The Annals of Statistics
2020, Vol. 48, No. 4, 2008–2030
https://doi.org/10.1214/19-AOS1874

SHARP INSTRUMENTS FOR CLASSIFYING COMPLIERS AND GENERALIZING CAUSAL EFFECTS

BY EDWARD H. KENNEDY[*], SIVARAMAN BALAKRISHNAN[**] AND MAX G'SELL[†]

Department of Statistics and Data Science, Carnegie Mellon University, [*]*edward@stat.cmu.edu;* [**]*siva@stat.cmu.edu;* [†]*mgsell@stat.cmu.edu*

It is well known that, without restricting treatment effect heterogeneity, instrumental variable (IV) methods only identify "local" effects among compliers, that is, those subjects who take treatment only when encouraged by the IV. Local effects are controversial since they seem to only apply to an unidentified subgroup; this has led many to denounce these effects as having little policy relevance. However, we show that such pessimism is not always warranted: it can be possible to accurately predict who compliers are, and obtain tight bounds on more generalizable effects in identifiable subgroups. We propose methods for doing so and study estimation error and asymptotic properties, showing that these tasks can sometimes be accomplished even with very weak IVs. We go on to introduce a new measure of IV quality called "sharpness," which reflects the variation in compliance explained by covariates, and captures how well one can identify compliers and obtain tight bounds on identifiable subgroup effects. We develop an estimator of sharpness and show that it is asymptotically efficient under weak conditions. Finally, we explore finite-sample properties via simulation, and apply the methods to study canvassing effects on voter turnout. We propose that sharpness should be presented alongside strength to assess IV quality.

1. Introduction. Instrumental variable (IV) methods are a widespread tool for identifying causal effects in studies where treatment is subject to unmeasured confounding. These methods have been used in econometrics since the 1920s [49], but have only been set within a formal potential outcome framework more recently [21, 31, 35]. Roughly speaking, an instrument is a variable that is associated with treatment, but is itself unconfounded and does not directly affect outcomes. An archetypal example is in randomized experiments with noncompliance, where initial randomization can be an instrument for the treatment that was actually received. IV methods are also used widely in observational studies, where investigators try to exploit natural randomness in, for example, treatment preference, distance or time. We refer to Baiocchi, Cheng, and Small [5], Hernán and Robins [18], Imbens [20] for a more comprehensive review and examples.

Despite their popularity and prevalence, instrument variable methods bring some difficulties that do not arise in studies of unconfounded treatments. In particular, without restricting treatment effect heterogeneity in some way or adding extra assumptions, one cannot identify average treatment effects across the entire population. For example, even in the simplest setting involving a randomized study with one-sided noncompliance (e.g., where subjects randomized to control cannot access treatment), the treatment effect is nonparametrically identified only among those who actually receive treatment.

One option then is to pursue bounds on the overall average treatment effect [6, 31, 35]. This approach is robust, but has been criticized on the grounds that the resulting inferences can

Received January 2018; revised May 2019.
MSC2020 subject classifications. 62G05, 62H30.
Key words and phrases. Causal inference, instrumental variable, noncompliance, observational study, generalizability.

be so imprecise that they are not helpful for making policy decisions. Others argue that even wide bounds are useful, by making explicit that any more precision would require further assumptions [37]. An alternative approach incorporates extra assumptions to achieve point identification. Classically, this was often accomplished via constant treatment effect assumptions within linear structural equation models. More recent generalizations allow for heterogeneous treatment effects and nonlinear models based on weaker homogeneity restrictions, for example, no effect modification by the instrument, or other no-interaction or parametric assumptions [36, 43]. However, as noted by Tchetgen Tchetgen and Vansteelandt [44], parametric identification can be problematic since it a priori restricts the effect of interest, and such functional form knowledge is not typically available in practice.

Yet another strategy instead assumes monotonicity [21, 35], which rules out the possibility that the instrument could encourage someone to take control when they would otherwise take treatment (i.e., rules out so-called defiers). This approach is unique in allowing nonparametric identification of a causal effect, but only a local effect among the subgroup of compliers, that is, those subjects who would only take treatment when encouraged by the instrument [2, 21]. These local average treatment effects (LATEs) have generated some controversy, since they are defined in an unidentified subgroup that is not directly observed; we refer to Imbens [20] and Swanson and Hernán [42] for a recent debate. The issue is that, for encouraged subjects, we never get to see whether they would have taken treatment if not encouraged, and vice versa for unencouraged subjects. Therefore, it is generally unknown whether any given subject is a complier or not.

One justification for continuing to pursue complier effects is that they allow something causal to be learned in broken or "second-best" studies with unmeasured confounding, even without restricting effect heterogeneity [19, 20]. In other words, although complier effects may not be an ideal target estimand, in reality most observational studies are confounded and so the ideal is not attainable. Despite this, one might argue, complier effects can still reveal a piece of the puzzle of the causal structure, and can in principle be used together with bounds on more standard effects.

However, such justification is not always convincing, yielding some lively debate. Robins and Greenland [37] stressed early on that the complier subgroup is not identified, and gave examples where complier effects are not of primary policy interest. Pearl [33] says the complier "subpopulation cannot be identified and, more seriously, it cannot serve as a basis for policies." Deaton [11] compares targeting local effects to the drunk who only looks for his keys near the lamppost, since that is where the light is. Swanson and Hernán [42] state that complier effects "only pertain to an unknown subset of the population," and that "as we do not know who is a complier, we do not know to whom our new policy should apply." These kinds of critiques suggest that generalization via complier effects is a hopeless endeavor. In this paper, we explore whether this is necessarily the case.

1.1. *Motivating example.* The most common way to judge an instrument's quality is by its strength, typically defined as the proportion of compliers $\mathbb{P}(C = 1)$ [5], where C is the unobserved indicator of complier status. However, consider Figure 1.

In this toy example, there is a single covariate $X \sim N(0, 1)$ and three candidate instruments, (Z_1, Z_2, Z_3). All three instruments have exactly the same strength, each yielding 30% compliers in the population. However, the available information about compliers changes drastically across the three cases. For the first instrument Z_1, it is only known that the probability of compliance is 30% for each subject, regardless of covariate value. Thus there is no additional information beyond the marginal strength; this is the worst-case setup often considered in critiques of complier-specific effects. However, consider the third instrument Z_3. For

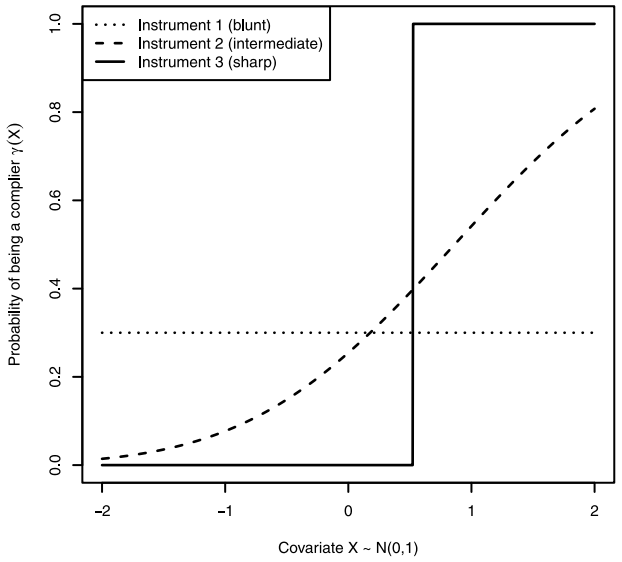


FIG. 1. *Compliance probability* $\gamma(x) = \mathbb{P}(C = 1 \mid X = x)$ *for three equally strong IVs.*

this instrument, the covariate X perfectly predicts compliance, so that $C = \mathbb{1}(X > 0.5244\ldots)$ and the complier-specific effect

$$\mathbb{E}\big(Y^{a=1} - Y^{a=0} \mid C = 1\big) = \mathbb{E}\big(Y^{a=1} - Y^{a=0} \mid X > 0.5244\ldots\big)$$

is in fact just a conditional effect within an observable subgroup. Therefore, when using Z_3 as an instrument all aforementioned concerns about local effects fall away completely. Importantly, this fact is not reflected at all in the strength of the instrument. It is also missed by the first-stage F-statistic, another common measure of instrument quality [8, 40], regardless of whether modeling assumptions are correct or not; we provide a simulated example in Appendix A [26]. The second instrument Z_2 is an intermediate between Z_1 and Z_3.

This example raises many interesting questions, which arise more generally in any instrument variable study. How can we quantify the extra information afforded by instruments like Z_2 relative to Z_1? Can we leverage this information to obtain more accurate guesses of who the compliers are? Can this help us go beyond local effects and instead identify effects in observable subgroups? The goal of this paper is to provide answers to these questions. Overall, we find that pessimism about local effects may be warranted in studies with blunt instruments. However, our work indicates that many concerns can be ameliorated or avoided in studies with sharp instruments, even if they are weak.

1.2. *Outline and contributions.* In this paper, we characterize sharp instruments as those that admit accurate complier predictions, and tight bounds on effects in identifiable subgroups. We present some notation and our assumptions in Section 2. In Section 3, we discuss the problem of classifying compliers. We propose several complier classification rules, derive their large-sample errors and discuss optimality and estimation. In Section 4, we discuss using instruments to bound effects in identifiable subgroups, characterize the subgroup that yields tightest bounds and propose corresponding estimators for these bounds. In Section 5, we propose a new summary measure of instrument quality called sharpness, which is separate from strength, and measures the variation in compliance explained by compliance scores. We show that sharper instruments yield better identification of compliers and tighter bounds on effects in identifiable subgroups, and present an efficient nonparametric estimator of the sharpness of an instrument. Our estimators are based on influence functions so as to yield fast

convergence rates and tractable inference even when relying on modern flexible regression methods; all methods are implemented in the `npcausal` package in R. Finally, in Section 6 we study finite-sample properties via simulation, and apply our methods in a study of effects of canvassing on voter turnout [17].

2. Notation and setup. We consider the usual instrumental variable setup, where one observes an i.i.d. sample $\{\mathbf{O}_1, \dots, \mathbf{O}_n\} \sim \mathbb{P}$ with

$$\mathbf{O} = (\mathbf{X}, Z, A, Y)$$

for covariates $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$, a binary instrument $Z \in \{0, 1\}$, a binary treatment $A \in \{0, 1\}$ and some outcome $Y \in [0, 1]$ of interest. We let Y^a denote the potential outcome [38] that would have been observed had treatment been set to $A = a$, and the goal is to learn about the distribution of the treatment effect $Y^{a=1} - Y^{a=0}$. We also need to define potential outcomes under interventions on the instrument. Thus let Y^{za} denote the potential outcome that would have been observed under both $Z = z$ and $A = a$, and similarly let A^z and $Y^z = Y^{zA^z}$ denote the potential treatment and outcome when the instrument is set to $Z = z$. In the statement of some of our results, we use the standard statistical big-O notation, as well as the shorthand $a \lesssim b$ to denote $a \leq Cb$ for some universal positive constant $C > 0$.

To ease the presentation, we let

$$\pi_z(\mathbf{x}) = \mathbb{P}(Z = z \mid \mathbf{X} = \mathbf{x}), \qquad \lambda_z(\mathbf{x}) = \mathbb{P}(A = 1 \mid \mathbf{X} = \mathbf{x}, Z = z),$$

denote the instrument propensity score and treatment regression, and let

$$\gamma(\mathbf{x}) = \lambda_1(\mathbf{x}) - \lambda_0(\mathbf{x})$$

denote the corresponding IV-difference.

We let $C = \mathbb{1}(A^{z=1} > A^{z=0})$ denote the latent variable indicating whether a subject is a complier, that is, whether a subject would respond to encouragement by the instrument. As mentioned in Section 1, C is not directly observed. Nonetheless, it is well known [1, 2, 18] that causal effects among compliers are nonparametrically identified under the following assumptions.

ASSUMPTION 1 (Consistency). $A = ZA^{z=1} + (1 - Z)A^{z=0}$ and $Y = ZY^{z=1} + (1 - Z)Y^{z=0}$.

ASSUMPTION 2 (Positivity). $\mathbb{P}\{\epsilon \leq \pi_z(\mathbf{X}) \leq 1 - \epsilon\} = 1$ for some $\epsilon > 0$.

ASSUMPTION 3 (Unconfounded IV). $Z \perp\!\!\!\perp (A^z, Y^z) \mid \mathbf{X}$.

ASSUMPTION 4 (Exclusion restriction). $Y^{za} = Y^a$.

ASSUMPTION 5 (Strong monotonicity). $\mathbb{P}(A^{z=1} < A^{z=0}) = 0$ and $\mathbb{P}(C = 1) \geq \epsilon > 0$.

(Note the lower-case indices z, a represent arbitrary values of the instrument and treatment.) We refer elsewhere [1, 2, 18] for a detailed discussion of the above assumptions, which are standard in the literature (as mentioned in Section 1, monotonicity is sometimes replaced by effect homogeneity or no-interaction assumptions). Assumptions 1–5 imply that the average effect among compliers (called the local average treatment effect, or LATE) with $\mathbf{V} = \mathbf{v}$ (for any subset $\mathbf{V} \subseteq \mathbf{X}$) is given by

$$\begin{aligned} &\mathbb{E}\big(Y^{a=1} - Y^{a=0} \mid \mathbf{V}, C = 1\big) \\ &\quad = \frac{\mathbb{E}\{\mathbb{E}(Y \mid \mathbf{X}, Z = 1) - \mathbb{E}(Y \mid \mathbf{X}, Z = 0) \mid \mathbf{V}\}}{\mathbb{E}\{\mathbb{E}(A \mid \mathbf{X}, Z = 1) - \mathbb{E}(A \mid \mathbf{X}, Z = 0) \mid \mathbf{V}\}}. \end{aligned} \tag{2.1}$$

This is the kind of local effect discussed in Section 1. Crucially, Assumptions 1–3 and 5 also imply that the chance of being a complier given covariates is given by

$$\mathbb{P}(C = 1 \mid \mathbf{X} = \mathbf{x}) = \gamma(\mathbf{x})$$

and so strength is given by $\mu \equiv \mathbb{P}(C = 1) = \mathbb{E}\{\gamma(\mathbf{X})\}$. The function $\gamma(\mathbf{x})$ has been termed the "compliance score" [3, 16, 25], and is an example of a "principal score" [13, 14, 24, 41]. Note that the principal score literature typically assumes independence between principal strata indicators (e.g., C) and potential outcomes, which we avoid here.

3. Classifying compliers. Heuristically, we propose calling instruments sharp when it is possible to predict compliance well, and obtain tight bounds on effects in identifiable subgroups. In this section, we discuss the first of these properties, that is, that of predicting the latent complier status C based on observed covariate information $\mathbf{X}$. We present several complier classification rules, characterize their errors and the relations between them, and discuss optimality. Finally, we present corresponding estimators and discuss estimation error and large-sample properties.

REMARK 1. Our view is that complier classification can be a valuable tool in practice, complementary to assessing compliance scores γ on their own. A first reason why it is pragmatic: it may be simply preferred (e.g., based on ease of interpretation) for practitioners to inspect a concrete set of likely compliers. Also, as we will discuss shortly, there is one particular classifier whose predicted compliers can act as surrogates for estimating any complier characteristic. Another pragmatic justification is that, statistically, complier classification is at least as easy as compliance score estimation: as in standard classification, one's score estimates could be severely biased and yet good classification error might still be attainable. For a trivial example, suppose $\gamma = 0$ for all $\mathbf{x}$ so there are no compliers, but estimated compliance scores $\widehat{\gamma} = 0.4$ everywhere and so are highly biased; even so, the classifier $\widehat{h} = \mathbb{1}(\widehat{\gamma} > t)$ is perfectly accurate for all $t \geq 0.4$.

Importantly, classification is also particularly crucial whenever decision-making is required. For example, from a policy perspective, encouraging noncompliers may be wasted effort since noncompliers will by definition have the same behavior regardless of encouragement. Thus one could consider the following two-stage treatment policy: first, compliance status is predicted, and then treatment is recommended only to those predicted compliers who are expected to benefit. Complier classification could also be useful for simultaneously minimizing noncompliance and increasing generalizability in experiments: for example, one could run a doubly randomized preference trial [32] where those subjects who are predicted to be compliers are randomized to the experimental arm with a higher probability, whereas predicted noncompliers are randomized to the observational arm with a higher probability. We aim to explore the use of complier classification in these specific decision-making contexts in detail in future work.

3.1. *Classifiers and properties.* As noted earlier, although compliers are not strictly identified it is possible to predict compliance status based on the fact that Assumptions 1–3 and 5 suffice to ensure that

$$\mathbb{P}(C = 1 \mid \mathbf{X} = \mathbf{x}) = \gamma(\mathbf{x}).$$

As stated in the following proposition, we can similarly identify the classification error $\mathcal{E}(h) = \mathbb{P}(C \neq h)$ for any given complier classification rule h, which we define as an arbitrary measurable function $h : \mathcal{X} \mapsto \{0, 1\}$ mapping the covariates to a binary prediction. As discussed further following (3.3), this proposition and subsequent results generalize in a natural way to classifiers that are stochastic.

PROPOSITION 1. *For any complier classification rule $h : \mathcal{X} \mapsto \{0, 1\}$, the corresponding classification error $\mathcal{E}(h) = \mathbb{P}\{C \neq h(\mathbf{X})\}$ is identified under Assumptions 1–3 and 5 as*

$$\mathcal{E}(h) = \mathbb{E}\big[\gamma(\mathbf{X})\big\{1 - h(\mathbf{X})\big\} + \big\{1 - \gamma(\mathbf{X})\big\}h(\mathbf{X})\big].$$

A proof of Proposition 1 and all other results can be found in the Appendix [26]. Although the compliance score has been discussed in the literature since at least Follmann [16] and Joffe et al. [25], we have not seen it used before for the specific purpose of predicting who the compliers are, nor have we seen any discussion of the error of this task. In contrast, most work seems to focus on the related but separate problem of estimating complier characteristics, such as $\mathbb{E}(\mathbf{X} \mid C = 1)$ [1, 5]. As explained above, we feel compliance classification is practically important and yet understudied, particularly for so-called sharp instruments that allow for accurate prediction. If the error $\mathcal{E}(h)$ can be made small, then it is possible to know who the compliers are quite precisely. A main point of this paper is to formalize this and show that it is possible for compliers to be accurately classified even with weak instruments.

The optimal classifier h_0 in terms of minimizing the error $\mathcal{E}(h)$ is given by the Bayes decision function

$$\operatorname*{arg\,min}_{h:\mathcal{X}\mapsto\{0,1\}} \mathcal{E}(h) = \mathbb{1}\big\{\gamma(\mathbf{x}) > 1/2\big\} \equiv h_0(\mathbf{x}). \tag{3.1}$$

The proof of this fact follows from the same logic as in standard classification problems [12]. Shortly, we will discuss estimation of the Bayes decision via the plug-in estimator $\mathbb{1}(\widehat{\gamma} > 1/2)$. One could also consider empirical risk minimizers of the form

$$\widehat{h} = \operatorname*{arg\,min}_{h\in\mathcal{H}} \widehat{\mathcal{E}}(h)$$

for an appropriate class $\mathcal{H}$ (e.g., linear classifiers) and estimator $\widehat{\mathcal{E}}(h)$ of the error. We leave this to future work, only considering plug-in classifiers in this paper.

Despite its simplicity and optimality (with respect to classification error), the rule h_0 may have some practically unfavorable properties in the setting of complier classification. In particular, the set of putative compliers returned by h_0 could have a very different size compared to the true set. We call classifiers strength-calibrated if they output sets with the same size as the true set.

PROPERTY 1. A complier classification rule $h : \mathcal{X} \mapsto \{0, 1\}$ is *strength-calibrated* if

$$\mathbb{P}\big\{h(\mathbf{X}) = 1\big\} = \mathbb{P}(C = 1). \tag{P1}$$

If for no other reason, strength calibration can be important in complier classification simply because strength $\mu = \mathbb{P}(C = 1)$ is such a fundamental quantity in instrumental variable problems. Strength is often the primary criterion used to judge instrument quality, since the more compliers there are, the more subjects there are for whom the local effect is relevant, and so the more meaningful and generalizable the effect is. Thus one might prefer to trade off some error for a classification rule that accurately reflects the underlying size of the complier population, for instance, in settings where achieving a minimum error threshold is sufficient, rather than precise minimization.

Similarly, it is possible that the optimal rule h_0 would never guess any compliers (i.e., $h_0 = 0$ with probability one), which could be unfavorable for a practical analysis. For example, suppose $\gamma = \mu = 49\%$, or that the covariate X was uniform and $\gamma(x) = x/2$. Then the optimal rule h_0 would return the empty set in both cases, even though the proportion of compliers is nearly one-half and a quarter, respectively. The empty set could be an unsatisfying result for a practitioner who was curious about identifying which particular subjects were compliers.

A simple strength-calibrated rule is given by the quantile-threshold classifier

$$h_q(\mathbf{x}) = \mathbb{1}\{\gamma(\mathbf{x}) > q\}, \tag{3.2}$$

where $q = F^{-1}(1-\mu)$, and $F(t) = \mathbb{P}\{\gamma(\mathbf{X}) \leq t\}$ is the cumulative distribution function of the compliance score. The rule h_q simply predicts that the $100\mu\%$ of subjects with the highest compliance scores are the compliers. That h_q is strength-calibrated follows since

$$\mathbb{P}(h_q = 1) = \mathbb{P}\{F(\gamma) > 1-\mu\} = 1-(1-\mu)$$

because $F(\gamma)$ follows a uniform distribution. Here, we have assumed there exists an exact (unique) quantile q such that $F(q) = 1-\mu$; when this does not hold, one could instead enforce a weaker condition like $\mathbb{P}(h=1) \geq \mathbb{P}(C=1)$. In the next subsection, we show that when there is a unique quantile, no other strength-calibrated rule can achieve a better classification error than h_q.

One could similarly consider rules of the form $h_t(\mathbf{x}) = \mathbb{1}\{\gamma(\mathbf{x}) > t\}$ for a generic $t \in [0, q]$, if a finer trade-off between classification error and size is required, for example, if the increase in classification error when moving from h_0 to h_q is too severe.

Another restriction that may be useful to consider in complier classification problems is that of ensuring the covariate distributions among the predicted and true compliers are the same. We call this distribution-matching.

PROPERTY 2. A complier classification rule $h : \mathcal{X} \mapsto \{0,1\}$ is *distribution-matched* if

$$\mathbb{P}\{\mathbf{X} \leq \mathbf{x} \mid h(\mathbf{X}) = 1\} = \mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid C = 1) \quad \forall \mathbf{x}. \tag{P2}$$

Distribution matching is useful as it allows practitioners to query the covariate distribution among predicted compliers to learn about the true complier distribution. This provides a user-friendly method for assessing complier characteristics, which can be an alternative to direct estimation via the identifying expressions given for example by Abadie [1]. Strength-calibration and distribution-matching together imply that $\mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid h = 0) = \mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid C = 0)$, so the statistician can also estimate prevalence ratios [5] like $\mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid C = 1)/\mathbb{P}(\mathbf{X} \leq \mathbf{x})$ by simply comparing predicted compliers to the whole sample, that is, by estimating $\mathbb{P}(\mathbf{X} \leq \mathbf{x} \mid h = 1)/\mathbb{P}(\mathbf{X} \leq \mathbf{x})$ for a distribution-matched classifier h.

In fact, we show in the next subsection that the only rule that is both strength-calibrated and distribution-matched is the stochastic classifier

$$h_s(\mathbf{x}) = \mathbb{1}\{\gamma(\mathbf{x}) > U\} \sim \text{Bernoulli}\{\gamma(\mathbf{x})\}, \tag{3.3}$$

where $U \sim \text{Unif}(0,1)$ is an independent draw from the uniform distribution on $[0,1]$. Note that h_s randomly predicts that a subject with covariates $\mathbf{x}$ is a complier with probability $\gamma(\mathbf{x})$. To be precise, since h_s is stochastic it should really also be indexed by U, as in $h_s(\mathbf{x}) = h_s(\mathbf{x}, U)$. It is implicit that any expectations $\mathbb{E}(h) = \mathbb{E}\{h(\mathbf{X}, U)\}$ are over both $\mathbf{X}$ and U.

3.2. *Classifier errors and relations.* In the following results, we characterize the errors of the classifiers h_q and h_s, show that they are optimal in the classes of strength-calibrated and distribution-matched classifiers, respectively, and relate their error to the minimal Bayes error $\mathcal{E}(h_0)$. Interestingly, the classification error for the stochastic classifier h_s takes a simple form, which equals the quadratic entropy, that is, the asymptotic error of a nearest neighbor classifier [10, 12].

THEOREM 1. *Suppose there is a unique* $(1-\mu)$ *quantile so that* $\mathbb{P}(\gamma > q) = \mu$. *Then for the quantile-threshold classifier* h_q *defined in* (3.2) *we have*

$$\mathcal{E}_q \equiv \mathcal{E}(h_q) = 2\mathbb{E}[\gamma(\mathbf{X})\mathbb{1}\{\gamma(\mathbf{X}) \leq q\}] \leq \mathcal{E}(h)$$

for any strength-calibrated $h : \mathcal{X} \mapsto \{0,1\}$ *with* $\mathbb{E}(h) = \mu$.

Further, the only classifier that is both strength-calibrated and distribution-matched is the stochastic classifier h_s defined in (3.3). Its error is given by

$$\mathcal{E}_s \equiv \mathcal{E}(h_s) = 2\mathbb{E}\{\gamma(\mathbf{X}) - \gamma(\mathbf{X})^2\}.$$

We prove Theorem 1 and all other results in the Appendix [26]. Since $\mathcal{E}_q \leq \mathcal{E}_s$ and $\mathcal{E}_s$ equals the asymptotic nearest-neighbor error, we can transport results from the standard classification setting accordingly. The following theorem from Cover and Hart [10], Devroye et al. [12] shows how these errors yield bounds on the optimal error $\mathcal{E}(h_0)$, and indicates how much worse they can be compared to $\mathcal{E}(h_0)$.

PROPOSITION 2 (Cover and Hart [10], Devroye et al. [12]). *Suppose there is a unique $(1-\mu)$-quantile q such that $\mathbb{P}(\gamma > q) = \mu$. Then the optimal classification error $\mathcal{E}(h_0)$ is bounded as*

$$\frac{1}{2}(1-\sqrt{1-2\mathcal{E}_s}) \leq \frac{1}{2}(1-\sqrt{1-2\mathcal{E}_q}) \leq \mathcal{E}(h_0) \leq \mathcal{E}_q \leq \mathcal{E}_s.$$

We further have the upper bound $\mathcal{E}_q \leq \mathcal{E}_s \leq \mathcal{E}(h_0)\{1-\mathcal{E}(h_0)\} \leq 2\mathcal{E}(h_0)$.

Proposition 2 follows from our Theorem 1 together with Theorem 3.1 of Devroye et al. [12], and shows that the errors of the stochastic and quantile classifiers can be quite informative about the optimal error $\mathcal{E}(h_0)$ of unconstrained classifiers. For example, if compliance status can be correctly predicted for 75% of the population with either classifier (e.g., $\mathcal{E}_s = 0.25$) then the optimal classifier can have no better than 86% accuracy. Theorem 1 further indicates that the errors $\mathcal{E}_q$ and $\mathcal{E}_s$ can never be worse than twice that of the best unconstrained classifier, which is particularly informative when $\mathcal{E}_q$ or $\mathcal{E}_s$ are not too large.

3.3. *Estimation.* The simplest way to estimate the proposed classification rules is via plug-in estimators. For example, the plug-in estimator of the Bayes decision function h_0 is given by

$$\widehat{h}_0(\mathbf{x}) = \mathbb{1}\{\widehat{\gamma}(\mathbf{x}) > 1/2\}. \tag{3.4}$$

Analogs of this estimator have been studied widely in the classification literature [4, 12]. However, the form of the Bayes classifier h_0, in our setting, brings some additional complications relative to the standard classification setting, since $\gamma(\mathbf{x}) = \lambda_1(\mathbf{x}) - \lambda_0(\mathbf{x})$ is a difference in regression functions. For example, the minimax convergence rate for estimating γ can depend not only on the smoothness of γ, but also on the smoothness of λ_z and π. This is an open problem and beyond the scope of this paper; nonetheless, we can still relate the error of $\widehat{h}_0$ to that of $\widehat{\gamma}$, as in standard classification problems. Specifically, as in Theorem 2.2 of Devroye et al. [12] we have

$$\mathcal{E}(\widehat{h}) - \mathcal{E}(h_0) \leq 2\|\widehat{\gamma} - \gamma\|,$$

where here and throughout we let $\|f\|^2 = \mathbb{P}(f^2) = \int f(\mathbf{o})^2 d\mathbb{P}(\mathbf{o})$ denote the squared $L_2(\mathbb{P})$ norm (in fact the above also holds replacing the L_2 with the L_1 norm). This shows that consistent estimation of the compliance score γ is enough to yield a consistent plug-in estimator of the rule h_0, in terms of classification error.

A plug-in estimator for the quantile rule h_q is given by

$$\widehat{h}_q(\mathbf{x}) = \mathbb{1}\{\widehat{\gamma}(\mathbf{x}) > \widehat{q}\}, \tag{3.5}$$

where $\widehat{q}$ is an estimate of the $(1-\mu)$ quantile of γ, that is, an estimate of q for which $\mathbb{P}(\gamma \leq q) = 1-\mu$. For example, one could use $\widehat{q} = \widehat{F}^{-1}(1-\widehat{\mu})$, for initial estimators $\widehat{F}$ and

$\widehat{\mu}$ of the distribution function and mean of the compliance score, respectively. In the next subsection, we will detail an efficient estimator of μ, which is doubly robust and can attain the minimax root-n convergence rate even if $(\widehat{\pi}, \widehat{\lambda}_z)$ converge at slower nonparametric rates. Finally, a plug-in estimator of the stochastic classification rule is given by

$$\widehat{h}_s(\mathbf{x}) = \mathbb{1}\{\widehat{\gamma}(\mathbf{x}) > U\} \tag{3.6}$$

for $U \sim \text{Unif}(0,1)$, so that $\widehat{h}_s \sim \text{Bernoulli}(\widehat{\gamma})$. We note that, although natural, the plug-in classifiers described above are not necessarily exactly strength-calibrated or distribution-matched when estimated from a finite-sample. For the plug-in estimators in (3.5) and (3.6), the next result gives a bound, relating excess classification error to error of the estimated compliance score (and quantile estimation error for $\widehat{h}_q$). For the quantile classifier, we require a margin condition [4], which controls the behavior of γ around the threshold q. Formally, we have the following condition.

ASSUMPTION 6 (Margin condition). For some $\alpha > 0$ and for all t, we have that

$$\mathbb{P}(|\gamma - q| \le t) \lesssim t^{\alpha}. \tag{3.7}$$

The margin condition requires that there are not too many compliance scores near the quantile q. This is essentially equivalent to the margin condition used in standard classification problems [4], optimal treatment regime settings [30, 46], as well as other problems involving estimation of nonsmooth covariate-adjusted bounds [27].

Overall, the following result shows that plug-in classifiers using accurate nuisance estimates have small excess error.

THEOREM 2. *Let $\widehat{h}_q$ and $\widehat{h}_s$ be the plug-in classifiers defined in* (3.5) *and* (3.6). *Then for $\widehat{h}_s$*

$$|\mathcal{E}(\widehat{h}_s) - \mathcal{E}_s| \le (\sqrt{1 - 2\mathcal{E}_s})\|\widehat{\gamma} - \gamma\|.$$

Furthermore, under the margin condition, for $\widehat{h}_q$ we have that

$$|\mathcal{E}(\widehat{h}_q) - \mathcal{E}_q| \lesssim (\|\widehat{\gamma} - \gamma\|_\infty + |\widehat{q} - q|)^{\alpha}.$$

REMARK 2. From a theoretical standpoint, we might consider if the margin assumption may be eliminated in the analysis of the plug-in quantile classifier. In Appendix D [26], we show that if we can obtain reasonable bounds on the errors $\|\widehat{\gamma} - \gamma\|_\infty$ and $|\widehat{q} - q|$, a slight modification of the plug-in quantile classifier in (3.5) achieves a similar guarantee without the margin assumption.

The next result shows a further unique property of $\widehat{h}_s$, which is that it can be used to estimate complier characteristics of the form $\theta = \mathbb{E}\{f(\mathbf{X}) \mid C = 1\}$, by simply computing corresponding averages in the group of predicted compliers with $\widehat{h}_s = 1$. For example, one might be interested in, for a given variable X_j, the complier-specific mean $f(\mathbf{X}) = X_j$ or distribution function $f(\mathbf{X}) = \mathbb{1}(X_j \le t)$. The proposed estimator is then given by

$$\widehat{\theta} = \mathbb{P}_n\{f(\mathbf{X}) \mid \widehat{h}_s(\mathbf{X}) = 1\} = \frac{\mathbb{P}_n\{f(\mathbf{X})\widehat{h}_s(\mathbf{X})\}}{\mathbb{P}_n\{\widehat{h}_s(\mathbf{X})\}}, \tag{3.8}$$

where $\mathbb{P}_n$ denotes the empirical measure so that sample averages can be written as $\mathbb{P}_n\{f(\mathbf{O})\} = \frac{1}{n}\sum_{i=1}^n f(\mathbf{O}_i)$. For simplicity, we suppose $\widehat{\gamma}$ is fit in a separate independent sample; this will be discussed in more detail after stating the result.

THEOREM 3. *Assume that f is bounded, then for the estimator $\widehat{\theta}$ defined in (3.8) we have that*

$$|\widehat{\theta} - \theta| = O_{\mathbb{P}}\Big(\frac{1}{\sqrt{n}} + \|\widehat{\gamma} - \gamma\|\Big),$$

whenever $\widehat{\gamma}$ is constructed from a separate independent sample.

Theorem 3 shows that $\widehat{\theta}$ is consistent as long as $\widehat{\gamma}$ is, and that the convergence rate is of the same order as a typical plug-in estimator. This gives an alternative to the weighting approach of Abadie [1]. Our approach only requires computing usual statistics among predicted compliers. In general, however, this approach will not be fully efficient for two reasons. The first is that $\widehat{\theta}$ is a plug-in estimator, not specially targeted to estimate θ well (partly evidenced by the first-order bias term $\|\widehat{\gamma} - \gamma\|$ in its convergence rate). We conjecture that $\widehat{\theta}$ might be able to attain full nonparametric efficiency under strong smoothness assumptions and for particular $\widehat{\gamma}$ estimators (e.g., kernel regression with undersmoothing). However, a more flexible approach would be to estimate θ with an appropriate doubly robust influence function-based estimator. The other reason the estimator $\widehat{\theta}$ is not fully efficient is because it uses only a single sample split, however, this can be remedied by swapping samples and averaging; we formally include this approach in our subsequent proposed estimators of effect bounds and sharpness. Despite disadvantages with respect to efficiency, the proposed plug-in estimator of θ might be favored in some settings for its simplicity.

4. Bounding effects in identifiable subgroups. In this section, we consider the second feature of so-called sharp instruments: obtaining tighter bounds on effects in identifiable subgroups, that is, subgroups defined not by principal strata (e.g., compliers) but by observed covariates. In the toy example from Section 1, we saw a case where the local effect actually reduced to such a subgroup effect (among those with $X > 0.5244\ldots$). This raises the question of when this can occur, and if it cannot, how to quantify the extent to which it can nearly occur. We derive bounds on effects in any identifiable subgroup and derive the corresponding bound length, and characterize the optimal subgroup that minimizes bound length, among all subgroups of a given size. Finally, we propose efficient nonparametric bound estimators, and describe their asymptotic properties.

4.1. *Bounds and bound length.* Define the treatment effect in an identifiable subgroup $\{\mathbf{x} : g(\mathbf{x}) = 1\}$ corresponding to an arbitrary measurable subgroup indicator $g : \mathcal{X} \mapsto \{0, 1\}$ as

$$\beta(g) = \mathbb{E}\big(Y^{a=1} - Y^{a=0} \mid g = 1\big).$$

Our first result gives bounds on this effect under the instrumental variable assumptions, for any given g. Before stating our result, let us first introduce some notation. Define

$$\beta_j(g) = \mathbb{E}\big\{\mathbb{E}(V_{j,1} \mid \mathbf{X}, Z = 1) - \mathbb{E}(V_{j,0} \mid \mathbf{X}, Z = 0) | g = 1\big\} \tag{4.1}$$

for $j \in \{l, u\}$ where

$$V_{u,1} = YA + 1 - A, \qquad V_{u,0} = Y(1 - A), \tag{4.2}$$

$$V_{l,1} = YA, \qquad V_{l,0} = Y(1 - A) + A. \tag{4.3}$$

With these definitions in place, we have the following result.

THEOREM 4. *Under Assumptions 1–5, and if $\mathbb{P}(Y \in [0, 1]) = 1$, the effect $\beta(g)$ in the identifiable subgroup defined by $g : \mathcal{X} \mapsto \{0, 1\}$ is bounded as*

$$\beta_l(g) \leq \beta(g) \leq \beta_u(g).$$

Theorem 4 generalizes the results of Balke and Pearl [6], Manski [31], Robins [35] to allow for covariate adjustment and conditional effects; these previous bounds are recovered by taking $\mathbf{X} = \varnothing$ and $g = 1$ with probability one. The logic used in the proof of Theorem 4 follows that of this earlier work. Specifically, as shown above, Assumptions 1–5 allow one to express $\beta(g)$ in terms of observed data quantities and two nonidentified terms of the form $\mathbb{E}(Y^{a=t} \mid A^{z=t} = 1 - t, g = 1)$ for $t \in \{0, 1\}$; bounds are obtained by replacing these latter quantities with their most extreme values of 0 and 1. Note the condition that $Y \in [0, 1]$ is immaterial as long as Y is bounded in some finite range $[y_{\min}, y_{\max}]$, since then one can work with $Y^* = (Y - y_{\min})/(y_{\max} - y_{\min}) \in [0, 1]$ and transform back.

An important consequence of Theorem 4 for our work is in the length of the corresponding bounds, which provides a basis for quantifying near-identification of effects $\beta(g)$ in identifiable subgroups. This length is given in the following corollary.

COROLLARY 1. *The length of the bounds in Theorem 4 for any subgroup h is*

$$\ell(g) \equiv \beta_u(g) - \beta_l(g) = \mathbb{E}\{1 - \gamma(\mathbf{X}) \mid g = 1\}.$$

Importantly, under Assumptions 1–5 we have

$$\ell(g) = \mathbb{P}(C = 0 \mid g = 1),$$

so the bound length is also interpretable as the proportion of noncompliers in the subgroup $\{\mathbf{x} : g(\mathbf{x}) = 1\}$. This fact was noted previously for marginal effects (i.e., when $g = 1$ with probability one) by Balke and Pearl [6], for example. It implies that the bounds on the subgroup effect $\beta(g)$ are strictly narrower than those on the average effect $\mathbb{E}(Y^{a=1} - Y^{a=0})$ whenever $\mathbb{P}(C = 0 \mid g = 1) < \mathbb{P}(C = 0)$, that is, whenever the proportion of noncompliers in the subgroup is less than the proportion overall. In Section 5, we frame this condition in a different way that shows how it is intimately related to our proposed notion of sharpness.

Corollary 1 further suggests exploring subgroups that minimize bound length. Among all possible subgroups, the one minimizing bound length is simply that which picks the subject(s) with the maximum compliance score, that is, $\arg\min_h \ell(g) = \mathbb{1}(\gamma = \gamma_{\max})$ for $\gamma_{\max} = \sup_{\mathbf{x} \in \mathcal{X}} \gamma(\mathbf{x})$. However, in general this subgroup will have negligible size (unless there is a non-trivial point mass at $\gamma_{\max}$), leading to estimates with necessarily high finite-sample error. This is similar to the potential disadvantages of the optimal classification rule h_0 discussed in Section 3.1. Therefore, as discussed there, it may be preferable to only consider subgroups of a particular minimum size. We let

$$\mathcal{G}(t) = \{g : \mathbb{P}(g = 1) = t\}$$

denote the set of all subgroups of a given size t, and we assume there exists a unique quantile ξ such that $\mathbb{P}(\gamma > \xi) = t$. The following result gives the form of optimal subgroups of a given size.

PROPOSITION 3. *Let $F(t) = \mathbb{P}(\gamma \leq t)$ denote the distribution function of the compliance score. Then the subgroup that minimizes bound length among all those of size at least t is given by*

$$\arg\min_{g \in \mathcal{G}(t)} \ell(g) = \mathbb{1}\{\gamma(\mathbf{X}) > F^{-1}(1 - t)\}.$$

Proposition 3 shows that, among all subgroups of size t, the subgroup that yields the tightest bounds is simply the group with the $100t\%$ highest compliance scores. This is perhaps

expected given the form of the bounds from Corollary 1, and their interpretation as a proportion of noncompliers. Note also that, once we restrict to subgroups of a given size t, the minimizers of complier classification error and bound length are the same, that is,

$$\arg\min_{g\in\mathcal{G}(t)} \mathcal{E}(g) = \arg\min_{g\in\mathcal{G}(t)} \ell(g) = \mathbb{1}\{\gamma(\mathbf{X}) > F^{-1}(1-t)\}.$$

Therefore, for subgroups of a given size, the problems of finding the classifier with best error and the subgroup with tightest bounds are equivalent, both leading to a version of the quantile classifier h_q from Section 3.1.

This suggests targeting novel subgroup effects of the form

$$\mathbb{E}\{Y^{a=1} - Y^{a=0} \mid \gamma(\mathbf{X}) > F^{-1}(1-t)\}.$$

These effects are similar in spirit to those proposed by Follmann [16], Joffe et al. [25], which are also conditional on the compliance score, but these prior works use parametric models and do not use quantiles. Thus, our proposed effects can be viewed as a nonparametric generalization.

4.2. *Estimation and inference.* Now we turn to estimation and inference for bounds on $\beta(g)$. We focus in particular on $\beta(h_q)$, that is, the effect among the $100\mu\%$ of the population with the highest compliance scores. Our bound estimators (and sharpness estimators presented in the next section) are built from the efficient influence function and use sample splitting. These tools are used to combat bias from nonparametric estimation of nuisance functions (e.g., the compliance score γ) and to allow arbitrary complex and flexible nuisance estimators to be used.

Influence functions are a central element of nonparametric efficiency theory. We refer to Bickel et al. [7], van der Vaart [48], van der Laan and Robins [47], Tsiatis [45] and others for more detailed information, and so just give some brief description here. The efficient influence function is important because its variance yields a benchmark for nonparametric efficiency, and because it can be used to construct estimators that are in some cases minimax optimal and efficient in nonparametric models. Such estimators are typically doubly robust or have general second-order bias, and so can attain parametric rates of convergence, even in high-dimensional settings where nuisance functions are estimated at slower rates via flexible nonparametric methods. Mathematically, the efficient influence function corresponds to the score function in a one-dimensional submodel that is least favorable, in the sense of having minimal Fisher information for the parameter of interest, across all submodels. We refer to the earlier references for more details.

To simplify notation in this section, for any random variable T we let

$$\varphi_z(T;\boldsymbol{\eta}) = \frac{\mathbb{1}(Z=z)}{\pi_z(\mathbf{X})}\{T - \mathbb{E}(T \mid \mathbf{X}, Z=z)\} + \mathbb{E}(T \mid \mathbf{X}, Z=z)$$

denote the uncentered efficient influence function for the parameter $\mathbb{E}\{\mathbb{E}(T \mid \mathbf{X}, Z=z)\}$, where $\boldsymbol{\eta} = \{\pi_z(\mathbf{X}), \mathbb{E}(T \mid \mathbf{X}, Z=z)\}$ denotes the relevant nuisance functions. We use $\boldsymbol{\eta}$ for nuisance functions generally, though the actual functions depend on the choice of T. In particular, we let

$$\nu_{j,z}(\mathbf{X}) = \mathbb{E}(V_{j,1} \mid \mathbf{X}, Z=1) - \mathbb{E}(V_{j,0} \mid \mathbf{X}, Z=0),$$

and let $\widehat{\nu}_{j,z}$ denote an estimate of $\nu_{j,z}$, for variables $V_{j,z}$ defined as in (4.2).

Following Chernozhukov et al. [9], Robins et al. [34], Zheng and van der Laan [50], we propose to use sample splitting to allow for arbitrarily complex nuisance estimators $\widehat{\boldsymbol{\eta}}$ and avoid empirical process conditions, by constructing the estimated $\boldsymbol{\eta}$ values for each subject

using data from only other subjects. Specifically, we split the data into K disjoint groups by drawing variables $(B_1, \ldots, B_n)$ independent of the data, with $B_i = b$ indicating that subject i was split into group b. For example, each B_i could be drawn uniformly from $\{1, \ldots, K\}$ or to ensure equally sized groups $(B_1, \ldots, B_n)$ could be drawn uniformly from the set of permutations of sequences containing n/K repetitions of each value of $b \in \{1, \ldots, K\}$. In our analysis, we focus on the former setting and treat K as a fixed constant. We first estimate the strength of the instrument by the weighted average of corresponding estimators across groups

$$\widehat{\mu} = \sum_{b=1}^{K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(B_i = b) \right\} \mathbb{P}_n^b\{\phi_\mu(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{\text{-}b})\} = \mathbb{P}_n\{\phi_\mu(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{\text{-}B})\},$$

where $\mathbb{P}_n^b$ denotes the subempirical distribution over the units $\{i : B_i = b\}$ in group b, that is, $\mathbb{P}_n^b\{f(\mathbf{O})\} = \sum_{i=1}^n f(\mathbf{O}_i)\mathbb{1}(B_i = b)/\sum_{i=1}^n \mathbb{1}(B_i = b)$, the function

$$\phi_\mu(\mathbf{O}; \boldsymbol{\eta}) = \varphi_1(A; \boldsymbol{\eta}) - \varphi_0(A; \boldsymbol{\eta}),$$

is the uncentered influence function for $\mu = \mathbb{E}(\gamma)$, and here $\widehat{\boldsymbol{\eta}}_{\text{-}b}$ denotes estimators of $\boldsymbol{\eta} = (\pi_z, \lambda_z)$ constructed using only those units with $B_i \neq b$. Then for $j \in \{l, u\}$ we propose estimating $\beta_j(h_q)$ with $\widehat{\beta}_j(\widehat{h}_q)$, where

$$\widehat{\beta}_j(\widehat{h}_q) = \mathbb{P}_n\big[\{\varphi_1(V_{j,1}; \widehat{\boldsymbol{\eta}}_{\text{-}B}) - \varphi_0(V_{j,0}; \widehat{\boldsymbol{\eta}}_{\text{-}B})\}\widehat{h}_{q,\text{-}B}(\mathbf{X})\big]/\mathbb{P}_n\{\widehat{h}_{q,\text{-}B}(\mathbf{X})\}$$

for $\widehat{h}_{q,\text{-}b} = \mathbb{1}(\widehat{\gamma}_{\text{-}b} > \widehat{q}_{\text{-}b})$ and $\widehat{q}_{\text{-}b}$ the $(1 - \widehat{\mu})$ quantile of $\widehat{\gamma}$ solving $\mathbb{P}_n^b\{\mathbb{1}(\widehat{\gamma}_{\text{-}b} > \widehat{q}_{\text{-}b})\} = \widehat{\mu}$ (at least up to $o_{\mathbb{P}}(1/\sqrt{n})$ error).

Before stating our next result, we define the remainder terms that appear in our result:

$$R_{1,n} = \|\widehat{\pi}_1 - \pi_1\| \Big(\max_z \|\widehat{\lambda}_z - \lambda_z\| + \max_z \|\widehat{\nu}_{j,z} - \nu_{j,z}\| \Big), \tag{4.4}$$

$$R_{2,n} = \big(\|\widehat{\gamma} - \gamma\|_\infty + |\widehat{q} - q| \big)^\alpha, \tag{4.5}$$

where $\alpha > 0$ is the margin exponent in (3.7). The next theorem gives the rate of convergence for our proposed estimator, as well as nonparametric conditions under which it is asymptotically normal and efficient.

THEOREM 5. *Assume that* $\mathbb{P}\{\epsilon \leq \widehat{\pi}_z(\mathbf{X}) \leq 1 - \epsilon\} = 1$ *for* $z = 0, 1$ *and some* $\epsilon > 0$, *and that* $\|\widehat{\pi}_1 - \pi_1\| + \max_z \|\widehat{\lambda}_z - \lambda_z\| + \max_z \|\widehat{\nu}_{j,z} - \nu_{j,z}\| + \mathbb{P}(\widehat{h}_q \neq h_q) = o_{\mathbb{P}}(1)$.

1. *If the margin condition holds for some* α, *then*

$$\widehat{\beta}_j(\widehat{h}_q) - \beta_j(h_q) = O_{\mathbb{P}}\Big(\frac{1}{\sqrt{n}} + R_{1,n} + R_{2,n} \Big).$$

2. *If it also holds that* $R_{1,n} + R_{2,n} = o_{\mathbb{P}}(1/\sqrt{n})$, *then*

$$\sqrt{n}\{\widehat{\beta}_j(\widehat{h}_q) - \beta_j(h_q)\} \rightsquigarrow N\big(0, \operatorname{var}\big[\{\varphi_1(V_{j,1}) - \varphi_0(V_{j,0})\}h_q - \beta_j(h_q)\phi_\mu\big]/\mu^2\big).$$

Theorem 5 shows that the error in estimating bounds on $\beta(h_q)$ consists of a doubly robust second-order term $R_{1,n}$ that will be small if either π_z or $(\lambda_z, \nu_{j,z})$ are estimated accurately, along with a term $R_{2,n}$ that will be small if the compliance score γ is estimated accurately, and particularly so depending on a margin condition.

If the exponent in the margin condition is too small, for example, $\alpha \leq 1$, then the proposed estimators will not in general be asymptotically normal or even $\sqrt{n}$-consistent, for example, if $\widehat{\gamma}$ and $\widehat{q}$ are estimated nonparametrically at slower than $\sqrt{n}$-rates. In general, we expect the margin condition to be weakest when the instrument is sharper, that is, α is likely larger

for sharper instruments, and smaller for more blunt instruments, since then γ is more flat and likely puts more mass around the quantile q. In Appendix F [26], we consider some examples, and illustrate for which α values the condition holds. An alternative approach would be to avoid the margin condition by instead targeting a smooth approximation of the nonsmooth functional $\beta_j(h_q)$, for example, in the same spirit as Kennedy et al. [29], or smooth but wider bounds.

Importantly, under the conditions of Theorem 5 that ensure asymptotic normality, one can use the approach of Imbens and Manski [22] to construct valid confidence intervals for the partially identified effect $\beta(h_q)$. We implement this in the `npcausal` R package.

5. Summarizing sharpness. So far we have discussed two primary features that make an instrument sharp: accurate prediction of compliers, and tight bounds on effects in identifiable subgroups. In this section, we present a new summary measure of sharpness that captures these two properties, separate and apart from strength. We characterize how this measure is related to the complier classification error and bound length quantities from previous sections, and discuss efficient nonparametric estimation and inference. We suggest our sharpness measure be reported alongside strength in practice.

5.1. *Proposed measure and properties.* To summarize sharpness, we use the proportion of variance in the instrument's compliance explained by covariates, specifically that proportion explained by the highest compliance score values; this is equivalent to the correlation between the true and predicted compliance status. Although we view this measure as a (strength-independent) summary of how well one can predict compliance and obtain tight bounds on identifiable subgroup effects, we refer to it as sharpness for simplicity.

DEFINITION 1. The *sharpness* ψ of instrument Z with latent compliance indicator C and compliance score γ is defined as

$$\psi = \frac{\operatorname{cov}(C, h_q)}{\operatorname{var}(C)} = \operatorname{corr}(C, h_q),$$

where $h_q = \mathbb{1}\{\gamma(\mathbf{X}) > F^{-1}(1-\mu)\}$ is the quantile classifier defined in (3.2), which selects subjects with the top $100\mu\%$ compliance scores.

We will now give some motivation and intuition for our proposed sharpness measure. First, as a ratio of covariances, it is easily interpretable as a measure of variance explained. In particular, it represents the proportion of variation in compliance explained by the highest $100\mu\%$ compliance scores (it is in the unit interval when γ is continuously distributed). In this sense, it can be viewed as a model-free and population version of a classical R^2 measure, indicating to what extent compliance can be predicted by covariates (through the compliance score). In fact, sharpness is also the slope of a population regression of compliance C on predicted compliance h_q. At one extreme, if the highest compliance scores do not predict compliance at all, that is, $C \perp\!\!\!\perp h_q$ (say if $\gamma \approx 0.5$ so that C is just a coin flip), then the sharpness measure is zero. Conversely, if compliance is perfectly predictable, that is, $C = \mathbb{1}(\gamma > q) = h_q$, then sharpness is one. For the toy example in Figure 1, the sharpness is 0%, 40% and 100% for instruments 1, 2 and 3, respectively.

One could substitute other classifiers for h_q and redefine sharpness as $\psi(h) = \operatorname{cov}(C, h)/\operatorname{var}(C)$ for some other $h : \mathcal{X} \mapsto \{0, 1\}$, such as h_0 or h_s discussed in Section 3. We focus on h_q for three main reasons: first, it is optimal among classifiers with size μ, that is,

$$\operatorname*{arg\,min}_{h \in \mathcal{G}(\mu)} \mathcal{E}(h) = \operatorname*{arg\,min}_{h \in \mathcal{G}(\mu)} \ell(h) = \operatorname*{arg\,max}_{h \in \mathcal{G}(\mu)} \psi(h) = h_q.$$

Second, using the classifier h_q yields simple and interpretable relationships between $\mathcal{E}(h_q)$, $\ell(h_q)$, and ψ, as will be discussed shortly; and finally, the classifier has an easy interpretation as selecting the highest $100\mu\%$ of compliance scores.

The proposed sharpness measure is further interpretable since, for any classifier $h : \mathcal{X} \mapsto \{0, 1\}$, we show in Appendix G that

$$\frac{\operatorname{cov}(C, h)}{\operatorname{var}(C)} = \mathbb{P}(h = 1 \mid C = 1) - \mathbb{P}(h = 1 \mid C = 0).$$

Thus, in addition to measuring variance explained, sharpness also measures the difference between true positive and false positive rates. In particular, for the quantile classifier we have that

$$\psi = \mathbb{P}\big(\gamma > F^{-1}(1-\mu) \mid C = 1\big) - \mathbb{P}\big(\gamma > F^{-1}(1-\mu) \mid C = 0\big).$$

This quantity is typically called the Youden index and is a popular summary measure of classifier performance [15, 39].

One might question the additional benefits of reporting sharpness ψ, beyond just the classification error $\mathcal{E}(h)$ or bound length $\ell(h)$. One crucial feature of ψ is that, unlike say classification error $\mathcal{E}(h)$, it is formally separate from instrument strength μ, in the sense of variation independence. This means, for example, that ψ—unlike $\mathcal{E}(h)$—cannot be small solely due to instrument strength (or lack thereof). As an illustrative example, consider an instrument for which $\gamma = 0.05$ with probability one. Then the optimal classifier in terms of prediction error is given by $h_0 = 0$, and this uninteresting rule classifies 95% of subjects correctly (an impressive error rate). However, this instrument has zero sharpness in the intuitive sense of the motivating example from Section 1, and this fact is not reflected by the classification error. In particular, with respect to both classification error and strength, the instrument with $\gamma = 0.05$ is virtually indistinguishable from one with $\gamma = \Phi(-2.7 + 1.4x)$ for $X \sim N(0, 1)$. Both yield approximately 5% classification error and strength, but in the latter case more information is available: we know that subjects with larger x values are more likely to be compliers; in fact, we have $\psi = \operatorname{cov}(C, h_q)/\operatorname{var}(C) \approx 50\%$ for the second instrument, compared to $\psi = 0$ for the first.

More formally, sharpness and strength are truly separate measures in the sense that they are variation independent in the presence of a continuous covariate. In particular, for an instrument with any given strength $\mu \in [\epsilon, 1-\epsilon]$ we can construct a congenial compliance score γ with any sharpness value $\psi \in [0, 1]$; conversely, for an instrument with any given sharpness $\psi \in [0, 1]$ we can construct a congenial compliance score with any strength value $\mu \in [\epsilon, 1-\epsilon]$. For example, suppose without loss of generality that $X \sim N(0, 1)$, which can be satisfied for any continuous covariate X^* with cumulative distribution function G via the transformation $X = \Phi^{-1}\{G(X^*)\}$ for Φ the $N(0, 1)$ distribution function. Then for $\gamma(x) = \Phi(b_0 + b_1 x)$ we can always find particular (b_0, b_1) values to satisfy $\mathbb{E}(C) = \mu$ and $\operatorname{cov}(C, h_q) = \psi\mu(1-\mu)$ for any $(\mu, \psi) \in [\epsilon, 1-\epsilon]^2$. For the case where $\psi = 0$ or $\psi = 1$, we can simply take $\gamma = \mu$ and $\gamma = \mathbb{1}\{x > \Phi^{-1}(1-\mu)\}$, respectively. More details are given in Section E of the Appendix, along with a plot to illustrate the bijective relationship between (μ, ψ) and (b_0, b_1).

Although compliance status C is not directly observed, sharpness is still identified under usual instrumental variable assumptions, simply because the compliance score is identified.

PROPOSITION 4. *Under Assumptions 1–3 and 5, sharpness is identified as*

$$\psi = \mathbb{E}\big\{\gamma(\mathbf{X})h_q(\mathbf{X}) - \mu^2\big\}/\mu(1-\mu).$$

Proposition 4 follows easily from the definition of sharpness together with the fact that $\mathbb{E}(C \mid \mathbf{X}) = \gamma(\mathbf{X})$, and is of course critical for constructing estimators of sharpness from observed data, which will be presented in the next subsection.

Having defined, motivated and identified the sharpness measure ψ, we now turn to characterizing its relation to classification error and bound length. The next result shows that, keeping strength fixed, sharper instruments yield more accurate complier classification and tighter bounds on identifiable subgroup effects.

THEOREM 6. *The classification error $\mathcal{E}(h_q)$ and bound length $\ell(h_q)$ can be expressed in terms of strength μ and sharpness ψ as*

$$\mathcal{E}(h_q) = 2\mu(1-\mu)(1-\psi),$$
$$\ell(h_q) = (1-\mu)(1-\psi).$$

The theorem indicates the precise relationship between complier classification error, bound length, strength and sharpness for h_q. The result follows from the fact that, defining $\psi(h) = \operatorname{cov}(C,h)/\operatorname{var}(C)$ for general classifiers h, we have

$$\mathcal{E}(h) = 2\mu(1-\mu)\{1-\psi(h)\} + (1-2\mu)(\mathbb{E}h - \mu),$$
$$\ell(h) = (1-\mu)\{1-\mu\psi(h)/\mathbb{E}h\}$$

together with the fact that $\mathbb{E}(h_q) = \mu$. Theorem 6 has several important consequences. First, it shows that strength and sharpness are fundamental aspects of the quality of an instrument, since together they completely determine the best error for classifying compliers and the tightest bounds on identifiable subgroup effects, among all classifiers/subgroups of size μ. It also shows that for fixed strength, sharper instruments yield better complier classification and tighter bounds on identifiable subgroup effects. As expected, perfect complier prediction $\mathcal{E}(h_q) = 0$ and point identification $\ell(h_q) = 0$ requires perfect sharpness $\psi = 1$ (note we must have $\mu \leq 1 - \epsilon$ because if $\mu = 1$ then $A = Z$, which means the instrument cannot be unconfounded if the treatment is confounded).

Of more practical relevance, Theorem 6 also shows that nonzero sharpness is an important sufficient condition for better complier prediction and tighter bounds on identifiable subgroup effects. Focusing first on complier prediction, we observe that if $\psi > 0$ then there exists a classifier that attains better error than the naive strength-calibrated classifier (which simply flips a coin with probability μ). This follows since if $\psi > 0$ then $\mathcal{E}(h_q) < 2\mu(1-\mu)$, which is the error of the rule $h \sim \text{Bern}(\mu)$. Further, since the classifier h_q attains a better error than the coin flip rule, then h_0 does as well, since the error of h_q is a lower bound for the latter. Turning our attention to bound lengths, we note that if $\psi > 0$ then there exists an identifiable subgroup (of size μ) yielding tighter bounds than those on the average treatment effect. This follows since nonzero sharpness $\psi > 0$ implies $\ell(h_q) < 1 - \mu$, which is the length of the bounds on the average treatment effect $\mathbb{E}(Y^{a=1} - Y^{a=0})$ as derived, for example, by Balke and Pearl [6], Manski [31], Robins [35]. The size of ψ indicates the percent reduction in the length of the bounds, for example, bounds on the subgroup effect $\beta(h_q)$ are precisely $100\psi\%$ tighter than those on the average treatment effect. The only way tighter bounds could be obtained would be to consider smaller subgroups.

In summary, the sharpness measure proposed in Definition 1 captures the proportion of variance in an instrument's compliance explained by the highest compliance scores. It is an interpretable and strength-independent reflection of (i) how accurately compliers can be classified and (ii) how tightly effects in identifiable subgroups can be bounded. We suggest that it be reported alongside strength in instrumental variable analyses; in the next subsection, we propose methods for estimation and inference.

5.2. *Estimation and inference*. Here, we propose an estimator for sharpness ψ that, like estimators from previous sections, uses influence functions to correct bias from nonparametric nuisance estimation and incorporates sample splitting to avoid empirical process restrictions. We refer back to Section 4.2 for more details and notation.

Our sharpness estimator relies on the strength estimator $\widehat{\mu} = \mathbb{P}_n\{\phi_\mu(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{\text{-}B})\}$ from Section 4, as well as an estimator $\widehat{\xi} = \mathbb{P}_n(\widehat{\phi}_{\xi,\text{-}B})$ of

$$\xi = \mathbb{E}(\gamma h_q),$$

where $\phi_\xi = \phi_\mu(\mathbf{O}; \boldsymbol{\eta}) h_q(\mathbf{X})$ and $\widehat{\phi}_{\xi,\text{-}b} = \phi_\mu(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{\text{-}b}) \widehat{h}_{q,\text{-}b}(\mathbf{X})$ are the corresponding influence function for ξ and its estimate. In particular, we estimate sharpness as

$$\widehat{\psi} = \frac{(\widehat{\xi} - \widehat{\mu}^2)}{\widehat{\mu}(1 - \widehat{\mu})}$$

which appropriately combines influence-function-based estimators of the corresponding terms from Definition 1 (i.e., the numerator is the estimator of the covariance between compliance C and the classifier h_q). To concisely state our results, we define the remainder terms:

$$R_{1,n} = \|\widehat{\pi}_1 - \pi_1\| \Big(\max_z \|\widehat{\lambda}_z - \lambda_z\| \Big),$$

$$R_{2,n} = \big(\|\widehat{\gamma} - \gamma\|_\infty + |\widehat{q} - q| \big)^{1+\alpha},$$

where once again $\alpha > 0$ is the margin exponent (see (3.7)). We note that in comparison with the remainder in (4.5) for the estimation of subgroup effects the remainder here for the estimation of sharpness is of lower order, that is, we are able to estimate sharpness at much faster rates. With these definitions in place, the next theorem gives corresponding convergence rates, as well as conditions under which $\widehat{\psi}$ is asymptotically normal and efficient.

THEOREM 7. *Assume that* $\mathbb{P}\{\epsilon \le \widehat{\pi}_z(\mathbf{X}) \le 1 - \epsilon\} = 1$ *for* $z = 0, 1$ *and some* $\epsilon > 0$, *and* $\|\widehat{\pi}_1 - \pi_1\| + \max_z \|\widehat{\lambda}_z - \lambda_z\| + \mathbb{P}(\widehat{h}_q \neq h_q) = o_{\mathbb{P}}(1)$.

1. *If the margin condition holds for some* $\alpha > 0$, *then*

$$\widehat{\psi} - \psi = O_{\mathbb{P}}\Big(\frac{1}{\sqrt{n}} + R_{1,n} + R_{2,n} \Big).$$

2. *If it also holds that* $R_{1,n} + R_{2,n} = o_{\mathbb{P}}(1/\sqrt{n})$, *then*

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N\Big(0, \operatorname{var}\Big[\frac{\{\phi_\mu h_q + q(\phi_\mu - h_q) - \xi\}}{(\mu - \mu^2)} + \frac{(2\mu\xi - \xi - \mu^2)}{(\mu - \mu^2)^2}(\phi_\mu - \mu) \Big]\Big).$$

Theorem 7 gives two main results. First, it shows that the proposed sharpness estimator is consistent with convergence rate that is second-order in nuisance estimation errors, under weak conditions (bounded IV propensity scores, consistent nuisance estimators and the margin condition). This means $\widehat{\psi}$ attains faster rates than those of its nuisance estimators, which comes from using influence functions for better bias correction than a general plug-in. We do not require any complexity or empirical process conditions, since we use sample splitting to separate the evaluation and estimation of the influence function. Second, Theorem 7 shows that if the second-order nuisance errors converge to zero at a faster than $\sqrt{n}$ rate, the estimator is asymptotically normal, and efficient by virtue of the fact that we are working in a nonparametric model (where the only influence function is the efficient one). This condition on the nuisance estimation is satisfied, for example, if $\alpha = 1$ and the nuisance estimators

converge at faster than $n^{1/4}$ rates; this can hold under nonparametric smoothness, sparsity or other structural conditions.

The asymptotic variance in the second part of Theorem 7 can be easily estimated with its corresponding plug-in, from which Wald-type confidence intervals can be constructed. Since such intervals may go outside the unit interval, we give an improved logit-transformed interval in Corollary 5 in the Appendix (which is implemented in the `npcausal` R package).

6. Simulations and illustration. In this section, we report the results of various simulations we performed to illustrate the finite-sample performance of our proposed estimators. We also analyze data from a study of canvassing effects on voter turnout [17] and study the sharpness of the instrument and explore some of its consequences.

6.1. *Simulation study.* To assess finite-sample performance, we considered simulations from the following model:

$$\begin{aligned} X &\sim N(0,1), \qquad C \mid X \sim \text{Bern}(\gamma) \quad \text{for } \gamma(x) = \Phi(b_0 + b_1 x), \\ Z \mid X, C &\sim \text{Bern}(\pi_1) \quad \text{for } \pi_1(x) = \text{expit}(x), \\ A &= CZ + (1-C)A^* \quad \text{for } A^* \mid X, C, Z \sim \text{Bern}(0.5), \\ Y &= AY^{a=1} + (1-A)Y^{a=0} \\ &\qquad \text{for } Y^a \mid X, C, Z, A \sim \text{Bern}\big(0.5 + (a-0.5)\beta\big), \end{aligned}$$

with (b_0, b_1) chosen to ensure given values (μ, ψ) of strength $\mu = 30\%$ and sharpness as detailed in Appendix E. This model satisfies Assumptions 1–5 and implies

$$\mathbb{E}\big(Y^{a=1} - Y^{a=0}\big) = \mathbb{E}\big(Y^{a=1} - Y^{a=0} \mid h_q\big) = \beta. \tag{6.1}$$

We used the proposed methods to classify compliers and estimate sharpness and bounds. Nuisance functions were estimated with correctly specified logistic regression models, with $K = 2$ sample splits. To assess performance, we used empirical error $\mathbb{P}_n(\widehat{h} \neq C)$ for each classifier; length of estimated bounds for parameters (6.1) with $\beta = 20\%$; and bias, RMSE and 95% CI coverage of the sharpness estimator. All code is in Appendix 1.

The simulations illustrate what our theory predicts, as illustrated in Table 1. Instruments with the same strength can yield drastically different complier classification error (between

TABLE 1
Simulation results across 500 *simulations* (*all figures are percentages*)

	Class. error			Bound length		Sharpness est.		
Setting	$\widehat{h}_0$	$\widehat{h}_q$	$\widehat{h}_s$	ATE	$\beta(h_q)$	Bias	SE	Cov
$n = 500$:								
$\psi = 0.2$	30.9	36.7	39.9	68.8	61.2	−9.4	13.5	96.9
$\psi = 0.5$	21.2	22.1	29.2	69.9	36.2	−1.4	13.9	98.2
$\psi = 0.8$	8.5	9.3	14.3	70.4	13.9	0.1	10.3	95.8
$n = 1000$:								
$\psi = 0.2$	30.0	35.1	39.6	70.1	59.7	−3.7	10.3	97.0
$\psi = 0.5$	20.6	21.4	28.9	69.9	35.4	−0.4	8.0	95.2
$\psi = 0.8$	8.4	8.8	13.6	70.1	14.4	−0.4	7.0	95.0
$n = 5000$:								
$\psi = 0.2$	29.6	33.7	39.4	70.0	56.4	−0.4	3.6	95.8
$\psi = 0.5$	20.5	21.0	28.1	70.1	34.9	0.4	3.1	95.2
$\psi = 0.8$	8.4	8.5	12.6	70.0	14.1	−0.1	3.1	94.6

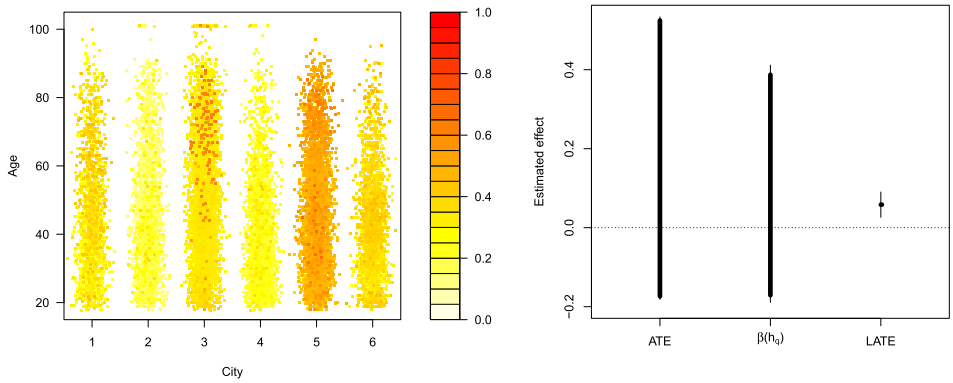
39.9% to 8.4% here) and subgroup effect bound lengths (between 13.9% to 70.4%) depending on sharpness. Our proposed sharpness estimator has minimal bias decreasing with sample size and confidence intervals attain nominal coverage (coverage was at least 95% for all bound estimators).

6.2. *Data analysis*. Here, we illustrate the proposed methods by analyzing data from a study of canvassing effects on voter turnout. Green et al. [17] conducted a study of $n = 18{,}933$ voters across six cities who were randomly assigned to receive encouragement to vote in local elections or not. Recall we are using an i.i.d. assumption; inference without this assumption is an important avenue of future research. Noncompliance arose since some voters who were assigned to receive encouragement could not be contacted. As a result, Green et al. [17] estimated the complier average effect, where here compliers are those people who would be encouraged only when assigned to be. Aronow and Carnegie [3] argue that the local estimand is of limited interest, since in this study compliance is less an inherent characteristic, and more a feature of the design and could change over time (e.g., multiple contacts could increase compliance). Thus it is of interest to identify compliers based on observed characteristics, so as to better generalize the study results by understanding to which subpopulation the effect corresponds.

In this study, the measured covariates include city indicators (Bridgeport, Columbus, Detroit, Minneapolis, Raleigh, St. Paul), party affiliation, prior voting history, age, family size, race and corresponding missingness indicators. We use our proposed methods to classify compliers, estimate bounds on the average treatment effect as well as the subgroup effect $\beta(h_q)$ and assess sharpness of the instrument (i.e., initial randomization). We used random forests (via the `ranger` R package) to estimate the nuisance functions with $K = 2$-fold sample splitting.

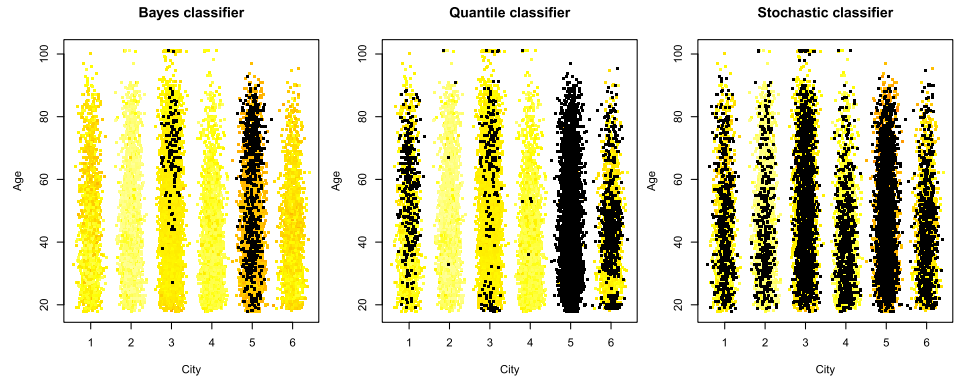
In Figures 2(a) and 2(c), we present estimated compliance scores and results from the three proposed complier classification methods, respectively. In both cases, we plot the voter's estimated compliance scores against two important covariates: the voter's city and age. The estimated compliance scores ranged from 8% to 69% across the voter population. Overall, the results indicate that the set of compliers is very likely to contain people from Raleigh (city 5), across a range of ages, as well as older voters in Detroit (city 3). Relative to the estimated Bayes classifier $\widehat{h}_0$, the quantile classifier $\widehat{h}_q$ classifies more voters as compliers (30% versus only 4%), mostly from Raleigh but also from Bridgeport and St. Paul. With the stochastic classifier $\widehat{h}_s$, it is somewhat more difficult to distinguish predicted compliers from the rest, based on city and age; however, one can still clearly see overrepresentation in Raleigh and St. Paul. The estimated error of the quantile classifier is $2\widehat{\mu}(1-\widehat{\mu})(1-\widehat{\psi}) = 33.3\%$, which yields bounds $27.2\% \pm 6.1\% = [21.1\%, 33.3\%]$ on the optimal error $\mathcal{E}(h_0)$ from Proposition 2.

Our nonparametric doubly robust analysis yielded an estimated local effect very similar to that of Green et al. [17] (5.7%, 95% confidence interval (CI): 2.5%–8.9%). However, we estimate that the instrument in this study was stronger than it was sharp, yielding $\widehat{\mu} = 30.1\%$ (95% CI: 29.2%–31.1%) but $\widehat{\psi} = 20.9\%$ (95% CI: 18.8%–23.2%). Figure 2(b) shows the estimated local effect, along with bounds on the average treatment effect and subgroup effect $\beta(h_q)$; we used the Imbens and Manski [22] approach to construct confidence intervals for the subgroup effects. Although the bounds for the subgroup effect are narrower than for the average treatment effect, they are still relatively wide due to the instrument not being very sharp. In particular, the bounds on $\beta(h_q)$ are 20% narrower than for the average treatment effect, but still cover zero; the estimated bounds on $\beta(h_q)$ are $[-17.1\%, 38.7\%]$ with 95% CI $[-18.9\%, 41.2\%]$.



(a) Estimated compliance scores

(b) Estimated effects



(c) Complier classification results, predicted compliers are marked in black.

FIG. 2. *Results from analysis of Green et al.* [17] *study of canvassing effects.*

7. Discussion. In this paper, we introduce a new measure of instrument quality, called sharpness, which measures the variation in an instrument's compliance explained by the covariates (in particular, by the compliance scores), and which reflects how well one can predict who compliers are and how tightly one can bound effects in identifiable subgroups. We propose complier classification rules and characterize their large-sample errors, as well as novel effects in identifiable subgroups defined by subjects with the highest compliance scores. We discuss nonparametric methods for estimating all of these quantities (classification rules, bounds and sharpness) and give general rates of convergence results, as well as conditions under which the methods are efficient. Finally, we have studied the methods via simulation, and applied them in a study of canvassing effects on voter turnout. Implementations of all our methods are publicly available in the `npcausal` R package.

There are several caveats to mention and ways in which our work could be generalized. Although we have allowed for complex covariate information, we have focused on the relatively simple setting where both the instrument and treatment are binary. The binary instrument restriction can be removed without changing the estimands and methods too much (although some nontrivial statistical complications could result, as noted, e.g., in Kennedy et al. [28]). A multivalued treatment, however, prevents (nonparametric) identification of the compliance score and even local treatment effects; therefore, removing this restriction would necessitate a substantially different approach, for example, involving estimands that are only partially identified without further assumptions. The same goes for removing the monotonicity restriction, a lack of which also prevents nonparametric identification. Although the binary/monotonic setup we consider here is widely used, it would be useful in future work to consider analogs

of sharpness for different instrumental variable models, such as those of Robins [36], Tan [43] that replace monotonicity with effect homogeneity restrictions. It would also be worthwhile to consider violations of the instrumental variable assumptions [6, 23] might affect sharpness.

In practice, we propose that sharpness should be assessed in instrumental variable studies, alongside strength. Sharp instruments can help yield more generalizable causal effects (via better prediction of compliers, and tighter bounds on effects in identifiable subgroups), which has been a prominent concern with standard instrumental variable methods. Given the substantial benefits of sharp instruments, this work also suggests new strategies for data collection and study design. Namely, one should aim to collect data not only on covariates that explain instrument assignment (so as to de-confound the instrument-treatment/outcome relationships for Assumption 3), but also on covariates that predict subjects' compliance behavior. Further, sharpness provides another factor to consider when choosing among instruments, in cases where numerous IVs are available (e.g., in A/B test settings involving many experiments with noncompliance). Importantly, both sharpness and strength can be assessed without outcome data; so if such data collection is costly, one can decide where to collect data on the basis of sharpness and strength.

Acknowledgments. The authors thank the Johns Hopkins Causal Inference Working Group, Luke Keele, Betsy Ogburn, Jamie Robins, Fredrik Sävje and Larry Wasserman for helpful comments and discussions.

This work was supported in part by NSF Grant DMS-17130003.

SUPPLEMENTARY MATERIAL

Supplement to "Sharp instruments for classifying compliers and generalizing causal effects" (DOI: 10.1214/19-AOS1874SUPP; .pdf). Proofs, extra details and R code is given in the Supplementary Materials.

REFERENCES

[1] ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Econometrics* **113** 231–263. MR1960380 https://doi.org/10.1016/S0304-4076(02)00201-4

[2] ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

[3] ARONOW, P. M. and CARNEGIE, A. (2013). Beyond LATE: Estimation of the average treatment effect with an instrumental variable. *Polit. Anal.* **21** 492–506.

[4] AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

[5] BAIOCCHI, M., CHENG, J. and SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33** 2297–2340. MR3257582 https://doi.org/10.1002/sim.6128

[6] BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1176.

[7] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

[8] BOUND, J., JAEGER, D. A. and BAKER, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Amer. Statist. Assoc.* **90** 443–450.

[9] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2016). Double machine learning for treatment and causal parameters. ArXiv preprint. Available at arXiv:1608.00060.

[10] COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13** 21–27.

[11] DEATON, A. (2010). Instruments, randomization, and learning about development. *J. Econ. Lit.* **48** 424–455.

[12] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

[13] DING, P. and LU, J. (2017). Principal stratification analysis using principal scores. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 757–777. MR3641406 https://doi.org/10.1111/rssb.12191

[14] FELLER, A., MEALLI, F. and MIRATRIX, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *J. Educ. Behav. Stat.* **42** 726–758.

[15] FLUSS, R., FARAGGI, D. and REISER, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biom. J.* **47** 458–472. MR2190469 https://doi.org/10.1002/bimj.200410135

[16] FOLLMANN, D. A. (2000). On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *J. Amer. Statist. Assoc.* **95** 1101–1109. MR1821718 https://doi.org/10.2307/2669746

[17] GREEN, D. P., GERBER, A. S. and NICKERSON, D. W. (2003). Getting out the vote in local elections: Results from six door-to-door canvassing experiments. *J. Polit.* **65** 1083–1096.

[18] HERNÁN, M. A. and ROBINS, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17** 360–372.

[19] IMBENS, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.* **48** 399–423.

[20] IMBENS, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statist. Sci.* **29** 323–358. MR3264545 https://doi.org/10.1214/14-STS480

[21] IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

[22] IMBENS, G. W. and MANSKI, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72** 1845–1857. MR2095534 https://doi.org/10.1111/j.1468-0262.2004.00555.x

[23] IMBENS, G. W. and RUBIN, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models **64** 555–574. MR1485828 https://doi.org/10.2307/2971731

[24] JO, B. and STUART, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Stat. Med.* **28** 2857–2875. MR2750169 https://doi.org/10.1002/sim.3669

[25] JOFFE, M. M., TEN HAVE, T. R. and BRENSINGER, C. (2003). The compliance score as a regressor in randomized trials. *Biostatistics* **4** 327–340.

[26] KENNEDY, E. H, BALAKRISHNAN, S. and G'SELL, M. (2020). Supplement to "Sharp instruments for classifying compliers and generalizing causal effects." https://doi.org/10.1214/19-AOS1874SUPP.

[27] KENNEDY, E. H., HARRIS, S. and KEELE, L. J. (2019). Survivor-complier effects in the presence of selection on treatment, with application to a study of prompt ICU admission. *J. Amer. Statist. Assoc.* **114** 93–104. MR3941240 https://doi.org/10.1080/01621459.2018.1469990

[28] KENNEDY, E. H., LORCH, S. and SMALL, D. S. (2019). Robust causal inference with continuous instruments using the local instrumental variable curve. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 121–143. MR3904782

[29] KENNEDY, E. H., MA, Z., MCHUGH, M. D. and SMALL, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1229–1245. MR3689316 https://doi.org/10.1111/rssb.12212

[30] LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

[31] MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev.* **80** 319–323.

[32] MARCUS, S. M., STUART, E. A., WANG, P., SHADISH, W. R. and STEINER, P. M. (2012). Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychol. Methods* **17** 244.

[33] PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

[34] ROBINS, J. LI, L., TCHETGEN TCHETGEN, E., and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. Probability and Statistics: Essays in Honor of David A. Freedman 335–421.

[35] ROBINS, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology*: *A Focus on AIDS* 113–159.

[36] ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185 https://doi.org/10.1080/03610929408831393

[37] ROBINS, J. M. and GREENLAND, S. (1996). Comment: Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 456–458.

[38] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

[39] SCHISTERMAN, E. F., PERKINS, N. J., LIU, A. and BONDELL, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* **16** 73–81.

[40] STAIGER, D. and STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. MR1445622 https://doi.org/10.2307/2171753

[41] STUART, E. A. and JO, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Stat. Methods Med. Res.* **24** 657–674. MR3428422 https://doi.org/10.1177/0962280211421840

[42] SWANSON, S. A. and HERNÁN, M. A. (2014). Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation [discussion of MR3264545]. *Statist. Sci.* **29** 371–374. MR3264549 https://doi.org/10.1214/14-STS491

[43] TAN, Z. (2010). Marginal and nested structural models using instrumental variables. *J. Amer. Statist. Assoc.* **105** 157–169. MR2757199 https://doi.org/10.1198/jasa.2009.tm08299

[44] TCHETGEN TCHETGEN, E. J. and VANSTEELANDT, S. (2013). Alternative identification and inference for the effect of treatment on the treated with an instrumental variable. In *Harvard University Biostatistics Working Paper Series*, *Paper* 166.

[45] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data. Springer Series in Statistics*. Springer, New York. MR2233926

[46] VAN DER LAAN, M. J. and LUEDTKE, A. R. (2013). Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. *UC Berkeley Division of Biostatistics Working Paper Series* **329** 1–96.

[47] VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York.

[48] VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1999). *Lecture Notes in Math.* **1781** 331–457. Springer, Berlin. MR1915446

[49] WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Stat.* **5** 161–215.

[50] ZHENG, W. and VAN DER LAAN, M. J. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series* **273** 1–58.