ELSEVIER
Journal of Clinical Epidemiology xxx (xxxx) xxx
Journal of Clinical Epidemiology

REVIEW

Preference-based instrumental variables in health research rely on important and underreported assumptions: a systematic review

Tarjei Widding-Havneraas[a,b,*], Ashmita Chaulagain[a,b], Ingvild Lyhmann[a,b], Henrik Daae Zachrisson[c], Felix Elwert[d], Simen Markussen[e], David McDaid[f], Arnstein Mykletun[a,g,h,i]

[a] *Centre for Research and Education in Forensic Psychiatry, Haukeland University Hospital, Bergen, Norway*
[b] *Department of Clinical Medicine, University of Bergen, Bergen, Norway*
[c] *Department of Special Needs Education, University of Oslo, Oslo, Norway*
[d] *Department of Sociology, University of Wisconsin-Madison, WI 53706, USA*
[e] *Ragnar Frisch Centre for Economic Research, Oslo, Norway*
[f] *Care Policy and Evaluation Centre, Department of Health Policy, London School of Economics and Political Science, London, UK*
[g] *Division of Health Services, Norwegian Institute of Public Health, Oslo, Norway*
[h] *Department of Community Medicine, UiT - The Arctic University of Norway, Tromsø, Norway*
[i] *Centre for Work and Mental Health, Nordland Hospital, Bodø, Norway*

Accepted 3 June 2021

Abstract

Objective: Preference-based instrumental variables (PP IV) designs can identify causal effects when patients receive treatment due to variation in providers' treatment preference. We offer a systematic review and methodological assessment of PP IV applications in health research.

Study Design and Setting: We included studies that applied PP IV for evaluation of any treatment in any population in health research (PROSPERO: CRD42020165014). We searched within four databases (Medline, Web of Science, ScienceDirect, SpringerLink) and four journals (including full-text and title and abstract sources) between January 1, 1998, and March 5, 2020. We extracted data on areas of applications and methodology, including assumptions using Swanson and Hernan's (2013) guideline.

Results: We included 185 of 1087 identified studies. The use of PP IV has increased, being predominantly used for treatment effects in cancer, cardiovascular disease, and mental health. The most common PP IV was treatment variation at the facility-level, followed by physician- and regional-level. Only 12 percent of applications report the four main assumptions for PP IV. Selection on treatment may be a potential issue in 46 percent of studies.

Conclusion: The assumptions of PP IV are not sufficiently reported in existing work. PP IV-studies should use reporting guidelines.

Keywords: Causal Inference; Quasi-Experimental Methods; Instrumental Variables; Provider-Preference; Comparative Effectiveness; Systematic Review

1. Introduction

Instrumental variables (IV) is a quasi-experimental method that can overcome unobserved confounding under suitable conditions [1],[2]. Originally developed in economics [3–7] it is becoming a popular method for evaluating causal effects in health research [8–12]. An IV is a variable that induces random variation in treatment, which can be used to identify treatment effects. Randomization in a double-blind randomized controlled trial (RCT) is often presented as the ideal IV [2], with considerable subject matter knowledge usually required to find/defend any IVs [13].

Provider preference IV (PP IV) designs use variation in clinical practice patterns as an IV. The PP IV premise is that variation in practice patterns reflect underlying provider treatment preferences that induce, from the pa-

Funding: Research Council of Norway (288585/IAR) and Western Norway Regional Health Authority (912197). The funders had no role in this study.

Competing interest: None.

* Corresponding author. Tel: +47 92 45 14 87. Fax: +47 55 97 33 21.
E-mail address: tarjei.widding-havneras@uib.no (T. Widding-Havneraas).

https://doi.org/10.1016/j.jclinepi.2021.06.006

What is new?

Key findings

PP IV methods are increasingly used in health research, and across specialties. PP IV methods can estimate treatment effects where RCTs are not feasible due to ethical or practical problems with randomization. Few applications of PP IV report all four main identifying assumptions.

What this adds to what is known?

This review provides an overview of applications of PP IV with novel data on clinical and academic area, reporting of assumptions, and potential selection on treatment bias.

What should change now?

Researchers should be more transparent in reporting assumptions when using PP as an IV and pay more attention to potential bias introduced by selecting on treatment.

tient's perspective, random variation in patients' treatment status, emulating a randomized trial for a subset of patients [14]. Korn and Baumrind [15] first proposed the use of variation in individual physicians' preference for specific treatments as an IV. IV relies on four assumptions, which for consistency we number according to Swanson and Hernán's [16] IV reporting guideline: IV must (A1) predict treatment status ("relevance"), (A2) affect outcome only through exposure ("exclusion"), and (A3) not share any unmeasured causes with the outcome ("unconfoundedness"). A fourth assumption is that treatment effects are either constant (A4c), homogeneous (A4h) or monotonic, i.e., the IV only affects treatment status in one direction (A4m).

The broader term "preference-based IV" was introduced by Brookhart and Schneeweiss [14] who specify PP IV studies assume (1) between-provider variation in use of treatments, (2) patient selection or assignment to providers is unrelated to providers' treatment preference, and (3) providers' use of one treatment is independent of use of alternative treatments that affect outcomes. Consequently, the promise of circumventing unobserved confounding with PP IV is dependent on important assumptions [14].

Provider preference is difficult to measure directly. Researchers often measure latent provider preference as the proportion of patients that receive treatment of interest or, in pharmacological studies, prescriptions issued before current prescription at physician, facility, or geographical region levels. Table 1 presents examples of common PP IV designs.

1.1. Estimands and interpretation

Suppose we are interested in the effect of a specific medication ($D = 1$), with treatment as usual as the alternative ($D = 0$), on mortality (Y). In observational data, treatment receipt is likely correlated with unobserved risk factors of the outcome, such as patient presentation, causing confounding bias. Random variation in provider preference (PP) for $D = 1$ over $D = 0$, e.g., measured as the physician's last prescription before current prescription [17], can be considered as an IV. The intuition behind IV is that random variation in PP represents a natural source of randomization in D. By isolating variation in D entirely due to PP, IV can be used to identify causal effects [1],[20]. Under A1-A3 and constant (A4c) or homogenous (A4h) effects the IV-estimator is consistent for the average treatment effect (ATE) [21]. When treatment effects vary over patients, e.g. stronger side effects experienced by older patients, and both IV and treatment are binary, then, under A1-A3 and A4m, IV estimates the local average treatment effect (LATE) for a latent subpopulation of "compliers," who always take the treatment that corresponds to the provider's preference [6]. A caveat with PP IV designs is that complier interpretation is complicated because patients may comply to varying extents [14]. There is substantial debate on the clinical and policy relevance of the LATE estimand [22],[23] relating to the validity and interpretation of PP IV under various constraints, to which we turn next.

1.2. Validity

Several scenarios can give rise to a valid PP IV. Figure 1 displays data-generating models, directed acyclic graphs (DAG) [24],[25], to illustrate common scenarios and challenges. Figure 1a, b and c provide valid PP IVs, while Figure 1d, e and f present PP IV violations. In all models, treatment, D, and outcome, Y, are confounded by unobserved variables, U, preventing identification of the causal effect of D on Y by covariate adjustment (e.g., regression, matching, or weighting), motivating the search for a suitable IV.

Figure 1a illustrates a model where PP does not share unobserved risk factors with Y (unconfoundedness), directly affects D (relevance), and does not affect Y via any mechanism other than via D (exclusion). This is often considered the best-case scenario for IV estimation. In Figure 1b PP is a so-called proxy IV, such as the provider's manifest (measurable) prescribing behavior, which is affected by the unobserved true provider-preference, U_{PP}, but does not itself cause D. Most PP IV designs follow the proxy design; it is difficult to elicit providers' true preferences [8]. Interpretation of proxy IV designs is difficult. U_{PP} is a continous variable, while PP is typically measured as a binary proxy. If treatment effects

Please cite this article as: T. Widding-Havneraas et al., Preference-based instrumental variables in health research rely on important and underreported assumptions: A systematic review, Journal of Clinical Epidemiology, https://doi.org/10.1016/j.jclinepi.2021.06.006

Table 1. Common examples of provider preference instrumental variables designs: Physician, facility and regional. Examples correspond to physician [17], facility [18], and regional [19] level PP IVs. The physician example is a binary PP IV representing the instantaneous provider preference, alternatively prescribing patterns can be averaged over time

Authors	Study topic	Instrumental variable
Wang et al. [17]	Effect of conventional vs. atypical antipsychotic medications (APM) on short-term mortality risk among elderly users.	Physician's preference for prescribing an atypical APM instead of conventional APM measured as the most recent APM prescription before the current prescription.
Dalsgaard et al. [18]	Effect of early ADHD medication on contacts with hospitals, emergency ward, and police, among children diagnosed with ADHD.	Facility variation in propensity to prescribe medication measured as the share of other treated children in the same cohort diagnosed at the same facility.
Emdin et al. [19]	Effect of referral to cardiology follow-up on post-discharge mortality among patients with systolic heart failure.	Regional variation in referral to cardiology follow-up defined as the proportion of patients referred to follow-up within a region.

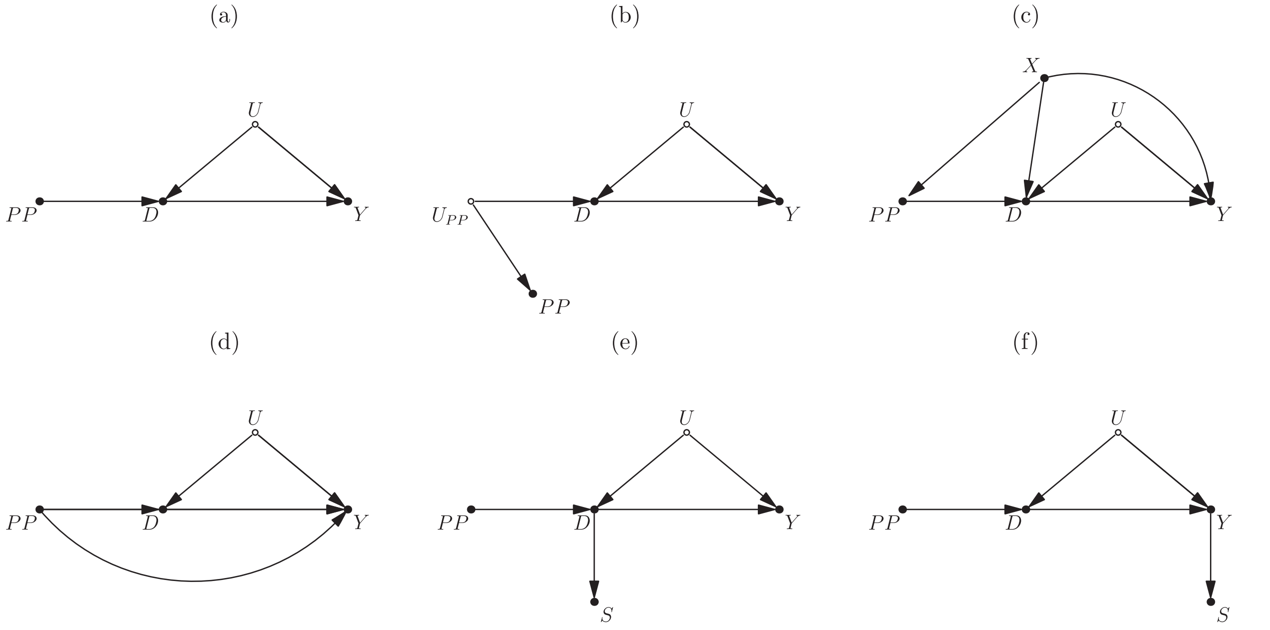

Figure 1. Models involving provider-preferences. PP is the measured provider-preference. U_{PP} is a latent (underlying) provider-preference. D is treatment. Y is the outcome. U are unobserved confounders. (a) PP with a valid causal IV. (b) PP is a valid proxy IV. (c) PP is a valid IV when X is controlled. (d) PP is not valid as the IV directly affects Y. (e) PP is not a valid IV if the sample is selected on S as a function of D. (f) PP is not valid if the sample is selected on S as a function of Y (such selection only leads to bias if there is an effect of D on Y).

are heterogeneous, this design will typically recover some weighted average of heterogenous treatment effects [2].

In Figure 1c PP is not random, as D and Y are confounded by X, e.g., shared patient and/or provider characteristics. Hence, all such confounders must be controlled for to meet A3. There is evidence of insufficient adjustment in PP IV applications [9]. When covariates are included, standard IV estimators obtain a variance-weighted average of covariate-specific LATEs [26] that may, however, be transformed into an unweighted LATE [27].

Figure 1d includes direct effects of PP on Y, which violates A3 and may occur if providers' preference for one treatment over another also leads them to treat patients differently in other ways [28]. In Figure 1e, sample selection, S, is a function of treatment, which violates A3. This occurs, for example, when a study restricts analyses to a subset of treatment options when more options are available [21],[29]. In this scenario, sample selection conditions on the descendant of a collider on the path $PP \rightarrow D \leftarrow U \rightarrow Y$, thereby opening a non-causal pathway between PP and Y [30–33]. Selection on the treatment can also occur in other models [21],[30],[32],[34],[35]. In Figure 1f, PP is not a valid IV as sample selection, S, is a descendant of Y [32]. For example, this bias may occur a study on the effects of ADHD medication on employment includes employed and unemployed but excludes non-employed people. In addition to scenarios in Figure 1, studies show (1) monotonicity (A4m) is unlikely to hold in certain PP IV applications [36]; (2) PP IVs can be biased when treatments are over/underused, as IV can over/underweight patients who may not need treatment in the former/latter case [14]; (3) IVs must be sufficiently strong to not induce weak IV bias ($F > 10$ in first stage regressions) [37], with recent work suggesting a considerably higher threshold [38].

PP IV is among the most commonly applied IVs in health research [8–10],[39], calling for scrutiny of current practice in view of recent studies pointing out potential issues with PP IV designs, including monotonicity [36] and bias from selecting on treatment [21]. Existing reviews either examine IVs in general [8–10,16,39] or are narrative PP IV reviews [40]. Here we contribute to the literature on PP IV through a systematic review focused on PP IV applications using a search strategy involving full-text mining [41], using databases that enable complete article text searches. We present novel data on applications, including academic and clinical areas, reporting of IV assumptions, potential bias from selection on treatment and strength of various PP IV definitions. The review's aim is to (i) provide an introduction to PP IV, (ii) systematically review applications of PP IV in health research, and (iii) evaluate current practice with PP IVs.

2. Methods

This systematic review adhered to the Preferred Reporting Item for Systematic Reviews and Meta-Analysis (PRISMA) guideline (Supplementary 1) [42] and is registered in PROSPERO (CRD42020165014).

2.1. Search strategy

We conducted a systematic search in ScienceDirect (full-text), SpringerLink (full-text), Medline (OVID) (title and abstract), and Web of Science (title and abstract). We had no language restrictions, but search words were restricted to English. As there are no pre-defined subject heading/keywords for PP IV, we also searched specific journals identified through database searches and prior knowledge of the literature: *American Journal of Epidemiology, International Journal of Epidemiology* (Oxford University Press Journals, full-text), *Health Economics* (Wiley Online Library, full-text), and *Epidemiology* (Wolters Kluwer, title and abstract). To identify additional relevant studies, we hand-searched reference lists of included studies. Key search words included "instrument* variable*", "provider", "physician", "prescribing", and "preference".

We combined all search results in EndNote X9 [43] and removed duplicates. All studies were imported into Covidence systematic review software [44] and remaining duplicates removed. Initial and full-text screening was conducted independently by two reviewers (TW and either AC or IL). Discrepancies regarding study inclusion were resolved through consensus. Search strategies and included study references are provided in Supplementary 2.

2.2. Eligibility

We included all empirical health research studies using quasi-experimental PP IV designs and real-world data. We defined PP IV as all applications where variation in treatment at either physician, facility or regional level are used as an IV to predict treatment status [9]. Eligible studies were peer-reviewed and used PP as an IV for any treatment in any population from the method's introduction (January 1, 1998) to last search date (March 5, 2020). We excluded all studies not applying PP as an IV in health research and studies only using simulated data.

2.3. Data extraction

A data extraction manual was developed for this review (Supplementary 3). Consistent data extraction was ensured by independently piloting 10 articles. We extracted data on publication year, country (data), sample definition, sample restricted to diagnosed population, academic discipline (first author's affiliation), clinical discipline (ICD-10 chapter), PP IV definition, PP IV category (physician, facility, or regional), treatment, outcome, F-statistics from first stage regressions, p-values for treatment effects, and authors specification of research question/objective and results (both in abstract). We also extracted data on whether studies used treatment as a sample-selection criteria, the application was part of a methodology paper, and whether multiple methods were used (triangulation). To ensure consistent data extraction, 20 percent of data was extracted in duplicate by two independent reviewers (TW and either AC or IL), with the remainder extracted by TW and cross-checked by another reviewer. We e-mailed authors when unable to find relevant information. Data available in Supplementary 4.

2.4. Quality assessment

We use reported assumptions necessary for valid IV designs as quality assessment and appraisal in line with existing reviews [39]. We extracted detailed data on A1-A3 and A4h/A4m [16] (Supplementary 3). Each condition was given a score of 1 if reported and 0 otherwise, so the maximum total score a study could obtain was 4. A1-A3 and A4h/A4m were extracted by two independent reviewers for 20 percent of studies (TW and either AC or IL), with the remainder extracted by TW. For A1-A3, a score of 1 was given if these assumptions were acknowledged or discussed, or, for A1 if the association between treatment and IV was reported and for A3 if covariates were included in IV-analyses (A3). For A4h/A4m, studies were coded 1 when reporting monotonicity or homogeneity.

2.5. Synthesis of results

We analyzed trends in PP IV use by publication year/topic, cross-tabulated data, and then used regression

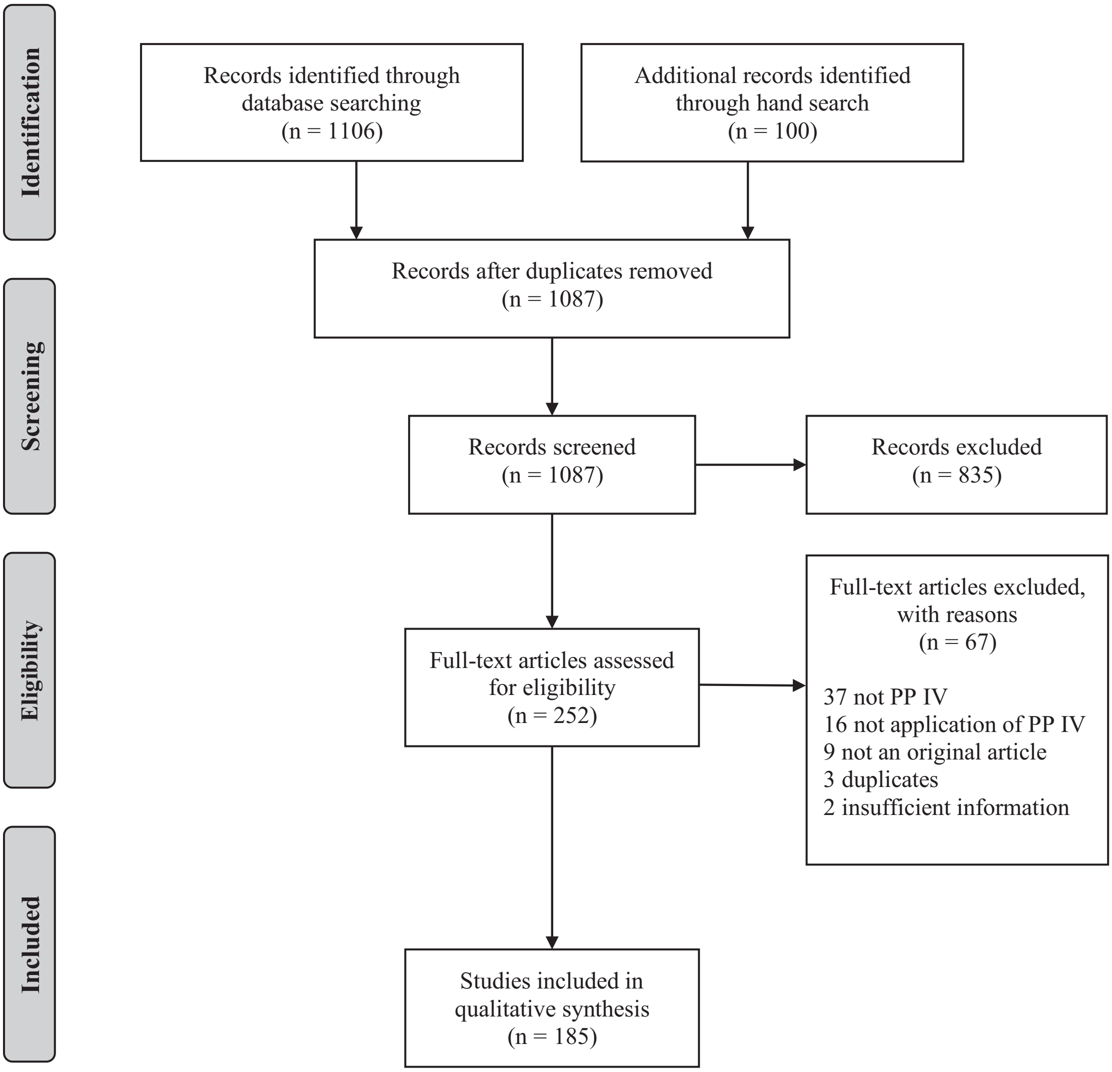
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 1106)
Additional records identified through hand search (n = 100)
Records after duplicates removed (n = 1087)
Records screened (n = 1087)
Records excluded (n = 835)
Full-text articles assessed for eligibility (n = 252)
Full-text articles excluded, with reasons (n = 67)
37 not PP IV
16 not application of PP IV
9 not an original article
3 duplicates
2 insufficient information
Studies included in qualitative synthesis (n = 185)

Figure 2. PRISMA flowchart.

models to test for change in mean reported assumptions score and proportion of significant findings over time. One way ANOVA and Kruskal Wallis tests examined support for differences in reported assumptions scores and *F*-statistics across disciplines, clinical areas, and PP IV categories. Stata SE 16.1. [45] was used for data analysis and visualization.

3. Results

1087 studies were identified and included in initial screening. 252 were assessed in full-text, with 185 meeting inclusion criteria (Figure 2). Figure 3A indicates the yearly number of PP IV studies in health research has increased.

3.1. Areas of application

PP IV methods were most commonly applied in medicine, followed by public health, and economics (Table 3). Most PP IV applications addressed treatment effects for samples with neoplasms, followed by circulatory disease, and mental/behavioral disorders. Together these comprised 55% of applications.

PP IV is increasingly used across clinical areas, with neoplasms consistently at the top (Figure 3B). Most studies use data from the US ($n = 122$, 39.5%), followed by UK ($n = 26$, 8.4%), Canada ($n = 21$, 6.8%), Japan ($n = 20$, 6.5%), and Sweden ($n = 13$, 4.2%) (Figure 3C). The included studies apply data from 33 countries.

3.2. Methodological assessment

The most frequently addressed identifying assumptions were A1 (relevance), A3 (confounding) and A2 (exclusion) (Table 2). Few reported A4h or A4m. Less than half of studies reported *F*-statistics from first stage regressions. We did not find support for a reduction in proportion of reported significant *p*-values over time (Supplementary

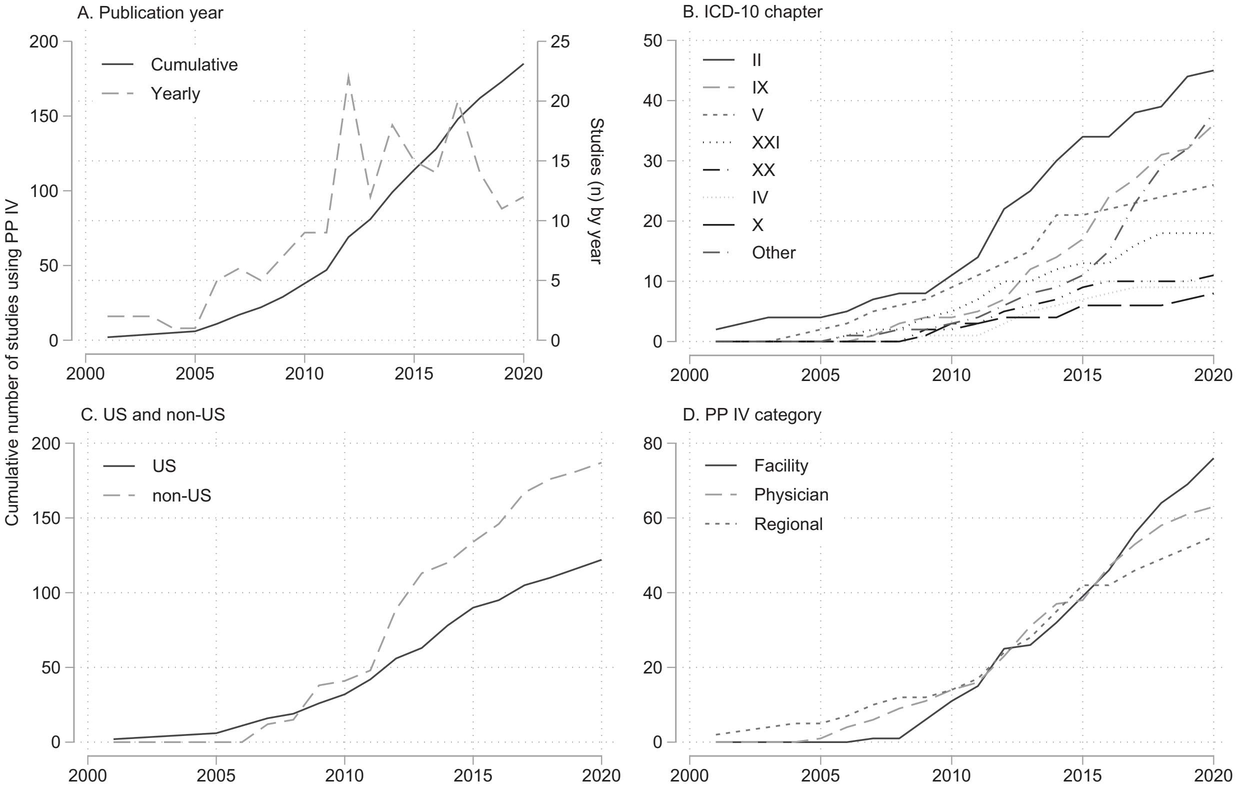

Figure 3. Cumulative trends in applications of preference-based instrumental variables by publication year, clinical area, country and PP IV definition. (A) Publication year with studies by year as right-side *y*-axis. (B) Clinical field by the study population's ICD-10 chapter: (II) Neoplasms, (IX) Diseases of the circulatory system, (V) Mental and behavioral disorders, (XXI) Factors influencing health status and contact with health services, (XX) External causes of morbidity and mortality, (IV) Endocrine, nutritional and metabolic diseases, (X) Diseases of the respiratory system, and remaining chapter codes in Other. (C) US and non-US applications by data used. (D) Provider-preference proxy category. One study may contain multiple PP IV categories, countries, and ICD-10 chapters and can be counted several times.

S2, Table S3). Moreover, bias from selecting on treatment may be an issue in several applications. Many studies use multiple methods to address research questions (triangulation), most commonly IV with multivariable regression or propensity score matching.

Mean quality assessment score (QA) was somewhat higher in public health and pharmacology than in medicine (Table 3). Small differences in QA scores were found between clinical areas and PP IV categories (Table 3). We found no support for a change in assumption reporting over time, and also no change after Swanson and Hernán's reporting guideline [16] was published (Supplementary 2, Table S3).

4. Discussion

4.1. Main findings

Our findings show PP IV is predominantly used to estimate treatment effects for cancer, cardiovascular diseases, and mental health, where valid causal inferences are crucial in planning treatments. Nearly half of the studies provided justification for using PP IV relative to RCT, underlining ways in which PP IV can contribute to causal evidence [11]. We identified more PP IV applications than existing reviews, perhaps due to mining full-texts in our initial search; however, several years have passed since publication of relevant systematic reviews [8],[9],[39].

The most common proxy definition of PP is variation in practice-patterns at facility-level, followed by physician- and regional-level. A mere 12 percent of studies report the four main assumptions (A1-A3 and A4h or A4m) for point identification of treatment effects with IV, while 73 percent report the three main assumptions (A1-A3) necessary for bounds on treatment effects. This is consistent with Swanson and Hernán's review [16] where relatively few reported the fourth assumption. We found considerable variation in how assumptions were reported, ranging from stating assumptions without further justification to careful delineations on validity concerns under given circumstances. The latter approach is encouraged as the validity of PP IV can vary considerably by context. For example, the validity of PP IV have been found to vary by database and definition of PP [46],[47], where the latter may also result in varying effect estimates [48]. Homogeneity (A4h) or monotonicity (A4m) should receive more attention in

Table 2. Methodological assessment of preference-based instrumental variables. Reported values in 185 studies. (1) Studies by total reported assumptions (range 0-4). (2) Recurring themes: RCTs not ethically feasible, impractical due to rare outcomes, generalizability, time-consuming data collection. (3) Multiple methods could be used in 173 studies, 12 studies focused on methodology. (4) Sample selection fully or partially a function of treatment. (5) 173 studies transparently reported 1524 p-values with median 3 per study. Percent weighted by reported p-values. (6) Median and interquartile range (IQR) weighted by F-statistics. 279 F-statistics reported by 86/194 applications. Median 3 per study. Range 4.2-109825. F-statistics above 104.7 (n/N, %): Overall 51/194 (26). By IV category: Physician 21/63 (33), Facility 12/76 (16), Regional 18/55 (33). Kruskal Wallis H test for differences in F-statistics across PP IV categories ($P = .07$).

Identifying assumptions (n, %)	
Stated or empirically verified relevance (A1)	180 (98)
Stated or discussed exclusion (A2)	157 (86)
Stated, discussed, or adjusted for covariates for unconfoundedness (A3)	178 (97)
Stated homogeneity (A4h)	1 (.5)
Stated monotonicity (A4m)	21 (11)
Quality assessment score (n, %)[1]	
1	11 (6)
2	18 (10)
3	134 (72)
4	22 (12)
Justification for using PP IV over RCT (n, %)[2]	86 (46)
Triangulation (n, %)[3]	133 (72)
Selection on treatment (n, %)[4]	85 (46)
p-value for treatment effect significant at 5% level (n/N, %)[5]	642/1524 (42)
Sample size (median, IQR)	31451 (6185-78531)
First stage F-statistic (median, IQR)[6]	270 (69-399)
F-statistic for physician PP IVs	399 (342-1871)
F-statistic for facility PP IVs	190 (29-949)
F-statistic for regional PP IVs	69 (26-135)

future studies given recent studies on how monotonicity in PP IVs can easily be violated [23],[36],[49].

PP IVs are most valuable in studies with considerable unobserved confounding, large sample size and a strong IV [50]. Most studies had relatively large sample sizes , generally much larger than applied in RCTs. Reported *F*-statistics suggest that PP IVs are relatively strong, albeit the wide range is similar to IVs in epidemiology more generally [8]. Potential reporting bias implies a cautious interpretation of our results. Around 40 percent of reported treatment effects were statistically significant at the 5% level.

Publication bias may be an issue for IV applications, and we believe a relevant next step could be to examine *z*-statistics like Brodeur et al. [51]. Our findings show that there is potential selection on treatment bias in 46 percent of PP IV applications, lending support to concerns raised in literature [21],[30]. Finally, many studies combine IV with other study designs which is particularly useful as the combination of multiple designs with various underlying assumptions can create a more comprehensive understanding of treatment effects [2],[52].

4.2. Strengths and limitations

This review was pre-registered in PROSPERO and was conducted accordingly, as described in the methods section. There are some limitations of this study. There is no keyword or search term that identifies with high specificity and sensitivity empirical studies applying PP IV approaches. We aimed to mitigate this limitation by full-text searches. Additional studies could probably have been identified by full-text search in additional journals. We applied only English-language search terms, which may have caused selection bias. Publication bias is difficult to assess when including studies across aims, subjects and disciplines, and no funnel-plot was attempted. Search for *p*-hacking strategies could identify publication bias [51], but we did not attempt this. As there is no developed methodological evaluation tool for critical appraisal of PP IV studies, unlike risk of bias appraisal in RCTs, our quality assessment and appraisal relies on the reporting of IV assumptions in line with existing IV-reviews [39].

4.3. Contribution

This review contributes to existing knowledge on PP IV in three ways. First, to our knowledge this is the first systematic review with an explicit focus on PP as an IV, which is warranted as this is among the most applied IVs in health research [8],[9],[39]. Second, we present novel data on PP IV applications on academic and clinical topics, reporting of IV assumptions, and potential selection

Table 3. Areas of applications and quality assessment score by academic discipline, clinical area, and PP IV category. Quality assessment (QA) score range from 0-4 reported assumptions. Discipline defined by first author's affiliation. Mean QA score differed by discipline (one-way ANOVA: P =.009). Pairwise comparisons show that public health differ from medicine ($P <$ .001), as do pharmacology ($P =$ .04). (2) Consists of general and other subdisciplines not specified. (3) Includes epidemiology and biostatistics. (4) Due to some studies including several ICD-chapter codes, the total is 191. Mean QA score differed by clinical areas (Kruskal Wallis H test: P =.029). Dunn's pairwise comparisons test supported following differences: IV, V, IX relative to II; X, XX relative to IV; X, XX, Other relative to V; X, XX, Other relative to IX. (5) Due to some studies using multiple PP IV categories, the total is 194. Mean QA score varied by PP IV categories (one-way ANOVA: P =.013). Pairwise comparisons showed that physician IVs differed from facility IVs ($P =$ .003).

	Areas of application	*Quality assessment score (n, %)*				
	n (%)	1	2	3	4	Mean
Discipline[1]						
Medicine[2]	60 (32)	7 (11.5)	8 (13.1)	44 (72.1)	2 (3.3)	2.7
Surgery	13 (7)	1 (7.7)	2 (15.4)	10 (76.9)	0 (0)	2.7
Pharmacology	13 (7)	0 (0)	1 (7.7)	10 (77.9)	2 (15.4)	3.1
Psychiatry	2 (1)	0 (0)	0 (0)	2 (100)	0 (0)	3
Public health[3]	80 (43)	3 (3.8)	4 (5)	57 (71)	16 (20)	3.1
Economics	17 (9)	0 (0)	3 (17.7)	12 (70.6)	2 (11.8)	2.9
Total	185 (100)	11 (5.8)	18 (9.5)	138 (73)	22 (11.6)	2.9
ICD-10 Chapter[4]						
Neoplasms (II)	45 (23)	1 (2.2)	6 (13.3)	37 (82.2)	1 (2.2)	2.8
Diseases of the circulatory system (IX)	36 (19)	2 (5.6)	0 (0)	27 (75)	7 (19.4)	3.1
Mental and behavioral disorders (V)	26 (13)	0 (0)	1 (3.9)	20 (77)	5 (19.2)	3.1
Factors influencing health status and contact with health services (XXI)	18 (9)	0 (0)	4 (22)	12 (66.7)	2 (11)	2.9
External causes of morbidity and mortality (XX)	11 (6)	2 (18.2)	1 (9.1)	8 (72.7)	0 (0)	2.5
Endocrine, nutritional, and metabolic diseases (IV)	9 (5)	1 (11.1)	0 (0)	5 (55.6)	3 (33.3)	3.1
Diseases of the respiratory system (X)	8 (4)	2 (25)	1 (12.5)	4 (50)	1 (12.5)	2.5
Other	38 (20)	3 (7.9)	5 (13.2)	26 (68.4)	4 (10.5)	2.8
Total	191 (100)	11 (5.8)	18 (9.4)	139 (72.8)	23 (12)	2.9
PP IV category[5]						
Facility	76 (39.2)	7 (9.2)	14 (18.4)	45 (59.2)	10 (13.2)	2.8
Physician	63 (32.4)	3 (4.8)	1 (1.6)	46 (73)	13 (20.6)	3.1
Regional	55 (28.4)	1 (1.8)	3 (5.5)	49 (89.1)	2 (3.6)	2.9
Total	194 (100)	11 (5.7)	18 (9.3)	140 (72.2)	25 (12.9)	2.9

on treatment bias, and discuss current practice. Third, we highlight specific design considerations raised in the PP IV methodological literature.

4.4. Implications

The credibility of the design requires transparent reporting. In line with former reviews on the use of IVs in health research [8], we find more explicit reporting of assumptions can make it easier to examine support for causal inference [see also, 16, 39, 54]. Future PP IV studies should draw on reporting guidelines [8,16], triangulation [2], DAGs [49], and falsification tests [54].

The PP variable must meet strong assumptions to be considered a plausible IV. When assumptions are violated, estimates can be biased in counterintuitive ways [2],[53]. Hence, there is a trade-off between accounting for unobserved confounding and introducing bias where the decision to apply PP IV should factor in strength of confounding and credibility of the PP IV [28],[53]. Moreover, IVs often have wide confidence intervals and may be prone to *p*-hacking compared to other common quasi-experimental designs [51]. While application of any method requires care, the combination of all aspects that go into valid causal inference from PP IVs warrant extra attention.

5. Conclusion

This systematic review provides evidence that PP is commonly used as an IV in health research, particularly for cancer, cardiovascular diseases and mental health, and presents novel data on methodological considerations. The review identified more applications of PP IV than existing reviews and expanded on reporting assumptions. We encourage authors and journals to emphasize reporting guidelines [8,16] in studies using PP IV. Empirical studies applying PP IV methods have merit to inform clinical and policy decisions on questions challenging or unfeasible to address with RCTs, but impact rests on the credibility of the study design.

Please cite this article as: T. Widding-Havneraas et al., Preference-based instrumental variables in health research rely on important and underreported assumptions: A systematic review, Journal of Clinical Epidemiology, https://doi.org/10.1016/j.jclinepi.2021.06.006

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.06.006.

CRediT authorship contribution statement

Tarjei Widding-Havneraas: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing, Funding acquisition. **Ashmita Chaulagain:** Conceptualization, Methodology, Investigation, Writing – original draft. **Ingvild Lyhmann:** Conceptualization, Methodology, Investigation, Writing – original draft. **Henrik Daae Zachrisson:** Conceptualization, Methodology, Supervision, Visualization, Writing – original draft. **Felix Elwert:** Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Simen Markussen:** Writing – original draft. **David McDaid:** Methodology, Writing – review & editing. **Arnstein Mykletun:** Conceptualization, Methodology, Supervision, Funding acquisition, Writing – review & editing.

References

[1] Morgan SL, Winship C. Counterfactuals and causal inference: Methods and principles for social research. Analytical methods for social research. 2 ed. New York: Cambridge University Press; 2015.
[2] Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC 2020.
[3] Wright PG. Tariff on animal and vegetable oils. New York: Macmillan Company; 1928.
[4] Haavelmo T. The statistical implications of a system of simultaneous equations. Journal of the Econometric Society 1943:1–12. Econometrica.
[5] Reiersøl O. Confluence analysis by means of instrumental sets of variables. Almqvist & Wiksell 1945.
[6] Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. J Am Stat Assoc 1996;91(434):444–55.
[7] Stock J, Trebbi F. Retrospectives: Who invented instrumental variable regression? J. Econ. Perspect. 2003;17(3):177–94.
[8] Davies NM, Smith GD, Windmeijer F, Martin RM. Issues in the Reporting and Conduct of Instrumental Variable Studies: A Systematic Review. Epidemiology 2013;24(3):363–9.
[9] Garabedian LF, Chu P, Toh S, Zaslavsky AM, Soumerai SB. Potential Bias of Instrumental Variable Analyses for Observational Comparative Effectiveness Research. Ann. Intern. Med. 2014;161(2):131–+.
[10] Cawley J. A selective review of the first 20 years of instrumental variables models in health-services research and medicine. J. Med. Econ. 2015;18(9):721–34.
[11] Bärnighausen T, Tugwell P, Røttingen J-A, Shemilt I, Rockers P, Geldsetzer P, et al. Quasi-experimental study designs series—paper 4: uses and value. J. Clin. Epidemiol. 2017;89:21–9.
[12] Glymour MM, Swanson AS, et al. Instrumental Variables and Quasi-Experimental Approaches. In: Lash TL, et al., editors. Modern Epidemiology. Wolters Kluwer; 2021. p. 677–709.
[13] Angrist JD, Krueger AB. Instrumental variables and the search for identification: From supply and demand to natural experiments. J. Econ. Perspect. 2001;15(4):69–85.
[14] Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. Int J Biostat 2007;3(1):14.
[15] Korn EL, Baumrind S. Clinician Preferences and the Estimation of Causal Treatment Differences. Statistical Science 1998;13(3):209–27.
[16] Swanson SA, Hernán M. Commentary: How to Report Instrumental Variable Analyses (Suggestions Welcome). Epidemiology 2013;24(3):370–4.
[17] Wang PS, Schneeweiss S, Avorn J, Fischer MA, Mogun H, Solomon DH, et al. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. N Engl J Med 2005;353(22):2335–41.
[18] Dalsgaard S, Nielsen HS, Simonsen M. Consequences of ADHD medication use for children's outcomes. Journal of Health Economics 2014;37:137–51.
[19] Emdin CA, Hsiao AJ, Kiran A, Conrad N, Salimi-Khorshidi G, Woodward M, et al. Referral for specialist follow-up and its association with post-discharge mortality among patients with systolic heart failure (from the National Heart Failure Audit for England and Wales). Am. J. Cardiol. 2017;119(3):440–4.
[20] Uddin MJ, Groenwold RH, Boer T, Belitser SV, Roes KC, Klungel OH. Instrumental Variable Analysis in Epidemiologic Studies: An Overview of the Estimation Methods. Pharmaceutica Analytica Acta 2015;6(4).
[21] Swanson SA, Robins JM, Miller M, Hernan MA. Selecting on Treatment: A Pervasive Form of Bias in Instrumental Variable Analyses. Am. J. Epidemiol. 2015;181(3):191–7.
[22] Imbens GW. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). J. Econ. Lit. 2010;48(2):399–423.
[23] Swanson SA, Hernán MA. Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation. Statistical Science 2014;29(3):371–4.
[24] Elwert F, *Graphical Causal Models*, in *Handbook of Causal Analysis for Social Research*, S.L. Morgan, Editor. 2013, Springer Netherlands: Dordrecht. 245-273.
[25] Hernán M, Robins J. Instruments for Causal Inference: An Epidemiologist's Dream? Epidemiology 2006;17(4):360–72.
[26] Angrist JD, Pischke J-S. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton: Princeton University Press; 2009.
[27] Frölich M. Nonparametric IV estimation of local average treatment effects with covariates. Journal of Econometrics 2007;139(1):35–75.
[28] Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Statistics in Medicine 2014;33(13):2297–340.
[29] Davies NM, Thomas KH, Taylor AE, Taylor GM, Martin RM, Munafò MR, et al. How to compare instrumental variable and conventional regression analyses using negative controls and bias plots. Int. J. Epidemiol. 2017;46(6):2067–77.
[30] Elwert F, Segarra E, *Instrumental Variables with Treatment-Induced Selection: Exact Bias Results.* arXiv preprint arXiv:2005.09583, 2020.
[31] Pearl J. Causality: Models, Reasoning and Inference. 2nd ed. Cambridge University Press; 2009.
[32] Hughes RA, Davies NM, Smith GD, Tilling K. Selection Bias When Estimating Average Treatment Effects Using One-sample Instrumental Variable Analysis. Epidemiology 2019;30(3).
[33] Elwert F, Winship C. Endogenous selection bias: the problem of conditioning on a collider variable. Annu. Rev. Sociol. 2014;40:31–53.
[34] Swanson SA. A Practical Guide to Selection Bias in Instrumental Variable Analyses. Epidemiology 2019;30(3):345–9.
[35] Ertefaie A, Small D, Flory J, Hennessy S. Selection Bias When Using Instrumental Variable Methods to Compare Two Treatments But More Than Two Treatments Are Available. Int J Biostat 2016;12(1):219–32.
[36] Swanson SA, Miller M, Robins JM, Hernan MA. Definition and

evaluation of the monotonicity condition for preference-based instruments. Epidemiology 2015;26(3):414–20.

[37] Staiger D, Stock JH. Instrumental Variables Regression with Weak Instruments. Econometrica 1997;65(3):557–86.

[38] Lee, D. S., McCrary, J., Moreira, M. J., Porter, J. *Valid t-ratio Inference for IV.* arXiv:2010.05058, 2020.

[39] Chen Y, Briesacher BA. Use of instrumental variable in prescription drug research with observational data: a systematic review. J. Clin. Epidemiol. 2011;64(6):687–700.

[40] Potter BJ, Dormuth C, Lorier JLe. A theoretical exploration of therapeutic monomania as a physician-based instrumental variable. Pharmacoepidemiology and Drug Safety 2020;29:45–52.

[41] Penning de Vries BBL, van Smeden M, Rosendaal FR, Groenwold RHH. Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice. J. Clin. Epidemiol. 2020;121:55–61.

[42] Moher D, Liberati D, Tetzlaff J, Altman DGfor the PRISMA group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine 2009;6(7):e1000097.

[43] EndNote X9, *Clarivate Analytics,* 2019.

[44] Covidence systematic review software, *Veritas Health Innovation,* 2020: Melbourne, Australia.

[45] StataCorp, *Stata Statistical Software: Release 16.1*. 2020, StataCorp LLC: College Station, TX.

[46] Uddin MJ, Groenwold RH, de Boer A, Afonso AS, Primatesta P, Becker C, et al. Evaluating different physician's prescribing preference based instrumental variables in two primary care databases: a study of inhaled long-acting beta2-agonist use and the risk of myocardial infarction. Pharmacoepidemiology & Drug Safety 2016;25(1):132–41 Suppl.

[47] Uddin MJ, Groenwold RH, de Boer A, Gardardsdottir H, Martin E, Candore G, et al. Instrumental variables analysis using multiple databases: an example of antidepressant use and risk of hip fracture. Pharmacoepidemiology & Drug Safety 2016;25(1):122–31 Suppl.

[48] Ionescu-Ittu R, Abrahamowicz M, Pilote L. Treatment effect estimates varied depending on the definition of the provider prescribing preference-based instrumental variables. J. Clin. Epidemiol. 2012;65(2):155–62.

[49] Swanson SA, Hernan MA. The challenging interpretation of instrumental variable estimates under monotonicity. Int J Epidemiol 2018;47(4):1289–97.

[50] Boef AGC, Dekkers OM, Vandenbroucke JP, le Cessie S. Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and level of confounding. J. Clin. Epidemiol. 2014;67(11):1258–64.

[51] Brodeur A, Cook N, Heyes A. Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. American Economic Review 2020;110(11):3634–60.

[52] Matthay EC, Hagan E, Gottlieb LM, Tan ML, Vlashov D, Adler NE, et al. Alternative causal inference methods in population health research: Evaluating tradeoffs and triangulating evidence. SSM - Population Health 2020;10:100526.

[53] Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiology and Drug Safety 2010;19(6):537–54.

[54] Labrecque J, Swanson SA. Understanding the Assumptions Underlying Instrumental Variable Analyses: a Brief Review of Falsification Strategies and Related Tools. Current Epidemiology Reports 2018;5(3):214–20.

Please cite this article as: T. Widding-Havneraas et al., Preference-based instrumental variables in health research rely on important and underreported assumptions: A systematic review, Journal of Clinical Epidemiology, https://doi.org/10.1016/j.jclinepi.2021.06.006