as a function in the R package RISCA [19].

The rest of this article proceeds as follows. We present the positivity-RT (PoRT) algorithm in the next section. Then, we reanalyze several datasets from recently published studies to illustrate its usefulness. Even if we focus on medical studies, PoRT's applicability is transportable to other domains such as economics or social sciences. The last section offers discussions and practical recommendations.

2 The Po(sitivity)-RT algorithm

2.1 A conceptual overview

The PoRT algorithm looks for combinations of a small number of predictors, where the exposed and unexposed samples are highly imbalanced (almost all the observations belong to one of the two groups), and the overall size of the subgroup is sufficiently high (Figure 1). It involves a two-step procedure. First, PoRT learns a succession of RTs according to different combinations of predictors. Second, it identifies imbalanced subgroups using two reading hyperparameters.

2.2 How are the RTs learned?

The algorithm learned the RTs using the R package *rpart* to predict the exposure on different combinations of predictors [20]. We define the first hyperparameter γ, which refers to the maximum number of predictors used to define the nodes. We stress that the predictors set must match the variables used for the adjustment since

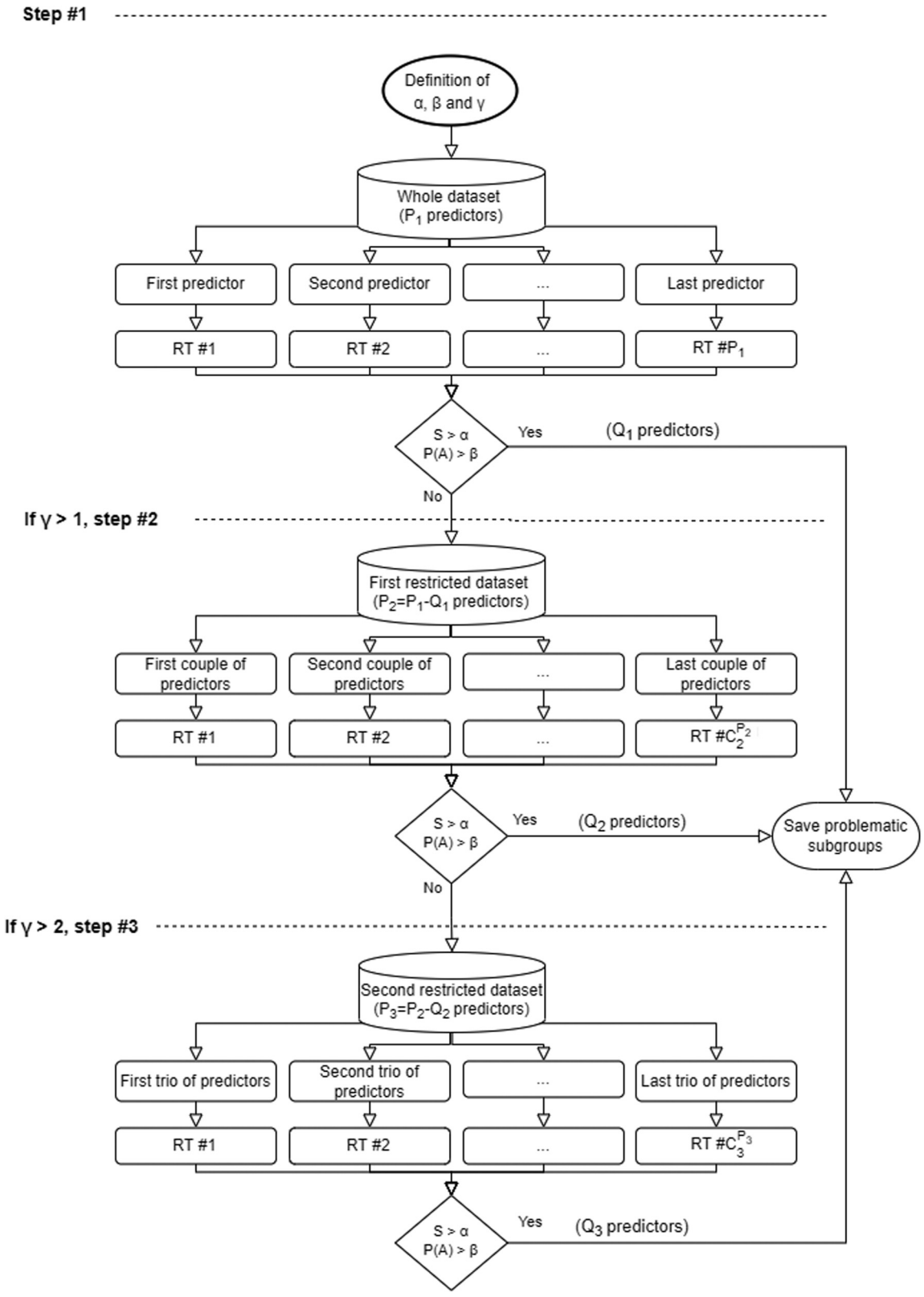

Figure 1: Flowchart presenting the PoRT algorithm. *P*(*A*): Exposed and unexposed proportions; *C*: Binomial coefficient representing the maximal number of RTs for the step; *S*: Proportion of the whole sample contained in the subgroup; *α*, *β*, and *γ* were user-supplied hyperparameters.

positivity is only relevant for them [2]. In contrast, including pure predictors of the exposure in the adjustment set increases the risk of positivity violations. For instance, if $\gamma = 1$, each RT predicted the exposure on one different predictor. If $\gamma = 2$, each RT predicted the exposure on one different predictor and then on a different combination of two predictors, and so on (Figure 1). Three other hyperparameters, specific to *rpart*, must be defined. First, the minimal number of individuals in a node to make it split. Second, the minimal number of individuals in the leaves for the parent node to be split. Third, the maximum depth of the tree consists of the maximum number of successive splits. As in *rpart*, we refer to these hyperparameters as *minsplit*, *minbucket*, and *maxdepth*. Figure S2 illustrates the impact of these hyperparameters on the RT's learning. The rules of splitting are nicely summarized in the study by Kang et al. [8].

2.3 How are the RTs read?

Consider (α, β) as the two user-defined parameters to define the positivity violation. Parameter α is the minimal proportion of the whole sample size to consider a problematic subgroup. Parameter β is the exposed or unexposed proportion under which one can consider an in-sample positivity violation [3]. A node will be considered problematic if the proportion of being (un)exposed is more extreme than β or $1 - \beta$ and if it contains at least α individuals of the whole sample (see Figure 2 for an illustration). Note that, once a node is flagged, its descendants are not considered to avoid uninformative nested subgroups (as shown in Figure 2b).

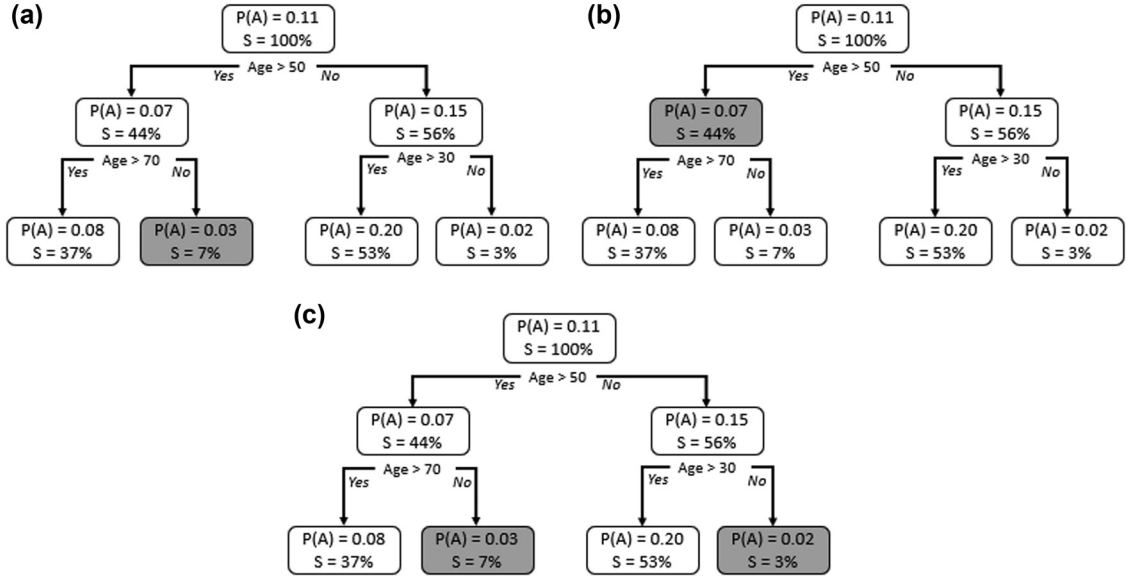

Figure 2: Impact of the hyperparameters α and β on the problematic subgroups' identification for a hypothetic tree. (a) $\alpha = 0.05$ and $\beta = 0.05$, (b) $\alpha = 0.05$ and $\beta = 0.10$, (c) $\alpha = 0.01$ and $\beta = 0.05$. The gray nodes correspond to subgroups violating the positivity assumption according to the parameters α and β. *P(A)*: Exposed or unexposed proportions, *S*: Proportion of the whole sample contained in the subgroup.

2.4 The algorithm definition

On the basis of hyperparameters previously defined, we used RTs to identify problematic subgroups. Figure 1 presents an overview of the proposed algorithm. The first step consists of learning one tree for each predictor and memorizing the nodes corresponding to problematic subgroups (which can be leaves or intermediate

nodes). If $\gamma = 1$, the algorithm stops. Otherwise, if at least one problematic subgroup is identified in the first step, the corresponding predictor(s) is(are) not considered in the next step, which estimates one tree for all possible couples of remaining predictors and memorizes the nodes corresponding to problematic subgroups. If $\gamma = 2$, the algorithm stops; otherwise, the third step consists of building one tree for all possible trios of remaining covariates not involved in the previously identified subgroups, and so on, until one reaches the value of γ.

2.5 The default values of the hyperparameters

To the best of our knowledge, there is no consensus on the precise definition of a subgroup presenting a positivity violation. Nevertheless, one can set $\alpha = 5\%$ as several authors suggest trimming the weights at this level in inverse probability weighting analyses [21–23]. Following D'Amour et al., one can also set $\beta = 5\%$ [3]. Note that this value is consistent with the propensity score literature [4,24]. By default, the last parameter is set to $\gamma = 2$ as positivity violations defined by three or more variables may not represent real practices. However, this could change depending on the field of study. Note that, to build the RTs, we used the default values defined in the R package *rpart* (version 4.1–15): at least 20 individuals for intermediate nodes, at least six individuals for leaves, and a maximum depth of 30 [20]. The trees were then not pruned to obtain the largest number of divisions.

The α hyperparameter could be increased as the sample size decreases. Indeed, random violations are more likely to occur in such situations. Similarly, alternative values can be considered for the β hyperparameter. For instance, Petersen et al. [4] and Crump et al. [25] argued in favor of $\beta = 1$ and 10%, respectively. Note that exposure prevalence can also influence this choice, with a small prevalence leading to a tighter definition of positivity and thus a smaller value of β. The parameter γ may also be increased with the sample size to refine the definition of subgroups. The *rpart*'s hyperparameters for splitting were set to small values to obtain the largest number of possible divisions. This limits the risk of missing some problematic subgroups due to a too restrictive splitting, but increases the risk of false positive results.

3 Applications

We reanalyzed four studies using the PoRT algorithm. The studies were all based on an inverse probability weighting approach [26–29]. We aimed to validate the algorithm's capacity to reidentify the positivity violations previously reported based on expert knowledge and to identify potential new violations. Therefore, we applied the PoRT algorithm to the entire cohort and used the adjustment set used by the authors as predictors (Table S1). We conducted complete case analyses. We categorized each continuous variable according to clinically relevant cutoffs before running the algorithm to avoid clinically insignificant subgroups (Table S1). In the case of new problematic subgroups, we (i) reanalyzed the data by excluding the subgroups, defined as plausible by an expert of the real decision process that generated the data, to evaluate the impact of shifting the targeted population, and (ii) compared the estimated PS of the individuals identified by PoRT to those excluded by the authors. We performed the main analysis with the default values ($\alpha = 5\%$, $\beta = 5\%$, and $\gamma = 2$), and we considered other values ($\alpha = 1$ or 10%, $\beta = 1$ or 10%) in sensitivity analyses. In a second experiment, we illustrated the importance of the γ hyperparameter by comparing the results with an RT using all predictors at once. In the last experiment, we evaluated the results of the PoRT algorithm without a prior categorization of the continuous predictors. We performed all analyses using R version 3.6.0 [30].

3.1 Barbiturates during intensive care for patients with traumatic brain injury

Léger et al. investigated the impact of barbiturates on mortality using a cohort composed of 1,088 patients admitted to intensive care units for traumatic brain injury [26]. This treatment may be offered to reduce

patients' intracranial hypertension (IH) and its consequences in terms of brain damage. The authors excluded individuals older than 70 years, for whom therapy was contraindicated. We identified nine subgroups potentially associated with in-sample positivity violations (Table 1). First, the PoRT algorithm confirmed patients older than 75 years as being problematic, a higher threshold than that suggested by the authors. Second, it indicated that patients without osmotherapy at admission had a relative frequency of barbiturates lower than 5%. This seems clinically relevant since barbiturates are a last-line therapy and should be proposed only after osmotherapy. Third, PoRT identified four subgroups composed of patients without IH at admission. Indeed, barbiturates are rarely proposed from a preventive perspective for high-risk patients. Note that this characteristic was not identified in the first iteration of PoRT because the barbiturates' relative frequency for patients without IH was 5.1%, only slightly above the 5% threshold. Among the three major violations with respect to the sample size, we found a substantial overlap between the patients (n = 534). They represent a broader subpopulation that should not be targeted in the study. Fourth, the algorithm detected that a lactatemia lower than 1 mmol/L, a simplified acute physiology score (SAPS) II score from 40 to 44 without severe trauma, and a SAPS II score from 25 to 55 along with a creatinine level between 50 and 60 mmol/L were associated with a relative frequency of barbiturates below 5%.

Figure 3 describes the results of the initial analysis performed by the authors and our reanalysis based on the restricted sample without the three previous subgroups associated with structural violations. In the initial study, the odds ratio (OR) was 2.2 (95% CI from 1.1 to 4.6). The exclusion of individuals without IH at admission reduced the sample size by almost 70%. The main impact was on the control group: the sample size decreased from 964 to 100 patients in this group versus 124 to 73 in the group under barbiturates. In the restricted sample (N = 173), the OR was 1.9 (95% CI from 1.0 to 3.5). The PSs for the discarded individuals ranged from 0.006 to 0.327, and more than 25% of these individuals presented a PS greater than the threshold of 0.05. In contrast, no individuals with a PS higher than 0.95 or lower than 0.05 (55.6% of the whole dataset) remained in the restricted dataset (i.e., after removing the problematic subgroups).

3.2 Kidney transplantations from marginal donors

Querard et al. compared grafts from standard and marginal donors, as defined by the expanded donor criteria [27]. Because of the shortage of kidneys for transplantation, kidney grafts harvested from standard donors are preferentially attributed to young recipients, resulting in a susceptible positivity issue. The authors did not consider recipient age among the eligibility criteria. In contrast, PoRT detected a problematic subgroup consisting of 391 recipients younger than 30 years (8.1% of the whole sample) with a marginal graft's relative frequency of 3.1% (Table S2).

The exclusion of these individuals did not change the effect size (Figure 3). In the restricted sample (N = 4,442), the hazard ratio (HR) was 1.3 (95% CI from 1.2 to 1.5). In the initial study (N = 4,833), the HR was 1.3 (95% CI from 1.1 to 1.6). The PSs of the discarded individuals ranged from 0.002 to 0.024, while 16.8% had an extreme PS, and 8.8% remained after discarding the problematic subgroups.

3.3 Hypothermic perfusion machine for marginal donors

Foucher et al. compared the use of a hypothermic perfusion machine to static cold storage in kidney transplantations from expanded donors [28]. The authors reduced the studied cohort to individuals older than 45 years because of a potential structural violation: the old-to-old graft allocation policy results in a lower susceptibility of younger candidates receiving marginal grafts. The PoRT algorithm did not detect this issue when using the entire cohort with no restriction on patient age (N = 1,978). Indeed, 32 (3.3%) and 44 (4.0%) individuals were younger than 45 years in the perfusion machine and cold storage groups, respectively. In contrast, PoRT suggested two problematic subgroups. First, individuals receiving transplants before 2014 in

Table 1: Subgroups of patients identified by the PoRT algorithm as potential sources of nonpositivity

Authors	Sample size	Problematic subgroup (*n*, %)	Identified by the authors	Clinically plausible
Léger et al. [26]	1,088	Age ≥ 75 years (73, 6.7)	Yes	Yes
		No osmotherapy at admission (732, 67.3)	No	Yes
		25 ≤ SAPS II score < 55 & 50 ≤ Creatinine < 60 (135, 12.4)	No	No
		No IH at admission nor history of head trauma (710, 65.3)	No	Yes
		No IH & severe trauma at admission (385, 35.4)	No	Yes
		No IH at admission & Creatinine <150 (740, 68.0)	No	Yes
		No IH at admission & SAPS II score <55 (532, 48.9)	No	Yes
		Lactatemia <1 (197, 18.1)	No	No
		40 ≤ SAPS II score <45 & no severe trauma (69, 6.3)	No	No
Querard et al. [27]	4,442	Recipient age <30 years (391, 8.1)	Yes	Yes
Foucher et al. [28]	1,978	Transplant before 2014 & A – D centers[a] (376, 19.0)	No	Yes
		Transplant before 2012 & CIT ≥20 h (101, 5.1)	No	Yes
Masset et al. [29]	383	None detected[b]	–	–

Abbreviations: BMI, body mass index; CIT, cold ischemia time and IH, intracranial hypertension. Creatinine and lactatemia were in mmol/L.
[a]The centers were anonymized. [b]Regardless of the sample and the adjustment set.

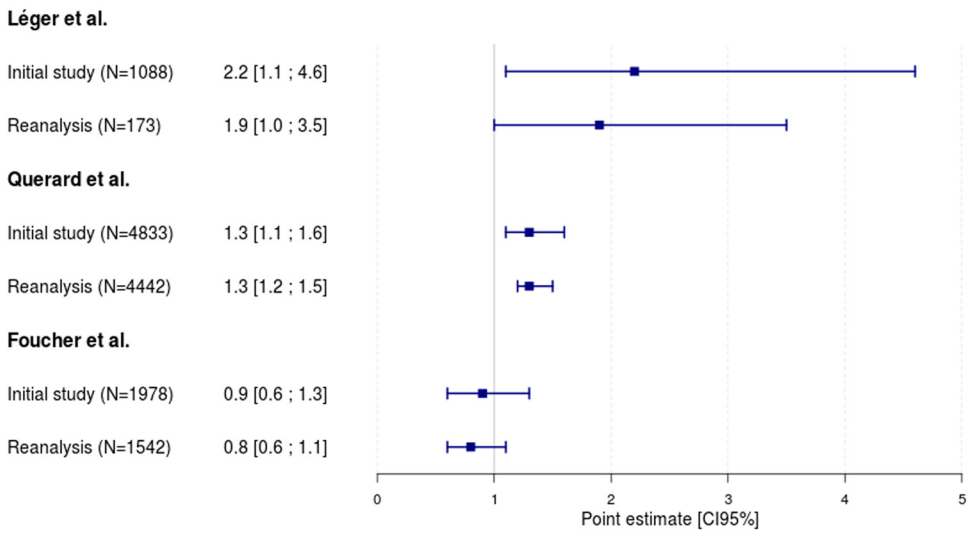

Figure 3: Comparisons of the results reported in the initial studies and those obtained by reanalyzing the data by excluding the patients associated with a structural violation of positivity. Léger et al. presented an OR, while the others presented a HR.

four centers had a prevalence of perfusion machines ranging from 0.8 to 4.7%. This violation appeared plausible because some hospitals were slow to adopt hypothermic perfusion machines, which were only introduced in 2010 in France. Second, patients with a cold ischemia time under 20 h who received a transplant before 2013 (N = 101) received a graft under static cold storage with a relative frequency lower than 5%, which also appeared plausible because the machines were preferentially attributed to situations with a long transport from donors to recipients. This preferential attribution became less strict with the increase in the availability of machines over time.

The exclusion of these two problematic subgroups with the inclusion of all recipients regardless of age shifted the HR from 0.9 (95% CI from 0.6 to 1.3) in the initial study to 0.8 (95% CI from 0.6 to 1.1) in the novel targeted population. The PSs of the discarded individuals ranged from 0.606 to 0.993, and more than 50% of these individuals presented a PS lower than the threshold of 0.95. In contrast, 11.9% of the individuals have an extreme PS and 4.6% remain after discarding the problematic subgroups.

3.4 Induction therapy in elderly kidney transplant recipients

Masset et al. compared the risk of several adverse events following a kidney transplant in elderly recipients depending on their induction therapy: antithymocyte globulins versus basiliximab [29]. We did not detect any positivity violations with PoRT, regardless of the event and adjustment set, confirming the authors' statements.

3.5 Sensitivity analyses: Impact of the α and β hyperparameters variation

As expected, lowering the value of parameter β decreased the number of subgroups identified (β = 1%, Table S3). In the study by Léger et al. [26], the problematic subgroups involving patient age and osmotherapy were no longer identified. Nevertheless, combinations of age, osmotherapy, and others were present, similar to the subgroups involving the IH in the main analysis, alerting against broader violations undetected with the current parameters. The subgroups identified in the studies by Querard et al. [27] and Foucher et al. [28] were also no longer identified due to the stringent definition of positivity.

With a broader definition (β = 10%, Table S4), we identified 11 new subgroups across the four studies. While their majority did not seem to have a plausible clinical interpretation, we identified the aforementioned subgroup of individuals without IH in the study by Léger et al. [26]. Importantly, such a choice can lead to the identification of broader subgroups for continuous variables, possibly incorrectly. For instance, we identified two subgroups in the study by Querard et al. [27] for recipients younger than 50 years and older than or equal to 65 years. These thresholds appear less plausible than that of 30 years.

We identified 6, 14, and 4 subgroups, leading to in-sample violations with parameter α at 1% in the studies by Léger et al. [26], Foucher et al. [28], and Masset et al. [29] (Table S5). Interestingly, in the study by Foucher et al. [28], recipient's age lower than 50 years was identified as problematic when associated with transplantation before 2013. This result might explain the exclusion criteria used by these authors.

Unexpectedly, a higher α did not always reduce the number of subgroups identified (Table S6). Four new subgroups were identified in the study by Léger et al. [26], notably for combinations including age. This finding can be explained by the fact that PoRT cannot identify the absence of IH with a β at 5%, but it could when β is set at 1%. The threshold for age at 5% was 75 years, which represents 6.7% of the whole sample. Because PoRT could not find this subgroup by searching those including at least 10% of the whole sample, it identified several other subgroups defined by age. This again alerts against an underlying broader violation undetected with the current parameters.

3.6 Comparison with an RT using all predictors at once

When using an RT with all the predictors at once, only one node was identified as problematic in the study by Léger et al. [26]. It corresponded to the absence of osmotherapy at inclusion (Table S7). For the study by Foucher et al. [28], this approach led to one problematic subgroup, also identified by our approach. The second subgroup identified by our approach remained unidentified. In the studies by Querard et al. [27] and Masset et al. [29], no subgroup was identified as problematic: the two approaches resulted in the same conclusions. Trees with the entire set of predictors were presented in Figures S3–S5.

The tree for the first study (Figure S3) presented several additional (potentially finer-grained) subgroups, like the individuals without osmotherapy nor IH at inclusion (or all descendants of this subgroup but one). However, the other main subgroup without IH at inclusion (with osmotherapy) is not identified. Since the target population was the individuals with IH, individuals without IH must be identified regardless of their osmotherapy status. By building only one tree with the entire adjustment set, one becomes highly dependent on the first splits, which may not be the most informative for the positivity. Moreover, considering the descendants of a previously flagged node increases the number of false positives and complicates the interpretation of the results.

3.7 Impact of a prior continuous categorization of predictors

Without prior categorization, PoRT detected seven problematic subgroups in the study by Léger et al. [26], including three considered clinically plausible (Table S8). In contrast, categorizing continuous predictors using clinically meaningful thresholds led to the identification of nine problematic subgroups, including six plausible according to domain experts (Table 1). In the study by Querard et al. [27], the subgroup identified in the main experiment was flagged again but with a threshold higher of 4 years. In the study of Foucher et al. [28], only one subgroup was identified out of the two in the main experiment. The missing subgroup involved the cold ischemia time, a variable challenging to split algorithmically due to its distribution. In the study by Masset et al. [29], PoRT identified three new problematic subgroups across the seven different datasets. However, these subgroups involve variables categorized automatically at uninterpretable thresholds. Therefore,

categorizing continuous predictors reduced the risk of false positives by identifying subgroups more coherent with clinical practice.

4 Discussion

The proposed PoRT algorithm uses three hyperparameters to identify subgroups affected by a positivity violation in the sample: α is the minimum size of the subgroup expressed as a percentage of the whole sample size, β is the maximum probability of being exposed or unexposed in the subgroup, and γ is the maximum number of predictors for the discrimination of subgroups. PoRT builds RTs using the classification and regression trees (CART) algorithm implemented in the *rpart* R package with its own hyperparameters [20]. CART has the advantage of providing an exposure relative frequency in each node, unlike alternative RT algorithms, such as C4.5 [18]. Other packages implementing CART could be used in PoRT, but the hyperparameters may differ.

In a reanalysis of data from four published studies, the PoRT algorithm confirmed two subgroups with a plausible violation, as stated by the authors [26,27]. It also confirmed the absence of problematic subgroups in one study [29]. Furthermore, PoRT identified nine new problematic subgroups, including six judged plausible by domain experts. These subgroups were initially missed by the authors but were validated by the same persons in light of these reanalyses. PoRT also invalidated the a priori choice in one study to restrict the population [28]. Note that making an accurate description of the problematic subgroups and their potential overlap (which may represent an even more problematic subpopulation) can help the investigators better define the eligibility criteria and, thus, a better-suited target population. For instance, the application to the dataset of Léger et al. [26] highlighted the advantage of increasing the level of γ to identify structural violations when β is close to the exposure prevalence. Previously, we emphasized that the predictors must match the intended adjustment set. Nevertheless, we can consider as additional predictors the center when a center effect is expected [31], and the calendar time when the individuals are included over an extended period. As shown in the application to the dataset of Foucher et al. [28], the probability of experiencing the exposure can change over time with the evolution of practices.

In addition to these concordant and discordant results illustrating the helpful support offered by the PoRT algorithm, we reanalyzed the data in consideration of these alternative targeted populations. Although the clinical conclusions were unaffected, we reported nonnegligible differences in the effect sizes, underlying the importance of the targeted population. Note that we used the same data to run PoRT and the PS analysis, and the width of the confidence interval might be considered with caution.

Other approaches can be employed for diagnosing positivity violations. First, as previously mentioned, Traskin and Small [16] and Karavani et al. [17] proposed tree-based approaches using all predictors simultaneously. Such approaches may suggest fewer problematic subgroups. For instance, only the individuals without osmotherapy at admission were identified by this approach in the study by Léger et al. [26], while both osmotherapy and IH statuses should be specified in the eligibility criteria. Second, Westreich and Cole recommended computing all possible contingency tables to check for the presence of empty cells [32]. Cole and Hernán warned against the presence of extreme weights when a propensity score is used [33]. Petersen et al. provided a bootstrap-based tool to quantify the amount of bias due to such violations [4]. In contrast to PoRT, the first approach is feasible only in lower-dimensional settings, while the two last approaches cannot easily identify the subgroups causing a positivity violation. Indeed, finding the problematic subgroups from the propensity score requires a recursive identification of nonoverlapping regions from the distribution of each adjustment variable. Furthermore, identifying subgroups defined by two or more variables is challenging. In our applications, PoRT suggested removing some individuals with nonextreme weights and keeping other individuals with extreme weights. Nevertheless, these weights are computed using logistic regressions, while PoRT has the advantage of not requiring parametric assumptions. Third, King and Zeng identify an overlapping multivariate space by a convex hull and suggest removing individuals outside this space [34]. Similarly, Fogarty et al. proposed an algorithm constructing a hyperrectangle around an identified overlapping multivariate space [35]. However, these approaches fail to identify subgroups inside this space [36]. In contrast,

Oberst et al. proposed an algorithm that suggests rules to redefine the target population, albeit the rules seem unnecessarily complex [36]. Comparing this last approach to PoRT on synthetic and real datasets is challenging and needs further work [37].

We focused on in-sample violations, which can appear for two reasons. First, due to the inclusion of individuals who theoretically could never receive one of the exposure values at the population level, one then speaks of structural violations. This incorrect inclusion of individuals might be due to some practices' heterogeneity (e.g., physicians not following the most recent standard of care guidelines). Second, one speaks of random violations because of sample-to-sample fluctuations [1,2]. Such a random violation is therefore more likely for small samples or with rare exposure. Because PoRT is a descriptive algorithm, it is agnostic about whether the lack of positivity is random or structural. It must be pipelined with human domain knowledge to discriminate random and structural violations and a causal estimator to draw inferences. PoRT aims to identify the observed lack of positivity in a sample without inference at a population level. Such generalizations may be made according to two complementary arguments. The first one is to gauge the random violation (i.e., sample-to-sample fluctuation) by statistical tests. However, such procedures will be associated with overfitting due to using the same sample to identify the problematic subgroups and test their significance. The correction of the type I error is therefore needed. The second one calls for domain knowledge to find an underlying cause for considering a violation at a population level (i.e., structural violation). Regardless of the qualification of the observed violation (random versus structural), its identification can help to redefine the target population. PoRT's strength for this task is to provide interpretable missing eligibility criteria. In addition, the algorithm can indicate to the investigators the number of random violations (i.e., those considered as unlikely to be structural). The investigators could then choose an analytical approach, such as g-computation or doubly robust estimators, which can extrapolate/interpolate in the strata without sufficient support in addition to the robustness analysis [1,26]. We stress that positivity is not solely related to propensity score-based approaches, even if a vast amount of the positivity literature is found in this context. Randomized clinical trials may also present random violations when an adjustment or stratification procedure is considered.

Some limitations must be noted. First, this algorithm will never replace a careful study design. It only constitutes a tool for questioning potential positivity violations in the sample and consequently changing the target population for which the observed sample has sufficient support. A bootstrap process can be envisaged for studying the sensitivity of PoRT to small changes in the data, a weakness of RTs [38]. Second, we considered time-fixed binary exposure. Further developments are needed to consider multimodal or even continuous exposure, as well as exposure regimens requiring a more complex positivity assumption [39,40]. Third, we did not perform any simulation study to assess the predictive performance of the algorithm. PoRT is above all a practical tool to help in precisely defining the target population. Finally, we did not investigate the optimal set of hyperparameters, another perspective to be addressed.

In conclusion, we illustrated that the PoRT algorithm represents a helpful tool to precisely determine the targeted population in causal studies by providing transparent and clear randomized trial-like exclusion criteria. It enables essential discussions about possible structural and random violations prior to data analysis. Therefore, we recommend its use with different hyperparameters combinations to obtain an overview of potential violations.

Acknowledgments: The authors would like to thank the members of the AtlanREA and DIVAT groups for their involvement in the study, the physicians who helped recruit patients, and all patients who participated in this study. We also thank the clinical research associates who participated in the data collection. The analysis and interpretation of these data are the responsibility of the authors.

Funding information: This work was partially supported by a public grant overseen by the French National Research Agency (ANR) to create the Common Laboratory RISCA (Research in Informatic and Statistic for Cohort Analyses, www.labcom-risca.com, reference: ANR-16-LCV1-0003-01). The funder had no role in the study design, the analysis and interpretation of data, the writing of the report, or the decision to submit the report for publication.

Author contributions: G.D. developed the algorithm, analyzed the data, and drafted the manuscript. A.C. designed the study, analyzed the data, and wrote the manuscript. All authors interpreted the data, revised the manuscript, and read and approved the final version of the manuscript. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The dataset from Léger et al. is available as supplementary material of the initial study (10.1002/bimj.202000323). The datasets from the three other studies are available from the DIVAT consortium (www.divat.fr/access-to-data/request-guidelines) upon reasonable request.

References

[1] Hernán M, Robins JM. Causal inference: What if?. Boca Raton: Chapman & Hall/CRC; 2020.
[2] Westreich D. Epidemiology by design: A causal approach to the health sciences. NY: Oxford University Press; 2020.
[3] D'Amour A, Ding P, Feller A, Lei L, Sekhon J. Overlap in observational studies with high-dimensional covariates. J Econom. 2021;221:644–54. doi: 10.1016/j.jeconom.2019.10.014.
[4] Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res. 2012;21:31–54. doi: 10.1177/0962280210386207.
[5] Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. Am J Epidemiol. 2019;188:250–7. doi: 10.1093/aje/kwy201.
[6] Kallus N, Santacatterina M. Optimal balancing of time-dependent confounders for marginal structural models. J Causal Inference. 2021;9:345–69. doi: 10.1515/jci-2020-0033.
[7] Hill J, Su Y-S. Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. Ann. Appl Stat. 2013;7:1386–420. doi: 10.1214/13-AOAS630.
[8] Kang J, Chan W, Kim M-O, Steiner PM. Practice of causal inference with the propensity of being zero or one: assessing the effect of arbitrary cutoffs of propensity scores. Commun Stat Appl Methods. 2016;23:1–20. doi: 10.5351/CSAM.2016.23.1.001.
[9] Lundberg I, Johnson R, Stewart BM. What is your estimand? Defining the target quantity connects statistical evidence to theory. Am Sociol Rev. 2021;86:532–65. doi: 10.1177/00031224211004187.
[10] Zhu Y, Hubbard RA, Chubak J, Roy J, Mitra N. Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. Pharmacoepidemiol Drug Saf. 2021;30:1471–85. doi: 10.1002/pds.5338.
[11] Nethery RC, Mealli F, Dominici F. Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. Ann Appl Stat. 2019;13:1242–67. doi: 10.1214/18-AOAS1231.
[12] Pirracchio R, Carone M, Rigon MR, Caruana E, Mebazaa A, Chevret S. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. Stat Methods Med Res. 2016;25:1968–54. doi: 10.1177/0962280213507034.
[13] Platt RW, Delaney JAC, Suissa S. The positivity assumption and marginal structural models: the example of warfarin use and risk of bleeding. Eur J Epidemiol. 2012;27:77–83. doi: 10.1007/s10654-011-9637-7.
[14] Didelez V. Commentary: Should the analysis of observational data always be preceded by specifying a target experimental trial. Int J Epidemiol. 2016;45:2049–51. doi: 10.1093/ije/dyw032.
[15] García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. Eur J Epidemiol. 2017;32:495–500. doi: 10.1007/s10654-017-0287-2.
[16] Traskin M, Small DS. Defining the study population for an observational study to ensure sufficient overlap: A tree approach. Stat Biosci. 2011;3:94–118. doi: 10.1007/s12561-011-9036-3.
[17] Karavani E, Bak P, Shimoni Y. A discriminative approach for finding and characterizing positivity violations using decision trees. 2019. doi: 10.48550/arXiv.1907.08127.
[18] Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J Clin Epidemiol. 2010;63:826–33. doi: 10.1016/j.jclinepi.2009.11.020.
[19] Foucher Y, Le Borgne F, Dantan E, Gillaizeau F, Chatton A, Combescure C. RISCA: Causal Inference and Prediction in Cohort-Based Analyses; 2019. https://CRAN.R-project.org/package = RISCA (20 October 2021, date last accessed) n.d.
[20] Therneau TM, Atkinson B rpart: Recursive Partitioning and Regression Trees. 2019. https://CRAN.R-project.org/package = rpart (20 October 2021, date last accessed) n.d.

[21] Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution—A simulation study. Am J Epidemiol. 2010;172:843–54. doi: 10.1093/aje/kwq198.
[22] Glynn RJ, Lunt M, Rothman KJ, Poole C, Schneeweiss S, Stürmer T. Comparison of alternative approaches to trim subjects in the tails of the propensity score distribution. Pharmacoepidemiol Drug Saf. 2019;28:1290–8. doi: 10.1002/pds.4846.
[23] Lee BK, Lessler J, Stuart EA. Weight Trimming and Propensity Score Weighting. PLOS ONE. 2011;6:e18174. doi: 10.1371/journal.pone.0018174.
[24] Zhou Y, Matsouaka RA, Thomas L. Propensity score weighting under limited overlap and model misspecification. Stat Methods Med Res. 2020;29:3721–56. doi: 10.1177/0962280220940334.
[25] Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96:187–99. doi: 10.1093/biomet/asn055.
[26] Léger M, Chatton A, Le Borgne F, Pirracchio R, Lasocki S, Foucher Y. Causal inference in case of near-violation of positivity: comparison of methods. Biometrical J. 2022;64:1389–403. doi: 10.1002/bimj.202000323.
[27] Querard AH, Le Borgne F, Dion A, Giral M, Mourad G, Garrigue V, et al. Propensity score-based comparison of the graft failure risk between kidney transplant recipients of standard and expanded criteria donor grafts: towards increasing the pool of marginal donors. Am J Transpl. 2018;18:1151–7. doi: 10.1111/ajt.14651.
[28] Foucher Y, Fournier M-C, Legendre C, Morelon E, Buron F, Girerd S, et al. Comparison of machine perfusion versus cold storage in kidney transplant recipients from expanded criteria donors: a cohort-based study. Nephrol Dial Transpl. 2020;35:1043–70. doi: 10.1093/ndt/gfz175.
[29] Masset C, Boucquemont J, Garandeau C, Buron F, Morelon E, Girerd S, et al. Induction therapy in elderly kidney transplant recipients with low immunological risk. Transplantation. 2020;104:613–22. doi: 10.1097/TP.0000000000002804.
[30] Core Team R. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
[31] Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. Ann Intern Med. 2001;135:112–23. doi: 10.7326/0003-4819-135-2-200107170-00012.
[32] Westreich D, Cole SR. Invited commentary: positivity in practice. Am J Epidemiol. 2010;171:674–7. doi: 10.1093/aje/kwp436.
[33] Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol. 2008;168:656–64. doi: 10.1093/aje/kwn164.
[34] King G, Zeng L. When can history be our guide? The pitfalls of counterfactual inference. Int Stud Q. 2007;51:183–210. doi: 10.1111/j.1468-2478.2007.00445.x.
[35] Fogarty CB, Mikkelsen ME, Gaieski DF, Small DS. Discrete optimization for interpretable study populations and randomization inference in an observational study of severe Sepsis mortality. J Am Stat Assoc. 2016;111:447–58. doi: 10.1080/01621459.2015.1112802.
[36] Oberst M, Johansson F, Wei D, Gao T, Brat G, Sontag D, et al. Characterization of overlap in observational studies. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. PMLR; 2020. p. 788–98.
[37] Boulesteix A-L, Lauer S, Eugster MJA. A plea for neutral comparison studies in computational sciences. PLOS ONE. 2013;8:e61562. doi: 10.1371/journal.pone.0061562.
[38] Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16:199–231. doi: 10.1214/ss/1009213726.
[39] Schnitzer ME, Platt RW, Durand M. A tutorial on dealing with time-varying eligibility for treatment: Comparing the risk of major bleeding with direct-acting oral anticoagulants vs warfarin. Stat Med. 2020;39:4538–50. doi: 10.1002/sim.8715.
[40] Rudolph JE, Benkeser D, Kennedy EH, Schisterman EF, Naimi AI. Estimation of the average causal effect in longitudinal data with time-varying exposures: The challenge of nonpositivity and the impact of model flexibility. Am J Epidemiol. 2022;191:1962–9. doi: 10.1093/aje/kwac136.