PNAS INAUGURAL ARTICLE | ECONOMIC SCIENCES MEDICAL SCIENCES

Instrumental variables methods reconcile intention-to-screen effects across pragmatic cancer screening trials

Joshua D. Angrist[a,1] and Peter Hull[b,1]

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2023.
Contributed by Joshua D. Angrist, received July 7, 2023; accepted November 3, 2023; reviewed by Amanda Kowalski and Robert W. Yeh

Pragmatic cancer screening trials mimic real-world scenarios in which patients and doctors are the ultimate arbiters of treatment. Intention-to-screen (ITS) analyses of such trials maintain randomization-based apples-to-apples comparisons, but differential adherence (the failure of subjects assigned to screening to get screened) makes ITS effects hard to compare across trials and sites. We show how instrumental variables (IV) methods address the nonadherence challenge in a comparison of estimates from 17 sites in five randomized trials measuring screening effects on colorectal cancer incidence. While adherence rates and ITS estimates vary widely across and within trials, IV estimates of per-protocol screening effects are remarkably consistent. An application of simple IV tools, including graphical analysis and formal statistical tests, shows how differential adherence explains variation in ITS impact. Screening compliers are also shown to have demographic characteristics similar to those of the full trial study sample. These findings argue for the clinical relevance of IV estimates of cancer screening effects.

clinical trials | causal effects | adherence | econometrics | treatment effects

Significance

As in real-world clinical settings, patients offered colonoscopy or sigmoidoscopy screening in pragmatic cancer screening trials may decide to skip it. The fact that many patients randomly assigned to screen remain unscreened in pragmatic trials (a problem called "nonadherence") complicates the interpretation of pragmatic trial results. We argue that econometric instrumental variables (IV) methods resolve the complications engendered by nonadherence. IV estimates from five trials show a stable reduction in colorectal cancer incidence from screening of almost 0.5 percentage points. A suite of graphical and statistical tools bolsters the clinical relevance of IV estimates of screening effects. Regrettably, many trials to date have failed to collect the data needed for IV analysis. But this is easily remedied in future trials.

The question of whether cancer screening improves health remains contentious—a fact highlighted by recent debates over mammography, prostate-specific antigen (PSA) screening, and colorectal cancer (CRC) screening.* Regularly screened patients tend to be healthier than those who opt out. But this observational comparison may be misleading: Patients and doctors who do and do not screen are likely to differ in many ways besides screening itself. When screening is randomly assigned, and those assigned to screening are indeed screened, any later difference in the health of screened and unscreened participants is almost certainly caused by screening. This fact motivates randomized screening trials, which offer screening to participants by lottery.†

Pragmatic randomized trials, meant to "measure effectiveness in routine clinical practice" (5), are particularly well suited to estimate the real-world impact of cancer screening on health. As in clinical practice, pragmatic trials allow patients and their doctors to be the ultimate arbiters of screening and other treatments. At the same time, evaluation of unpleasant and time-consuming medical interventions under real-world conditions is often complicated by the fact that many patients fail to take their doctor's advice. It is one thing to randomize the opportunity to screen, quite another to randomize screening itself. Consequently, pragmatic cancer screening trials typically report intention-to-screen (ITS) effects that compare those randomized to receive an offer of screening with a control group that receives no such offer.

Free colonoscopies—now there's an offer! Indeed, when it comes to pragmatic trials for colonoscopy and sigmoidoscopy, the share of participants randomized to a screening invitation who are actually screened can be worryingly low. Data from five screening trials, summarized in Table 1, bear this out. In four sigmoidoscopy trials, adherence ranges from a low of 58% in the Italian SCORE study to a high of 87% in the American PLCO study. The more invasive colonoscopy screenings offered to patients in Poland, Norway, and Sweden in the NordICC pragmatic trial resulted in even lower adherence, with only 42% of those randomly offered a colonoscopy completing one.

Nonadherence in NordICC, which showed little mortality benefit alongside reduced CRC incidence, recently sparked a debate over the clinical relevance of screening trial findings (6). In one of many letters responding to the Bretthauer et al. (7) report on

*See, e.g., Kowalski (1, 2) for discussions of mammography, Hayes and Barry (3) on PSA, and references in the CRC studies cited below.

†A recent NEJM editorial on CRC screening (4) notes that "[non-randomized] studies probably overestimate the real-world effectiveness of colonoscopy because of the inability to adjust for important factors such as incomplete adherence to testing and the tendency of healthier persons to seek preventive care."

Author contributions: J.D.A. and P.H. designed research; performed research; analyzed data; and wrote the paper.

Reviewers: A.K., University of Michigan; and R.W.Y., Harvard University.

Competing interest statement: J.D.A. serves on the board of Avela, an ed-tech startup; holds equity in Avela; and has authored textbooks describing the methods used in the manuscript.

[1]To whom correspondence may be addressed. Email: angrist@mit.edu or peter_hull@brown.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2311556120/-/DCSupplemental.

Published December 15, 2023.

Table 1. Sigmoidoscopy and colonoscopy screening trials

Trial	NordICC (1)	NORCCAP (2)	SCORE (3)	UKFSST (4)	PLCO (5)
Countries	Poland, Norway, Sweden	Norway	Italy	UK	U.S.
Screening period	2009 to 2014	1999 to 2001	1995 to 1999	1994 to 1999	1993 to 2001
Initial screening type	Colonoscopy	Sigmoidoscopy + FOBT	Sigmoidoscopy	Sigmoidoscopy	Sigmoidoscopy
Follow-up screening type	Biopsy	Colonoscopy, Biopsy	Colonoscopy	Colonoscopy	Second sigmoidoscopy (after 3/5 y)
Participants identification	Population registry	Population registry	Survey	Survey	Survey
Median Follow-up Years	10.0	10.9	10.5	11.2	11.9
Participants (N)	84,585	98,792	34,272	170,038	154,900
Range	55 to 64	50 to 64	55 to 64	55 to 64	55 to 74
Invitation ratio	0.33	0.21	0.50	0.34	0.50
Adherence rate	0.42	0.63	0.58	0.71	0.87

Notes: This table summarizes key features of the trials analyzed here. These trials randomly assigned an invitation to screening in the form of flexible sigmoidoscopy or colonoscopy. Half of the NORCCAP-treated group was invited for sigmoidoscopy, with the rest invited for both sigmoidoscopy and a fecal occult blood test (FOBT). The PLCO trial offered a second screening 3 or 5 y after the initial screening. This second screening had an adherence rate of 0.51. The adherence rate is the proportion screened in the group invited for screening. The number of participants counts only subjects with follow-up data. The PLCO protocol includes two invitations to screen, an initial sigmoidoscopy and a second sigmoidoscopy 3 to 5 y later for all subjects initially invited to screen. Follow-up screening invitations in NORCCAP, SCORE, and UKFSST are based on polyp detection in the initial screening. NordICC does not provide follow-up screenings.

NordICC, Winawer (8) asks, "are these intention-to-treat observations applicable to other clinical environments?" In cancer screening trials, randomization of invitations to screen ensures ITS effects are free of selection bias—meaning they are unconfounded by pre-treatment differences between those assigned to screening and control groups. Yet, as ref. 8 suggests, nonadherence can make ITS effects hard to compare across studies and even harder to apply to public health policy. Intuitively, low adherence dilutes ITS effects by including in the treatment group people whose screening behavior is unaffected by a randomized opportunity to screen. The number needed to screen in order to prevent cancer may therefore be well below the number that must be offered screening in a trial.

A role for offer adherence in mediating ITS effects is suggested by the first column of Table 2, which reports estimated ITS effects on CRC incidence 10 to 12 y after random assignment for the five trials summarized in Table 1. In principle, screening reduces CRC incidence by revealing precancerous abnormalities in the colon, which can then be removed.[‡] In practice, ITS effects on CRC incidence vary widely. The estimated ITS effect in the NordICC trial is 0.19 percentage points (reported as −0.0019 in the table), while data from the UKFSST trial yield an ITS estimate that's nearly twice as big at 0.37 points. These estimates are precise enough that the difference between them is unlikely to be due to chance (the P value for the null of equality is 0.02).

At first blush, systematic differences in ITS effects for a common or similar intervention would seem to threaten the external validity—and therefore the clinical relevance—of individual studies. Of course, medical interventions may affect different populations differently. The NordICC and UKFSST study populations are broadly similar, however, both involving men and women aged 55 to 64 in European countries that offer low-cost access to modern medical services. On the other hand, the impact of colonoscopy screening examined in NordICC might exceed that of less invasive and less sensitive sigmoidoscopy screening examined in UKFSST. But the ITS results in this regard present a puzzle, since the estimated CRC incidence reduction due to NordICC colonoscopy offers is far below the the estimated CRC incidence reduction yielded by UKFSST sigmoidoscopy offers.

This article shows that divergent ITS estimates—across trials, across sites within trials, and even across variations on a similar treatment—can be reconciled by instrumental variables (IV) methods that make adherence the mediator of trial effects. The next section sketches the IV approach to causal inference. The IV estimand, known to econometricians as a local average treatment effect (LATE), is shown to be a type of per-protocol effect that captures the average screening effect for subjects induced to screen by virtue of their trial participation. In some cases, this can be interpreted as the number needed to screen to prevent 1 CRC case. Section 2 uses IV to estimate screening effects on CRC incidence. Substantial variability in ITS estimates notwithstanding, LATE estimates are remarkably consistent across and within the five studies in Table 1. The fact that adherence explains variation in ITS impact, while LATEs are reasonably stable, bolsters the case for seeing IV estimates as clinically relevant.[§] In support of this claim, Section 3 deploys three IV tools not previously applied in this context: visual instrumental variables, overidentification testing, and complier characteristics. Section 4 summarizes our argument and draws some conclusions.

1. The IV Advantage

A. Casting Causal Effects. Consider a pragmatic trial offering CRC screening by lottery to a population of experimental subjects indexed by i. Let $Z_i \in \{0, 1\}$ be a dummy variable indicating experimental screening offers (also called invitations) and let $S_i \in \{0, 1\}$ be a dummy variable indicating post-randomization screening completion. Subjects are free to decline or ignore screening offers, while some not invited for screening through the trial may be screened elsewhere. The possibility of nonadherence is reflected in the fact that $S_i \neq Z_i$ for some (and perhaps many) subjects. CRC incidence, denoted by dummy variable $Y_i \in \{0, 1\}$, is measured for all subjects after offers are made in the trial.

A potential outcomes model is used to define the causal effects of interest in our setting (and many others; see, e.g., ref. 11). Let dummy variable Y_{0i} indicate the CRC status of subject i when she is unscreened, while Y_{1i} indicates CRC incidence when i is

[‡]Our follow-up horizon matches that in ref. 7. We focus on CRC incidence over mortality because estimates for the former are more precise, a point noted by Bretthauer et al. (9).

[§]Angrist and Meager (10) makes an analogous point in the context of schooling-related interventions in developing countries, where mediating instrumental variables gauge program implementation.

Table 2. IV estimates of screening effects on CRC incidence

				Per-protocol		
	Control mean (1)	ITS (2)	First stage (Adherence) (3)	IV/LATE (4)	As-treated (5)	PP omitting Never-takers (6)
NordICC	0.0110	−0.0019	0.4197	−0.0044	−0.0021	−0.0024
		(0.0006)	(0.0020)	(0.0017)	(0.0009)	(0.0010)
NORCCAP	0.0114	−0.0022	0.6297	−0.0035	−0.0025	−0.0024
		(0.0007)	(0.0027)	(0.0013)	(0.0009)	(0.0009)
SCORE	0.0179	−0.0032	0.5784	−0.0055	−0.0050	−0.0051
		(0.0009)	(0.0019)	(0.0024)	(0.0014)	(0.0015)
UKFSST	0.0161	−0.0037	0.7114	−0.0052	−0.0051	−0.0051
		(0.0005)	(0.0013)	(0.0008)	(0.0006)	(0.0006)
PLCO	0.0171	−0.0037	0.8660	−0.0042	−0.0038	−0.0040
		(0.0006)	(0.0012)	(0.0007)	(0.0006)	(0.0006)

Notes: This table reports the estimated effect of colonoscopy screening on 10-y colorectal cancer (CRC) incidence. The instrument is a randomly assigned invitation to undergo colonoscopy/sigmoidoscopy screening; treatment is defined as receiving colonoscopy/sigmoidoscopy screening. The outcome variable indicates CRC diagnosis 10 to 12 y after random assignment. Column 1 reports mean CRC incidence in the group not offered screening. Column 2 reports the reduced-form (intention-to-screen; ITS) effects of screening invitation on CRC incidence; column 3 reports the first-stage effect of screening invitation on screening. The IV estimate reported in column 4 is the ratio of ITS to first stage. Column 5 reports the as-treated effects of undergoing screening on CRC incidence; column 6 reports effects of invitation to screening on CRC incidence, omitting those that were invited but did not undergo screening. Robust SEs appear in parentheses.

screened. Only one of these potential outcomes is ever observed for a given subject, depending on the value of S_i. In particular, observed CRC incidence can be written:

$$Y_i = Y_{0i} + S_i(Y_{1i} - Y_{0i}). \quad [1]$$

The difference in potential outcomes by screening status, $Y_{1i} - Y_{0i}$, is the causal effect of screening on individual i. This is never seen for any one person, since we only see one of Y_{0i} or Y_{1i} for each i. Randomization of Z_i makes Z_i independent of both Y_{0i} and Y_{1i}.

Although individual causal effects are unknowable, randomized trials with full adherence reveal average effects. Specifically, when $S_i = Z_i$ for all i, a comparison of the average Y_i in the samples of screened ($S_i = 1$) and unscreened ($S_i = 0$) groups give the average screening effect, $E[Y_{1i} - Y_{0i}]$:

$$E[Y_i|S_i = 1] - E[Y_i|S_i = 0] = E[Y_{1i}|Z_i = 1] - E[Y_{0i}|Z_i = 0] = E[Y_{1i}] - E[Y_{0i}] = E[Y_{1i} - Y_{0i}].$$

The first equality follows from the potential outcomes model and the assumption that $S_i = Z_i$; the second follows from the random assignment of Z_i, which makes this independent of potential outcomes; the third follows from the fact that the expectation of a difference is the corresponding difference in expectations.

When screening itself is effectively randomized (because of full adherence), the unconditional average screening effect $E[Y_{1i} - Y_{0i}]$ also equals the average effect of screening on the screened:

$$E[Y_{1i} - Y_{0i}] = E[Y_{1i} - Y_{0i}|S_i = 1].$$

This quantity answers the question of whether those who are screened have lower average CRC incidence than they would have suffered in a counterfactual scenario in which they are unscreened. Clinicians and public health officials often prioritize this measure of impact, which reveals the extent to which people screened in a trial can expect to have fewer cancers as a result of screening. Moreover, with a dummy variable outcome like CRC incidence, the reciprocal of $E[Y_{1i} - Y_{0i}|S_i = 1]$ is the epidemiological "number needed to screen": the number of patients that must be screened, on average, to prevent one CRC case (12). To see this, note that a single CRC case is prevented by screening N^* such that $1 = E[Y_{1i} - Y_{0i} \mid S_i = 1] \times N^*$. The number needed to screen is therefore $N^* = E[Y_{1i} - Y_{0i} \mid S_i = 1]^{-1}$.

B. A Little LATE. For many subjects in screening trials, treatment received diverges from treatment assigned. Subjects who are especially healthy, worried, or well informed may be most likely to respond to a randomized invitation to screen. In such scenarios, screening S_i is no longer randomly assigned though it is still correlated with the randomized screening offers, Z_i. We model this correlation using potential adherence. Specifically, let S_{1i} denote a dummy variable indicating i's screening status when offered screening, while S_{0i} denotes a dummy indicating i's screening status when not offered. Potential adherence determines screening status according to:

$$S_i = S_{0i} + Z_i(S_{1i} - S_{0i}). \quad [2]$$

The causal effect of screening offers on an individual's screening behavior is the difference in potential adherence, $S_{1i} - S_{0i}$.

The local average treatment effects model, introduced in Imbens and Angrist (13) and Angrist et al. (14), categorizes trial participants on the basis of potential adherence. In randomized screening trials, screening compliers are subjects for whom $S_{1i} = 1$ and $S_{0i} = 0$. In the vernacular of screening trials, compliers are subjects who adhere to the screening status to which they are randomly assigned. Subjects for whom $S_{1i} = S_{0i} = 0$ or $S_{1i} = S_{0i} = 1$ are either never or always screened, regardless of Z_i. The LATE framework presumes that the trial population includes at least some compliers.

The LATE setup also assumes away the possibility of a perverse response in which trial participants are screened only when not invited for screening but are not screened when invited. In other words, we assume no subject has $S_{1i} = 0$ and $S_{0i} = 1$. Given this monotonicity assumption, $C_i = S_{1i} - S_{0i}$ is a dummy variable that equals one for compliers and is zero otherwise. Monotonicity is surely satisfied when those not offered screening have no other access to it, since $S_{0i} = 0$ then equals zero for all i. More generally, monotonicity is satisfied when invitations to screen necessarily

make screening more attractive and accessible to some subjects, with no effect on screening status for subjects not invited to screen.

A final LATE assumption is called an exclusion restriction. In our context, the exclusion restriction says that randomized invitations to screen have no effect on CRC incidence other than by boosting the likelihood of screening.[¶] Like monotonicity, this assumption is plausible in pragmatic screening trials where screening offers have no intrinsic value beyond possibly encouraging screening. Given exclusion, the random assignment of screening offers makes Z_i independent of the set $(Y_{0i}, Y_{1i}, S_{0i}, S_{1i})$. An ITS analysis leverages this independence to estimate the average effect of screening offers on CRC incidence. More ambitiously, IV takes us from ITS offer effects to the effect of screening itself.

The journey from ITS to screening effects starts by combining Eqs. **1** and **2** to show that randomized offers determine outcomes according to:

$$Y_i = Y_{0i} + S_{ji}(Y_{1i} - Y_{0i}) \text{ when } Z_i = j.$$

Because Z_i is independent of $(Y_{0i}, Y_{1i}, S_{0i}, S_{1i})$, this representation can be used to write ITS as:

$$\begin{aligned} &E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0] \\ &= E[Y_{0i} + S_{1i}(Y_{1i} - Y_{0i})] - E[Y_{0i} + S_{0i}(Y_{1i} - Y_{0i})] \\ &= E[(S_{1i} - S_{0i})(Y_{1i} - Y_{0i})] = E[C_i(Y_{1i} - Y_{0i})] \equiv \rho. \end{aligned} \quad \textbf{[3]}$$

In an IV context, the ITS effect denoted by ρ is called the reduced-form effect of treatment assignment. In a screening trial with less than full adherence, the reduced form averages $Y_{1i} - Y_{0i}$ for compliers (for whom $C_i = 1$) with zeros for subjects whose screening status is unchanged by Z_i (for whom $C_i = 0$). Hence, ITS understates the magnitude of the effect of screening itself.

Along with CRC incidence, screening status becomes an additional outcome in the LATE framework. The causal effect of screening offers on screening status is given by a comparison of conditional average screening rates analogous to the comparison generating the reduced form:

$$E[S_i \mid Z_i = 1] - E[S_i \mid Z_i = 0] = E[S_{1i} - S_{0i}] \equiv \pi. \quad \textbf{[4]}$$

In an IV context, π is called the first-stage effect of intended treatment assignment. Because monotonicity makes $S_{1i} - S_{0i} = C_i$ a dummy variable, π is the probability of compliance:

$$\pi = E[S_{1i} - S_{0i}] = Pr(C_i = 1).$$

The first stage captures the extent to which ρ is diluted by nonadherence. In a trial where few subjects take screening offers, C_i is mostly zero and, consequently, the reduced form is necessarily small. As long as the first stage is nonzero, however, some subjects offered a chance to screen take it. By dividing ρ by π, IV adjusts for dilution due to nonadherence, transforming the reduced form into a screening effect. This is formalized by using Eqs. **3** and **4** and the fact that C_i is a dummy to write:

$$\begin{aligned} \frac{\rho}{\pi} = \frac{E[C_i(Y_{1i} - Y_{0i})]}{E[S_{1i} - S_{0i}]} &= \frac{E[Y_{1i} - Y_{0i} \mid C_i = 1]Pr(C_i = 1)}{Pr(C_i = 1)} \\ &= E[Y_{1i} - Y_{0i} \mid C_i = 1]. \end{aligned} \quad \textbf{[5]}$$

LATE, defined as $E[Y_{1i} - Y_{0i} \mid C_i = 1]$, is the average causal effect of screening on screening compliers. Given monotonicity, random assignment of screening offers, and the exclusion restriction, the ratio of reduced-form offer effects to first-stage offer effects is the average causal effect of screening on experimental subjects screened when randomized to receive screening offers (but not otherwise). From a public-health perspective, the reciprocal of LATE gives the number needed to screen per cancer averted in the population that's responsive to screening opportunities.

LATE can be consistently estimated by replacing conditional expectations with sample averages on the left side of the formulas for π and ρ, above.[#] But the link between LATE and IV is of practical as well as conceptual value. In practice, IV estimates and the associated standard errors are easily computed using two-stage least squares (2SLS), an IV estimator described in Section 3. Powerful and flexible 2SLS estimators accommodate covariates and multiple instruments (both arise, for instance, in stratified trials in which offers are made at different rates in different strata). 2SLS also provides an immediate path to off-the-shelf statistical inference.[‖]

In comments on the relevance of IV adjustments for nonadherence, Hernán and Robins (24) notes that it is usually impossible to name individual LATE compliers in a study population, since only one of S_{1i} and S_{0i} is observed for any one subject. Even in a trial in which no randomized controls cross-over to receive screening (so $S_{0i} = 0$ for all i), the identity of compliers among subjects not offered treatment remains hidden since we don't know S_{1i} when $Z_i = 0$. Yet, just as readers of medical journals must remain ignorant of treated subjects' identities, researchers and other observers need not identify individual compliers. Rather, these observers are likely most interested in the distribution of complier characteristics. Are compliers mostly old or mostly young? Mostly male or mostly female? Do they have pre-existing conditions that predispose them to take advantage of screening? Are complier populations so unusual that the external validity of IV estimates is limited? The IV tools detailed in Section 3 answer these questions.

C. LATE, Effects on the Screened, and Per-Protocol the Old-Fashioned Way. Trial analysts distinguish intent-to-treat effects from per-protocol effects, typically defined as "the effect that would have been observed had all trial participants followed the trial protocol" (25). LATE is also per-protocol effect, but not for everyone: as Eq. **5** shows, ρ/π gives the average causal effect of screening among experimental subjects screened as a result of the trial—that is, for screening compliers. The complier population constitutes the subset of the study population that follows a trial protocol in the field.

Importantly, when all subjects not offered screening remain unscreened, LATE equals the average effect of screening on everyone in the study population who is screened. In other words, with no control-group crossovers into screening (as in most of the CRC screening trials analyzed below), LATE is an average causal effect in the population for which $S_i = 1$. This is a consequence of the fact that, in general, two sorts of subjects are screened:

[¶] Exclusion is formalized with the help of double-indexed potential outcomes. Let $Y_i(d, z)$ denote the outcome realized for subject i when $D_i = d$ and $Z_i = z$. Exclusion asserts that $Y_i(d, 0) = Y_i(d, 1) = Y_{di}$ for each $d \in \{0, 1\}$.

[#] The term "consistent" is used here in the statistician's sense: Sample moments and smooth functions thereof converge in probability to the corresponding population quantities as the sample size grows to infinity.

[‖] IV ideas applied to randomized trials appear in alternate forms in social science and medicine without referencing IV or potential outcomes. Bloom (15) adjusts trial data for treated never-takers. Newcombe (16) derives an adjustment for randomized trials with control-group crossovers. Hearst et al. (17) uses similar reasoning to obtain effects of Vietnam-era military service using the American draft lottery. Baker and Lindeman (18) and Baker et al. (19) use maximum likelihood to derive an IV-type adjustment for nonadherence in a model for Bernoulli outcomes. Some analyses of screening trials, including Atkin et al. (20) and Segnan et al. (21), reference an adherence adjustment due to Cuzick et al. (22). Also focusing on Bernoulli outcomes, the latter derives a maximum likelihood estimator that adjusts risk ratios for nonadherence. The Cuzick et al.'s (22) estimator is an instance of results in Imbens and Rubin (23), which uses IV to compute marginal distributions of potential outcomes for compliers.

- Those with $S_{0i} = 1$, in which case monotonicity implies $S_{1i} = 1$ as well. Ref. 14 calls this group, which is screened regardless of Z_i, always-takers.
- Compliers who are offered screening, a group for which $Z_i = 1$ and $S_{1i} - S_{0i} = S_{1i} = 1$.

In screening trials in which no controls are screened, the fact that $S_{0i} = 0$ for all i means there are no always-takers. Hence, only the second group, compliers with $Z_i = 1$, are screened. Moreover, because Z_i is randomly assigned, effects on compliers offered screening are the same as LATE for all compliers.

The result that LATE equals the average effect of screening on all screened subjects in a trial with no always-takers is formalized by writing:

$$\begin{aligned} E[Y_{1i} - Y_{0i} \mid S_i = 1] &= E[Y_{1i} - Y_{0i} \mid S_{1i} = 1, Z_i = 1] \\ &= E[Y_{1i} - Y_{0i} \mid S_{1i} = 1] \\ &= E[Y_{1i} - Y_{0i} \mid C_i = 1]. \end{aligned} \quad [6]$$

The first equality uses the fact that, with no always-takers, $S_i = S_{1i}Z_i$; the second uses the fact that Z_i is independent of potential outcomes and potential adherence; the third uses the fact that $C_i = S_{1i} - S_{0i} = S_{1i}$ when $S_{0i} = 0$ for everyone.

When applied to a randomized screening trial, the LATE theorem, Eq. **5**, turns only on the claims that random assignment to screening: i) makes screening more likely on average, ii) never inhibits screening, and iii) affects outcomes solely by making screening more likely. This contrasts with the arguments underpinning old-fashioned per-protocol adjustments. An "as-treated" analysis (as in, e.g., refs. 26 and 27) ignores experimental random assignment, comparing outcomes by screening status, S_i, as if the latter was randomized as intended. But comparisons of $E[Y_i \mid S_i = 1]$ and $E[Y_i \mid S_i = 0]$ in a trial with partial adherence are confounded for the same reason that comparisons by treatment status in cohort studies are confounded. In a pragmatic trial, where patients and their doctors freely choose adherence, potential outcomes cannot be presumed to be independent of adherence.

An alternative non-IV estimation strategy (seen, e.g., in ref. 7) compares all randomly assigned controls to treated subjects who are screened as intended. This amounts to a comparison of $E[Y_i \mid S_i = 1, Z_i = 1]$ with $E[Y_i \mid Z_i = 0]$, which differs from an as-treated analysis in that it discards subjects for whom $S_{1i} = 0$, rather than moving them to a putative control group defined by screening status. Ref. 14 labels subjects with $S_{1i} = 0$ "never-takers" because they remain unscreened regardless of their assignment. When no one assigned to control is screened, the per-protocol estimator discarding never-takers is given by:

$$\begin{aligned} &E[Y_i \mid S_i = 1, Z_i = 1] - E[Y_i \mid Z_i = 0] \\ &= E[Y_{1i} \mid S_{1i} = 1, Z_i = 1] - E[Y_{0i} \mid Z_i = 0] \\ &= E[Y_{1i} \mid S_{1i} = 1] - E[Y_{0i}], \end{aligned}$$

where the last equality uses random assignment of Z_i. Because the two expectations contrasted here involve different groups, this is not an apples-to-apples comparison.**

It is remarkable and even surprising that in a trial with no control-group crossovers, IV estimates

$$\frac{\rho}{\pi} = E[Y_{1i} - Y_{0i} \mid S_i = 1],$$

while, at the same time, $E[Y_i \mid S_i = 1, Z_i = 1] - E[Y_i \mid Z_i = 0]$ is compromised by selection bias. When the decision to comply is a matter of choice rather than chance, omission of never-takers is consequential because adherents in the group invited to screening may be special. In the NORCCAP trial, for instance, the Holme et al. (28) supplement notes that "some baseline characteristics (e.g., gender, area of residency, ethnic background, income, education, and marital status) are strong predictors of adherence." The resulting selection bias can go either way. For instance, NORCCAP adherents are relatively educated, and therefore likely to be relatively healthy whether screened or not. But they are also older and more likely to be male, elevating risk. Beyond such demographic differences, adherence may be motivated by chronic health concerns such as diabetes, a history of polyps, or a family history that elevates CRC risk.

2. Colonoscopy Screening Trials

A. Background. We apply IV to estimates from five trials meant to gauge the impact of CRC screening. Trials considered here include the four featured in an influential meta-analysis (29) plus NordICC, for which 10-y widely noted follow-up results were recently released (7). Screening treatments evaluated in these trials include colonoscopy (which examines the entire colon), sigmoidoscopy (which examines the lower colon and is relatively rare in the United States), and sigmoidoscopy plus fecal occult blood testing (FOBT).

The five trials of interest recruited and screened subjects in various ways. NordICC participants were drawn from population registries in Poland, Norway, and Sweden, with the sample limited to men and women 55 to 64 y of age who had not previously undergone screening, excluding people diagnosed with CRC. NordICC is the only one of our trials to offer initial colonoscopy screening rather than initial sigmoidoscopy. NORCCAP likewise randomly assigned participants directly from the Norwegian population registry, offering sigmoidoscopy in one group and sigmoidoscopy plus FOBT in another (we pool these treatment groups). The other three trials randomly assigned treatment to people who expressed interest in participating in a screening trial when surveyed. Finally, the American PLCO trial offered 2 sigmoidoscopy screening examinations to subjects recruited in various ways by mostly university-based cancer screening centers. Table 1 summarizes these and other key facts related to study populations, screening modalities, trial design, and adherence.††

Recent applications of IV methods to CRC screening trials include methodological studies such as Swanson et al. (25), which illustrates an IV-inspired bounding computation using NORCCAP; and Lee et al. (30), which uses PLCO data to illustrate a new IV procedure for estimation of survival models. Substantive trial analyses using IV include Holme et al. (28, 31, 32), which report IV estimates for the NORCCAP trial; Senore et al. (33), which reports IV estimates for the SCORE trial; and ref. 7, which comments briefly on an IV-based "sensitivity analysis." As far as we know, published IV analyses of the four European trials to date fail to note that IV recovers average causal effects on all screened subjects. Except for the two methodological contributions noted above, most screening trial reports feature traditional per-protocol estimates—comparing subjects by treatment received rather than IV. We aim to explain why IV analysis, which shares with ITS a focus on random assignment, offers a uniquely compelling solution to the

**The resulting selection bias is isolated by writing:

$$E[Y_{1i} \mid S_{1i} = 1] - E[Y_{0i}] = E[Y_{1i} - Y_{0i} \mid S_{1i} = 1] + \{E[Y_{0i} \mid S_{1i} = 1] - E[Y_{0i}]\}.$$

††See ref. 7 for a description of NordICC. Juul's (29) meta-analysis (which ignores differential adherence) details the other trials examined here.

adherence problem. We also show how IV tools can be deployed to establish the clinical relevance of IV estimates.

B. IV Estimates. Randomized screening invitations reduced CRC incidence in each of the screening trials summarized in Table 1. This is documented in Table 2, which reports reduced-form ITS estimates of the effect of screening offers on CRC incidence, along with associated SEs.[‡‡] Incidence reductions range from a low of 0.19 percentage points in the NordICC trial to highs of 0.37 percentage points in the UKFSST and PLCO trials. These estimated reductions are significantly different from zero and substantial in relative terms, amounting to roughly 20% of mean CRC incidence in the control groups (reported in column 1 of Table 2).

As with the corresponding reduced-form estimates, first-stage adherence (reported in column 3 of Table 2) varies considerably across trials. The IV estimates shown in column 4 of the table adjust for nonadherence by dividing reduced-form estimates by the corresponding first-stage estimates. The fact that IV estimates, at around 0.0045, on average, are larger than ITS estimates boosts the case for screening as a cancer mitigation strategy. The LATE interpretation of IV implies that the population induced to screen by efforts to promote screening can expect to enjoy cancer risk reductions given by the larger IV estimates rather than the diluted ITS effects. In other words, when weighing trade-offs presented by screening, IV estimates capture the benefit most relevant for patients and their doctors. Moreover, from a public health perspective, the number needed to screen among NordICC compliers is 227, roughly half the number needed to invite to screening (455) reported in ref. 7.

Importantly, IV flattens much of the cross-trial variation in impact. Focusing again on the contrast between NordICC and UKFSST, for which ITS estimates differ by a factor of two, the corresponding IV estimates differ by only 0.08 percentage points—a much smaller and statistically insignificant gap. Likewise, the gap between average ITS for the two Norwegian trials and average ITS for the relatively precisely estimated UKFSST and PLCO trials falls from a statistically significant 0.16 points ($P = 0.007$) to a statistically insignificant 0.08 points ($P = 0.52$) for IV. We return to this pattern in the discussion of IV tools, below.

In two of the five trials, old-fashioned as-treated and per-protocol analysis omitting never-takers miss the IV impact estimate. This is shown in columns 5 and 6 of Table 2: old-fashioned per-protocol effects, amounting to 0.21 and 0.24 percentage points in the NordICC trial, are much closer to the ITS estimate than to the markedly larger IV estimate of 0.44 percentage points. Likewise, old-fashioned per-protocol analysis of NORCCAP data yields estimates around 0.24 percentage points, close to the ITS effect of 0.22, while the corresponding IV estimate is 59% larger (0.35). This shortfall in as-treated estimates may be explained by the fact that experimental subjects who take up screening are older and more likely to have risk-elevating health concerns than the overall study population.

Discrepant per-protocol estimates for NordICC and NORCCAP may also reflect differences in the way these samples were recruited (both drew subjects from population registries, while other trials identified relevant subjects using surveys.) Old-fashioned per-protocol estimates for the three other trials are similar to the corresponding IV estimates, suggesting selection bias is not a foregone conclusion. Without IV estimates as a point of comparison, however, we'd never know for sure. IV adjusts for nonadherence without risk of selection bias from both unobserved or observed factors.

3. Establishing Clinical Relevance: An IV Toolkit

A regression of the reduced-form ITS estimates in column 2 of Table 2 on the corresponding first-stage estimates (reported in column 3) yields an R^2 of around 0.63. This descriptive fact hints at the possibility that adherence explains much of the variance in ITS effects. Three IV tools—visual instrumental variables (VIV), overidentification testing, and the distribution of complier characteristics—help examine this claim. The results of this examination support the external validity, and therefore the clinical relevance, of IV estimates of CRC screening effects.

A. Visualizing IV. The IV toolkit is applied to estimates from five trials and to estimates for experimental strata in three of the trials. The parameters to be reconciled are pairs of reduced-form and first-stage coefficients (ρ_j, π_j) indexed by $j = 1, \ldots, J$. The corresponding estimates are denoted by $\hat{\rho}_j$ and $\hat{\pi}_j$. Within-trial results are the reduced-form and first-stage estimates for three NordICC countries (Poland, Norway, and Sweden), two NORCCAP regions (Oslo and Telemark), and 10 PLCO centers. Adding full-sample estimates for SCORE and UKFSST, while deleting data points for the full NordICC, NORCCAP, and PLCO samples to avoid duplication, leaves a total of $J = 17$ pairs of estimates.

VIV provides a graphical summary of the variation in $(\hat{\rho}_j, \hat{\pi}_j)$ along with an overall estimate of screening effects. Note first that if screening effects are similar across trials, reduced-form and first-stage parameters are roughly proportional:

$$\rho_j \approx \lambda \pi_j;\ j = 1, \ldots J; \qquad [7]$$

where λ is the common LATE for screening compliers. This proportionality hypothesis motivates a linear regression of estimated reduced forms on estimated first stages, with no intercept:

$$\hat{\rho}_j = \lambda \hat{\pi}_j + \eta_j. \qquad [8]$$

Regression residual η_j reflects estimation error in $\hat{\rho}_j$ and $\hat{\pi}_j$, as well as approximation error when the proportionality restriction fails due to screening effect heterogeneity.[§§]

VIV plots $\hat{\rho}_j$ against $\hat{\pi}_j$, along with the line of best fit suggested by Eq. **8**. The slope of this line is an estimate of the common LATE for screening. This estimate is consistent for λ when the proportionality restriction Eq. **7** holds exactly, since η_j is a function of estimation error with probability limit zero as sample sizes grow. Otherwise, $\hat{\lambda}_{VIV}$ estimates a weighted average of trial- and strata-specific LATEs given by $\lambda_j = \rho_j / \pi_j$.

When the VIV slope is estimated by weighted least squares with weights proportional to the sample size times the within-trial variance of Z_i, $\hat{\lambda}_{VIV}$ is a two-stage least squares (2SLS) estimator of λ.[¶¶] 2SLS is a powerful and flexible estimation strategy that combines multiple instruments to produce a single, more precise IV estimate than would be obtained using the instruments one at a time. 2SLS also accommodates covariates—in this case, a set of dummy variables indicating the observations contributed by

[‡‡] Except for PLCO, for which we obtained anonymized microdata, estimates and SEs reported here are computed using published trial results. See *SI Appendix*, section 2 for details.

[§§] Applications of VIV to model validation include Angrist (34) and Angrist et al. (35). Angrist and Pischke (36) sketches the underlying econometric theory.

[¶¶] *SI Appendix*, section 3 details 2SLS and derives these weights. Intuitively, 2SLS weights reflect the fact that, under classical regression assumptions, the variance of the reduced-form estimate for each trial is inversely proportional to trial size times the within-trial variance of the instrumental variable, Z_i.

each trial and stratum in a data set that stacks data from all trials and strata.

Fig. 1*A* shows a VIV plot for the 17 groups examined here; whiskers in the plot indicate 95% CIs for reduced-form estimates. While some reduced-form estimates are more precise than others, overall, they tend to decline linearly with estimated first-stage adherence rates (The reduced-form and first-stage estimates plotted in this figure appear in *SI Appendix*, Table 1, along with estimated SEs). Fit with no intercept and using 2SLS weights, the VIV regression line in the figure has slope $\hat{\lambda}_{VIV} = -0.0047$, with an estimated SE of 0.0019. This estimate of the effect of screening on CRC incidence is close to the median of the group-specific IV estimates in Table 2.

The VIV line fits both cross-trial and within-trial estimates remarkably well. Low NordICC adherence, for instance, is associated with modest cancer reductions while high PLCO adherence is associated with larger CRC declines. NORCCAP, SCORE, and UKFSST, with middling adherence, also yield middling CRC impact. This consistent pattern is especially striking in view of the fact that NordICC assigned initial colonoscopy screening, while other trials offered sigmoidoscopy. It is also noteworthy that within NordICC the leftmost blue triangle marks low adherence and impact for Poland, with both impact and adherence roughly twice as large for Norway. A very noisy estimate for Sweden (reflecting a small sample size) sits well above the VIV line but is not statistically distinguishable from it. A few outlying screening effects for PLCO likewise have CIs covering the 2SLS line.

In marked contrast with reduced-form ITS estimates, IV estimates of CRC screening LATEs are unrelated to adherence. This is documented in Fig. 1*B*, which plots $\hat{\lambda}_j = \hat{\rho}_j/\hat{\pi}_j$ against first-stage adherence. The line fit to these points (again using 2SLS weights, though now allowing for an intercept) has a slope of indistinguishable from zero with a SE of 0.0039. In other words, adjusting ITS estimates for differential adherence fully explains the strong negative relationship between adherence and impact seen in Panel *A*.

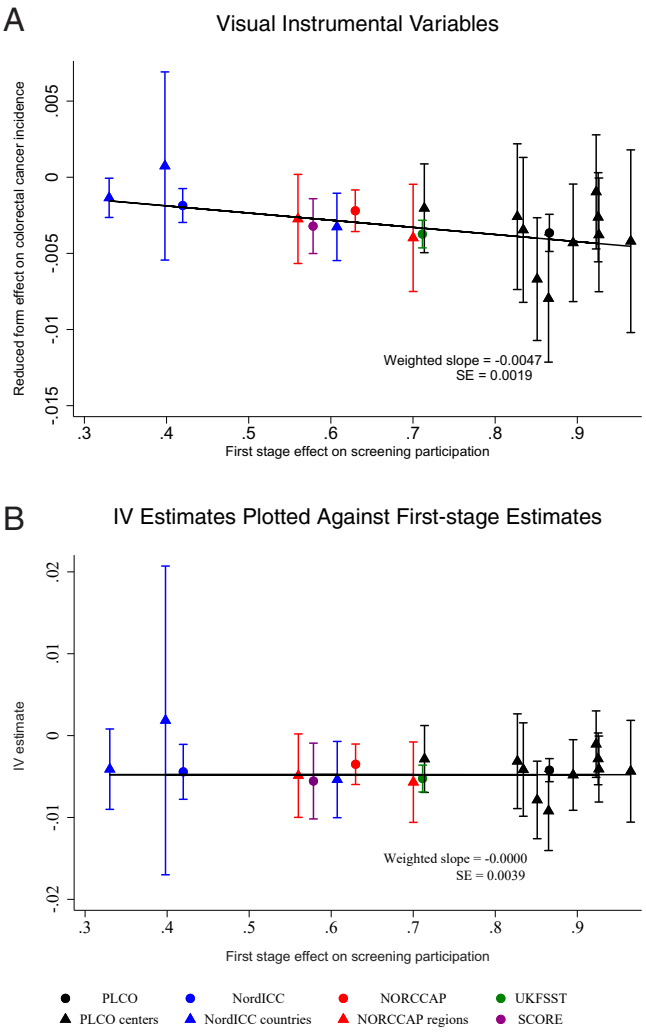

Fig. 1. Panel *A* of this figure plots reduced-from estimates of effects of screening invitations on colorectal cancer (CRC) diagnosis against first-stage estimates for 17 groups derived from 5 pragmatic trials. Panel *B* plots the corresponding IV estimates of screening effects. The mediating variable for IV is screening participation. Samples are for 3 NordICC countries, 2 NORCCAP regions, 10 PLCO centers, and for all randomized participants in each of UKFSST, NORCCAP, SCORE, NordICC, and PLCO. These trials randomly offered participating subjects sigmoidoscopy or colonoscopy screening, in populations that are otherwise unlikely to screen. CRC incidence is measured 10 to 12 y after random assignment. Regression lines plotted in the figures are weighted by $N_j p_j (1 - p_j)$ where N_j is the sample size and p_j is the offer rate. The VIV line in Panel *A* is fit without an intercept. Whiskers mark 95% CIs.

B. Overidentification Testing. The VIV line plotted in Fig. 1*A* yields a good though imperfect fit. Under the proportionality hypothesis expressed by Eq. **7**, anything less than a perfect fit is due to sampling variance in the underlying estimates. Can the fact that the VIV fit is imperfect indeed be put down to sampling variance alone? An overidentification test statistic answers this question.

In the context of the estimates in our VIV plot, the overidentification test is a goodness-of-fit statistic that can be written:

$$\hat{T} = \sum\nolimits_j (1/\hat{\sigma}_j^2)(\hat{\rho}_j - \hat{\lambda}_{VIV}\hat{\pi}_j)^2, \qquad [9]$$

where $\hat{\sigma}_j^2$ denotes the estimated sampling variance of $\hat{\rho}_j - \hat{\lambda}_{VIV}\hat{\pi}_j$. Under the proportionality null hypothesis, $\hat{T}$ has an asymptotic chi-square distribution with degrees of freedom given by the number of restrictions being tested. A single trial is enough to compute one LATE; two trials can be used to estimate two LATEs. The proportionality restriction implying that these are equal yields $2 - 1 = 1$ degree of freedom. More generally, when data from J trials and strata are used to estimate a single λ, we're imposing (and therefore testing) $J - 1$ restrictions. The null hypothesis is rejected when the overidentification test statistic is surprisingly large relative to a $\chi^2(J-1)$ distribution. In other words, the test rejects when deviations from the VIV line in Fig. 1*A* are too large to be attributed to sampling variance in the estimates (Ref. 36 details the theory behind overidentification testing. In an antecedent of the overidentification test applied here, Glasziou (37) tests for homogeneity of IV estimates in a meta-analysis of the effects of mammography in five breast cancer screening trials).

Over-identification test statistics, reported in Table 3, along with associated degrees of freedom and *P*-values, indicate that the proportionality hypothesis fits the reduced-form and first-stage estimates well (*SI Appendix*, section 4 details the calculation of these test statistics). The first row of the table reports test results for all groups used to fit the VIV line in Fig. 1, yielding a test value of around 12 and a *P*-value of 0.74. Test statistics in remaining rows evaluate the proportionality restriction across the five trials while pooling strata within trails, and for estimates across strata within NordICC, NORCCAP, and PLCO. Consistent with the impression made by the figure, no test weighs against the hypothesis of a stable per-protocol screening effect. By contrast,

Table 3. Overidentification tests

	Test statistic (1)	D.f. (2)	*P*-value (3)
All sites	12.03	16	0.74
All studies	1.79	4	0.77
NordICC countries	0.60	2	0.74
NORCCAP regions	0.05	1	0.83
PLCO centers	10.08	9	0.34

Notes: This table reports overidentification test statistics computed as described in *SI Appendix*, section 4, along with the associated degrees of freedom and *P*-values. The NordICC overidentification test statistic compares IV estimates for 3 countries (Poland, Norway, and Sweden). The NORCCAP overidentification test statistic compares IV estimates for 2 regions (Oslo and Telemark). The PLCO overidentification test statistic compares IV estimates across 10 PLCO screening centers.

an analogous comparison of reduced-form ITS estimates (testing whether these are constant across 5 trials) generates a P-value of 0.06; dropping the relatively imprecise SCORE reduced form sharpens the rejection of constant ITS effects further ($P = 0.03$).

C. Characterizing Compliers. IV overcomes the problem of selection bias in old-fashioned per-protocol estimates, but self-selection into adherence can still limit the clinical relevance of IV estimates. If, for instance, a particular demographic group is substantially under-represented among compliers, LATEs might be seen as being of limited value for this group. On the other hand, when all groups of interest are well-represented among LATE compliers, IV estimates of screening effects are more likely to predict screening effects beyond the trials that produced them. Our third IV tool is a simple estimators of complier characteristics.

With no always-takers, complier characteristics are revealed by the characteristics of screened participants. To be precise, consider a screening trial that collects data on subject characteristics, such as demographic information, socioeconomic background, and baseline health, summarized in a covariate vector with generic element X_i. In general, the complier mean of this characteristic is defined as $E[X_i \mid C_i = 1]$. When $S_{0i} = 0$ for all i and Z_i is independent of X_i, $E[X_i \mid C_i = 1] = E[X_i \mid S_i = 1]$. This point parallels the result highlighted in Section C, showing that LATE equals the average screening effect on the screened in a trial with no control-group crossovers.[##]

Allowing for control-group crossovers, we must contend with the fact that $C_i = S_{1i} - S_{0i}$ is unobserved, since only one of the two potential adherence variables is seen for each subject. Even so, complier means are easily estimated. To see this, consider an IV estimand (reduced form divided by first stage) with S_iX_i replacing the outcome variable Y_i. Because X_i is independent of Z_i, this new IV estimand can be simplified as:

$$\begin{aligned}\frac{E[S_iX_i \mid Z_i = 1] - E[S_iX_i \mid Z_i = 0]}{E[S_i \mid Z_i = 1] - E[S_i \mid Z_i = 0]} &= \frac{E[S_{1i}X_i] - E[S_{0i}X_i]}{E[S_{1i}] - E[S_{0i}]} \\ &= \frac{E[(S_{1i} - S_{0i})X_i]}{E[S_{1i} - S_{0i}]} \\ &= E[X_i \mid C_i = 1]. \quad \textbf{[10]}\end{aligned}$$

Complier mean X_i equals LATE for the effects of S_i on dependent variable S_iX_i.

[##] When treatment assignment rates differ within strata, as in the NORCCAP trial studied here, screening offers are independent of X_i only within strata. A consequence of this is that complier means may diverge from treated means even without always-takers. This point is fleshed out in *SI Appendix*, section 5, which shows how to compute complier means in stratified trials.

In the five CRC trials considered here, compliers have demographic characteristics broadly representative of trial participants at large. This is documented in Fig. 2, which compares complier means with the average X_i in full study samples for dummy variables indicating female and younger participants, and, for NORCCAP, a dummy indicating Oslo residents. Although there are some differences (compliers tend to be older and are more likely to be male), both demographic groups are well-represented among compliers in each study. Oslo residents are somewhat underrepresented among NORCCAP compliers, but not dramatically so. A similar computation comparing baseline health of PLCO compliers with those in the overall study sample also shows these groups to be comparable.

4. Summary and Conclusions

IV analysis of cancer screening trials offers an easily navigated path from ITS effects of screening invitations to credible per-protocol estimates of the causal effects of screening itself. Applied to five CRC screening trials with substantial nonadherence, IV methods reconcile divergent ITS effects with an estimated CRC incidence reduction from screening of nearly half a percentage point. Efforts to promote CRC screening would do well to feature this as the expected benefit for subjects who screen. It is also noteworthy that the US Preventive Services Task Force (USPSTF) marks trial evidence down due to "inconsistency of findings across individual studies" (https://www.uspreventiveservicestaskforce.org/uspstf/about-uspstf/methods-and-processes/grade-definitions). IV estimates showing consistent effects on subjects actually screened may therefore prompt an evidence quality upgrade.

Economists have long used IV to address nonadherence and other sources of selection bias in wide-ranging settings. Although IV ideas have also filtered into medical statistics, dissemination on the clinical side has been surprisingly slow. The gap across disciplines partly reflects missing data. For instance, a fair proportion of the PLCO control-group appears to have been screened some time after random assignment. Yet, estimates using PLCO data (including ours) ignore this fact since information on screening for the full study sample is unavailable (38). Kowalski's (2) IV analysis of the CNBSS mammography screening experiment uses information on screening among controls, but CNBSS appears to

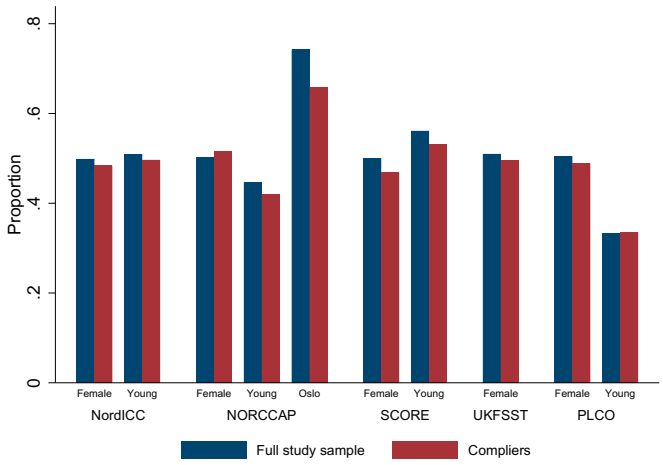

Fig. 2. This figure compares the sex, age, and region distribution for full study samples and screening compliers. Complier means are computed as described in the text. Bars show sample proportions (dummy variable means) in the groups indicated on the x-axis. Young refers to age group 50 to 54 for NORCCAP and to 55 to 59 for NordICC, SCORE, and PLCO.

be the only trial cited in the USPSTF mammography guidelines that identifies all such always-takers.

Short-sighted data collection is not limited to cancer screening; the landmark mRNA COVID-19 vaccine trial likewise neglects information on post-randomization vaccination among most controls (39). In addition to promoting use of IV, we hope that our work encourages routine monitoring of treatment status for all trial subjects, identifying treatments received in both experimental and control groups, whether treated per protocol or not.

Data, Materials, and Software Availability. Data used here are drawn from published and supplemental information in refs. 7, 20, 21, and 28. See Supplementary Information for details. In addition, PLCO microdata were obtained from https://cdas.cancer.gov/plco/ (40), obtained as described in https://cdas.cancer.gov/learn/plco/instructions/, received on Feb 15, 2023.

ACKNOWLEDGMENTS. We thank Edoardo Botteri, Amy Finkelstein, Guido Imbens, Amanda Kowalski, Emily Oster, and Robert Yeh for helpful comments. Carol Gao provided excellent research assistance. The Massachusetts Institute of Technology Institutional Review Board (Committee on the Use of Humans as Experimental Subjects, the Committee on the Use of Humans as Experimental Subjects) deems the research reported here to be exempt from review.

Author affiliations: [a]Department of Economics and National Bureau of Economic Research, Massachusetts Institute of Technology, Cambridge, MA 02142; and [b]Department of Economics and National Bureau of Economic Research, Brown University, Providence, RI 02912

1. A. E. Kowalski, Mammograms and mortality: How has the evidence evolved? *J. Econ. Perspect.* **35**, 119–140 (2021).
2. A. E. Kowalski, Behaviour within a clinical trial and implications for mammography guidelines. *Rev. Econ. Stud.* **90**, 432–462 (2023).
3. J. H. Hayes, M. J. Barry, Screening for prostate cancer with the prostate-specific antigen test: A review of current evidence. *JAMA* **311**, 1143–1149 (2014).
4. J. A. Dominitz, D. J. Robertson, Understanding the results of a randomized trial of screening colonoscopy. *New Engl. J. Med.* **387**, 1609–1611 (2022).
5. M. Roland, D. J. Torgerson, Understanding controlled trials: What are pragmatic trials? *BMJ* **316**, 285 (1998).
6. B. Gyawali, A controversial trial: Exposing misunderstandings of NordICC. *Medscape* (2022). https://www.medscape.com/viewarticle/982479 (Accessed 24 May 2023).
7. M. Bretthauer *et al*., Effect of colonoscopy screening on risks of colorectal cancer and related death. *New Engl. J. Med.* **387**, 1547–1556 (2022).
8. S. J. Winawer, Colonoscopy screening and colorectal cancer incidence and mortality. *New Engl. J. Med.* **388**, 376 (2023).
9. M. Bretthauer, M. Løberg, M. F. Kaminski, Colonoscopy screening and colorectal cancer incidence and mortality. *New Engl. J. Med.* **388**, 376 (2023).
10. N. Angrist, R. Meager, Implementation matters: Generalizing treatment effects in education. Available at SSRN 4487496 (2023).
11. G. W. Imbens, D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, Cambridge, 2015).
12. C. M. Rembold, Number needed to screen: Development of a statistic for disease screening. *BMJ* **317**, 307–312 (1998).
13. G. W. Imbens, J. D. Angrist, Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–475 (1994).
14. J. D. Angrist, G. W. Imbens, D. B. Rubin, Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* **91**, 444–455 (1996).
15. H. S. Bloom, Accounting for no-shows in experimental evaluation designs. *Eval. Rev.* **8**, 225–246 (1984).
16. R. Newcombe, Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur. *Stat. Med.* **7**, 1179–1186 (1988).
17. N. Hearst, T. B. Newman, S. B. Hulley, Delayed effects of the military draft on mortality. *New Engl. J. Med.* **314**, 620–624 (1986).
18. S. G. Baker, K. S. Lindeman, The paired availability design: A proposal for evaluating epidural analgesia during labor. *Stat. Med.* **13**, 2269–2278 (1994).
19. S. G. Baker, B. S. Kramer, K. S. Lindeman, Latent class instrumental variables: A clinical and biostatistical perspective. *Stat. Med.* **35**, 147–160 (2016).
20. W. S. Atkin *et al*., Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: A multicentre randomised controlled trial. *Lancet* **375**, 1624–1633 (2010).
21. N. Segnan *et al*., Once-only sigmoidoscopy in colorectal cancer screening: Follow-up findings of the Italian randomized controlled trial-SCORE. *J. Natl. Cancer Inst.* **103**, 1310–1322 (2011).
22. J. Cuzick, R. Edwards, N. Segnan, Adjusting for non-compliance and contamination in randomized clinical trials. *Stat. Med.* **16**, 1017–1029 (1997).
23. G. W. Imbens, D. B. Rubin, Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.* **64**, 555–574 (1997).
24. M. A. Hernán, J. M. Robins, Per-protocol analyses of pragmatic trials. *New Engl. J. Med.* **377**, 1391–1398 (2017).
25. S. A. Swanson *et al*., Bounding the per-protocol effect in randomized trials: An application to colorectal cancer screening. *Trials* **16**, 541 (2015).
26. G. Chêne *et al*., Intention-to-treat vs. on-treatment analyses of clinical trial data: Experience from a study of pyrimethamine in the primary prophylaxis of toxoplasmosis in HIV-infected patients. *Control. Clin. Trials* **19**, 233–248 (1998).
27. D. L. Packer *et al*., Effect of catheter ablation vs. antiarrhythmic drug therapy on mortality, stroke, bleeding, and cardiac arrest among patients with atrial fibrillation: The CABANA randomized clinical trial. *JAMA* **321**, 1261–1274 (2019).
28. Ø. Holme *et al*., Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: A randomized clinical trial. *JAMA* **312**, 606 (2014).
29. F. E. Juul *et al*., 15-Year benefits of sigmoidoscopy screening on colorectal cancer incidence and mortality: A pooled analysis of randomized trials. *Ann. Int. Med.* **175**, 1525–1533 (2022).
30. Y. Lee, E. H. Kennedy, N. Mitra, Doubly robust nonparametric instrumental variable estimators for survival outcomes. *Biostat (Oxford, England)* **24**, 518–537 (2023).
31. Ø. Holme *et al*., Effectiveness of flexible sigmoidoscopy screening in men and women and different age groups: Pooled analysis of randomised trials. *BMJ* **356**, 1–8 (2017).
32. Ø. Holme *et al*., Long-term effectiveness of sigmoidoscopy screening on colorectal cancer incidence and mortality in women and men: A randomized trial. *Ann. Int. Med.* **168**, 775–782 (2018).
33. C. Senore *et al*., Long-term follow-up of the Italian Flexible Sigmoidoscopy Screening Trial. *Ann. Int. Med.* **175**, 36–45 (2022).
34. J. D. Angrist, Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *Am. Econ. Rev.* **80**, 313–336 (1990).
35. J. D. Angrist, P. A. Pathak, R. A. Zárate, Choice and consequence: Assessing mismatch at Chicago exam schools. *J. Public Econ.* **223**, 104892 (2023).
36. J. D. Angrist, J. S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, Princeton, NJ, 2009).
37. P. Glasziou, Meta-analysis adjusting for compliance: The example of screening for breast cancer. *J. Clin. Epidemiol.* **45**, 1251–1256 (1992).
38. R. E. Schoen *et al*., Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. *New Engl. J. Med.* **366**, 2345–2357 (2012).
39. H. M. El Sahly *et al*., Efficacy of the mRNA-1273 SARS-CoV-2 vaccine at completion of blinded phase. *New Engl. J. Med.* **385**, 1774–1785 (2021).
40. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial, PLCO Colorectal Datasets. Cancer Data Access System. https://cdas.cancer.gov/datasets/plco/22/. Accessed 15 February 2023.