IV METHODS RECONCILE INTENTION-TO-SCREEN EFFECTS ACROSS PRAGMATIC CANCER SCREENING TRIALS

Joshua Angrist
Peter Hull

WORKING PAPER 31443

NBER WORKING PAPER SERIES

IV METHODS RECONCILE INTENTION-TO-SCREEN EFFECTS ACROSS PRAGMATIC CANCER SCREENING TRIALS

Joshua Angrist
Peter Hull

Working Paper 31443
http://www.nber.org/papers/w31443

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2023, Revised September 2023

We thank Edoardo Botteri, Amy Finkelstein, Guido Imbens, Amanda Kowalski, and Emily Oster for helpful comments. Carol Gao provided excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

IV Methods Reconcile Intention-to-Screen Effects Across Pragmatic Cancer Screening Trials
Joshua Angrist and Peter Hull
NBER Working Paper No. 31443
July 2023, Revised September 2023
JEL No. C21,C26,C93,I12

ABSTRACT

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 actually 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 broadly similar to those of the full trial study sample. These findings argue for the clinical relevance of IV estimates of cancer screening effects.

Joshua Angrist
Department of Economics, E52-436
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
angrist@mit.edu

Peter Hull
Department of Economics
Box B, Brown University
Providence RI 02912
and NBER
peter_hull@brown.edu

1 Introduction

The question of whether cancer screening improves health remains contentious—a fact highlighted by the debate over mammography, prostate-specific antigen (PSA) screening, and colorectal cancer (CRC) screening.[1] 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 don't screen are likely to differ in many ways besides the 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.[2]

Pragmatic randomized trials, meant to "measure effectiveness in routine clinical practice" [Roland and Torgerson, 1998], 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, the evaluation of unpleasant and time-consuming medical interventions under real-world conditions is usually complicated by the fact that many patients fail to take their doctor's advice. It's 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 comparing groups 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 saw 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 [Gyawali, 2022]. In one of many letters responding to the Bretthauer et al. [2022] report on NordICC, correspondent Winawer [2023] 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're unconfounded by pre-treatment differences between those assigned to screening and control groups. Yet, as Winawer [2023] suggests, nonadherence makes ITS effects hard to compare across studies and even harder to apply to public health policy. Intuitively, low adherence dilutes ITS effects by including people whose screening behavior is unaffected by the offer in the treatment group. The number need to *screen* to prevent cancer

[1] See, e.g., Kowalski [2021, 2023] for discussions of mammography, Hayes and Barry [2014] on PSA, and references in the CRC studies cited below.

[2] A recent NEJM editorial on CRC screening (Dominitz and Robertson, 2022) 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."

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 10-12 years 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.[3] In practice, ITS effects on incidence vary. The estimated ITS effect in the NordICC trial is 0.19 percentage points (reported as -0.0019 in the table), while data from the UKFFST trial yield an ITS estimate that's twice as big. Given the statistical precision with which these effects are estimated, with robust standard errors as small as 0.0005, these large differences in impact are unlikely to be due to chance alone.

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 55-64 in European countries offering 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. Section 3 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.[4] In support of this claim, Section 4 deploys three IV tools not previously applied in this context: visual instrumental variables, overidentification testing, and complier characteristics. Section 5 summarizes our argument and draws some conclusions.

2 The IV Advantage

2.1 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

[3] Our followup horizon matches that in Bretthauer et al. [2022]. We focus on CRC incidence over mortality because estimates for the former are more precise, a point noted by Bretthauer, Løberg and Kaminski [2023].

[4] Angrist and Meager [2023] makes an analogous point in the context of schooling-related interventions in developing countries, where mediating instrumental variables gauge program implementation.

(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 (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., Imbens and Rubin [2015]). 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 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:

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

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 it 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, $E[Y_{1i} - Y_{0i}]$:

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

The first equality follows from the potential outcomes model and $S_i = Z_i$; the second follows from the random assignment of Z_i; 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 $E[Y_{1i} - Y_{0i}]$ further 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're unscreened. Clinicians and public health officials often prioritize this measure of impact, which reveals whether people screened in a trial can expect to have fewer cancers as a result of screening. 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." This is the number of patients that must be screened, on average, to prevent one CRC case [Rembold, 1998].[5]

[5] With no always-takers, CRC case counts fall by $N\lambda$ when N are screened and LATE is λ. To prevent one CRC case, screen N^* such that $1 = N^*\lambda$. The number needed to screen is therefore $N^* = \frac{1}{\lambda}$.

2.2 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's 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:

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

The causal effect of screening offers on screening behavior is given by $S_{1i} - S_{0i}$.

The local average treatment effects (LATE) model, introduced in Imbens and Angrist [1994] and Angrist, Imbens and Rubin [1996], 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're 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 randomized 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.[6] 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 all potential outcomes in 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 equations (1) and (2) to show that randomized offers determine outcomes according to:

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

[6] 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 $d \in \{0, 1\}$.

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

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

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 that generating the reduced form:

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

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

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

Intuitively, π 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 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 equations (3) and (4) and the fact that C_i is a dummy to write:

$$\frac{\rho}{\pi} = \frac{E[C_i(Y_{1i} - Y_{0i})]}{E[S_{1i} - S_{0i}]} = \frac{E[Y_{1i} - Y_{0i} \mid C_i = 1]Pr(C_i = 1)}{Pr(C_i = 1)} = E[Y_{1i} - Y_{0i} \mid C_i = 1]. \tag{5}$$

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 exclusion, 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.[7] 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 discussed more in Section 4. 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).

[7]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 in large samples.

2SLS also provides an immediate path to off-the-shelf statistical inference.[8]

Hernán and Robins [2017] note that it's 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 i. Even in a trial in which no randomized controls cross over to receive 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, reaearchers and other observers need not identify individual compliers. Rather, these observers are 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? IV tools detailed in Section 4 answer these questions.

2.3 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" [Swanson et al., 2015]. LATE is also per-protocol effect, but not for everyone: as equation (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 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:

- Those with $S_{0i} = 1$, in which case monotonicity implies $S_{1i} = 1$ as well. Angrist, Imbens and Rubin [1996] call 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, $S_{0i} = 0$ for all i meaning 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.

[8] IV ideas applied to randomized trials appear in alternate forms in social science and medicine without referencing IV or potential outcomes. Bloom [1984] adjusts trial data for treated never-takers using equation (6), below. Newcombe [1988] derives an adjustment for randomized trials with control-group crossovers. Hearst, Newman and Hulley [1986] use similar reasoning to obtain effects of Vietnam-era military service using the American draft lottery. Baker and Lindeman [1994] and Baker, Kramer and Lindeman [2016] 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. [2010] and Segnan et al. [2011], reference an adherence adjustment due to Cuzick, Edwards and Segnan [1997]. Also focusing on Bernoulli outcomes, the latter derives a maximum likelihood estimator that adjusts risk ratios for nonadherence. The Cuzick, Edwards and Segnan [1997] estimator is an instance of results in Imbens and Rubin [1997], which uses IV to compute marginal distributions of potential outcomes for 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:

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

The first equality uses the fact that $S_i = S_{1i}Z_i$ with no always-takers; 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 result (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., Chêne et al. [1998] and Packer et al. [2019]) 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 are unlikely to be independent of adherence.

An alternative non-IV estimation strategy (seen, e.g., in Bretthauer et al. [2022]) 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. Angrist, Imbens and Rubin [1996] call 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:

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

using random assignment of Z_i. Because the two conditional expectations contrasted on the second line involve different groups, this is not an apples-to-apples comparison.[9]

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

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

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. Specifically, 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.

[9] The resulting selection bias appears as the second term in:

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

In the NORCCAP trial, for instance, the Holme et al. [2014] 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 healthier in the absence of screening. But they're also older and more likely to be male, elevating risk. Beyond observable 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.

3 Colonoscopy Screening Trials

3.1 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 a recent meta-analysis [Juul et al., 2022] plus NordICC, for which 10-year follow-up results were recently released [Bretthauer et al., 2022]. 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 US), 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 years 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. Fewer participants in the U.K. (UKFSST) and Italian (SCORE) trials were referred to follow-up colonoscopy screening as a result of an initial sigmoidoscopy screening. 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 regarding study populations, screening modalities, trial design, and adherence.[10]

Recent applications of IV methods to CRC screening trials include methodological studies by Swanson et al. [2015], which illustrates an IV-inspired bounding computation using NORCCAP; and Lee, Kennedy and Mitra [2023], which uses PLCO data to illustrate a new IV procedure for estimation of survival models. Substantive trial analyses using IV include Holme et al. [2014, 2017, 2018], which report IV estimates for the NORCCAP trial; Senore et al. [2022], which reports IV estimates for the SCORE trial; and Bretthauer et al. [2022], which comments briefly on an IV-based "sensitivity analysis." As far as we know, IV analyses of the four European trials to date fail to

[10] See Bretthauer et al. [2022] for a description of NordICC. The Juul et al. [2022] meta-analysis (which ignores differential adherence) details the other trials examined here.

note that IV recovers average causal effects on all screened subjects. Except for the two methodological contributions, most of these (and many other) trial reports feature traditional per-protocol estimates—comparing subjects by treatment received rather than by randomized treatment assigned. We aim to explain why IV analysis, which shares with ITS a focus on random assignment, offers a uniquely compelling solution to the adherence problem. We also show below how IV tools can be deployed to establish the clinical relevance of IV estimates.

3.2 IV Estimates

Randomized screening offers reduced CRC incidence in each of the screening trials summarized in Table 1. This is documented in the second column of Table 2, which reports reduced-form ITS estimates of the effect of screening offers on CRC incidence along with associated standard errors.[11] 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 non-offered control groups (reported in the first column of the table).

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 forms by first stages. The fact that IV estimates are larger than ITS estimates (around 0.45 percentage points on average) 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, half the number needed to *invite* to screening reported in Bretthauer et al. [2022].

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 64% larger (0.36). The shortfall in as-treated estimates may be explained by the fact that experimental subjects who take up screening are older and are more likely to have risk-elevating health concerns than the overall study population.

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

[11] Except for PLCO, for which we obtained anonymized microdata, estimates and standard errors reported here are computed using the published trial results. See Appendix A.2 for details.

of selection bias from unobserved as well as from observed factors.

4 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 support the external validity, and therefore the clinical relevance, of IV estimates of CRC screening effects.

4.1 Visualizing IV

The IV toolkit is applied to estimates from five trials and the 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 offers 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 assumed to be similar across trials, reduced forms and first stages are roughly proportional:

$$\rho_j \approx \lambda \pi_j; \;\; j = 1, ...J; \tag{7}$$

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:

$$\hat{\rho}_j = \lambda \hat{\pi}_j + \eta_j. \tag{8}$$

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.[12]

VIV plots $\hat{\rho}_j$ against $\hat{\pi}_j$, along with the line of best fit suggested by (8). The slope of this line yields an estimate of the common LATE for screening, an estimated denoted $\hat{\lambda}_{VIV}$. This estimate is consistent for λ when the proportionality restriction (7) holds exactly, since η_j is then a sample average with probability limit of 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

[12] Applications of VIV to model validation include Angrist [1990] and Angrist, Pathak and Zárate [2023]. Angrist and Pischke [2009] sketch the underlying econometric theory.

of λ.[13] 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 dummies indicating data from each trial and stratum in a data set that stacks data from all trials and strata.

Panel A of Figure 1 shows a VIV plot for the 17 groups examined here; whiskers in the plot indicate 95% confidence intervals for reduced-form estimates. While some reduced-form estimates are more precise than others, overall, these estimates appear to be a linearly decreasing function of estimated first-stage adherence rates.[14] 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 standard error of 0.0017. This estimate of the effect of screening on CRC incidence is close to the median of the group-specific IV estimates reported 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 drops. 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 an 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 confidence intervals 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 Panel B of Figure 1, 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 effectively zero with a standard error of 0.0039. In other words, adjusting ITS estimates for differential adherence fully explains the strong negative relationship plotted in Panel A.

4.2 Overidentification Testing

The VIV line plotted in Panel A of Figure 1 yields a good but imperfect fit. Under the proportionality hypothesis expressed by (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 a formal test.

In the context of the estimates in our VIV plot, the overidentification test is a goodness-of-fit

[13] Appendix A.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.

[14] The reduced-form and first-stage estimates plotted in this figure appear in Appendix Table A1, along with estimated standard errors.

statistic that can be written:

$$\hat{T} = \sum_j (1/\hat{\sigma}_j^2)(\hat{\rho}_j - \hat{\lambda}_{VIV}\hat{\pi}_j)^2, \tag{9}$$

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 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 Panel A of Figure 1 are too large to be attributed to sampling variance in the estimates.[15]

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.[16] The first row of the table reports test results for all groups used to fit the VIV line in Figure 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.

4.3 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 consists of 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. 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 2.3, that LATE equals the average screening

[15] Angrist and Pischke [2009] detail the theory behind overidentification testing. In an antecedent of the overidentification test applied here, Glasziou [1992] tests for homogeneity of IV estimates in a meta-analysis of the effects of mammography in five breast cancer screening trials.

[16] Appendix A.4 details test statistic calculation.

effect on the screened in a trial with no control-group crossovers.[17]

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 individual i. Even so, complier means are easily estimated. To see this, consider an IV estimand 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:

$$\frac{E[S_iX_i \mid Z_i = 1] - E[S_iX_i \mid Z_i = 0]}{E[S_i \mid Z_i = 1] - E[S_i \mid Z_i = 0]} = \frac{E[S_{1i}X_i] - E[S_{0i}X_i]}{E[S_{1i}] - E[S_{0i}]}$$
$$= \frac{E[(S_{1i} - S_{0i})X_i]}{E[S_{1i} - S_{0i}]} = E[X_i \mid C_i = 1]. \tag{10}$$

Complier mean X_i is therefore given by the LATE theorem applied to dependent variable S_iX_i.

In the five CRC trials considered here, compliers have demographic characteristics broadly representative of trial participants at large. This is documented in Figure 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), all demographic groups are well-represented among compliers in each study. Oslo residents are somewhat underrepresented among NORCCAP compliers, but not dramatically so.

5 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's also noteworthy that the U.S. Preventive Services Task Force (USPSTF) marks trial evidence down due to "Inconsistency of findings across individual studies."[18] 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, progress 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. Yet, estimates using PLCO data (including ours) ignore this fact since information on screening for the full study sample is unavailable [Schoen et al., 2012]. The Kowalski [2023] IV analysis of the CNBSS mammography screening experiment uses information on screening among controls, but

[17] 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 Appendix A.5, which shows how to compute complier means in stratified trials.

[18] See https://www.uspreventiveservicestaskforce.org/uspstf/about-uspstf/methods-and-processes/grade-definitions.

CNBSS appears to be the only trial cited in the USPSTF mammography guidelines that identifies 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 controls [El Sahly et al., 2021]. 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 provided per protocol or otherwise.

References

Angrist JD. 1990. "Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records." *American Economic Review*, 80(3): 313–336.

Angrist JD, Imbens GW. 1995. "Two-stage least squares estimation of average causal effects in models with variable treatment intensity." *Journal of the American Statistical Association*, 90(430): 431–442.

Angrist JD, Imbens GW, Rubin DB. 1996. "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association*, 91(434): 444–455.

Angrist JD, Pathak PA, Zárate RA. 2023. "Choice and consequence: Assessing mismatch at chicago exam schools." *Journal of Public Economics*, 223: 104892.

Angrist JD, Pischke JS. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Angrist N, Meager R. 2023. "Implementation Matters: Generalizing Treatment Effects in Education." *Available at SSRN 4487496*.

Atkin WS, et al. 2010. "Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial." *The Lancet*, 375(9726): 1624–1633.

Atkin W, et al. 2017. "Long term effects of once-only flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible Sigmoidoscopy Screening randomised controlled trial." *The Lancet*, 389(10076): 1299–1311.

Baker SG, Kramer BS, Lindeman KS. 2016. "Latent class instrumental variables: a clinical and biostatistical perspective." *Statistics in Medicine*, 35(1): 147–160.

Baker SG, Lindeman KS. 1994. "The paired availability design: a proposal for evaluating epidural analgesia during labor." *Statistics in Medicine*, 13(21): 2269–2278.

Bloom HS. 1984. "Accounting for no-shows in experimental evaluation designs." *Evaluation Review*, 8(2): 225–246.

Bretthauer M, Løberg M, Kaminski MF. 2023. "Colonoscopy Screening and Colorectal Cancer Incidence and Mortality." *New England Journal of Medicine*, 388(4): 376–376. Response to letters.

Bretthauer M, et al. 2022. "Effect of colonoscopy screening on risks of colorectal cancer and related death." *New England Journal of Medicine*, 387(17): 1547–1556.

Chêne G, et al. 1998. "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." *Controlled Clinical Trials*, 19(3): 233–248.

Cuzick J, Edwards R, Segnan N. 1997. "Adjusting for non-compliance and contamination in randomized clinical trials." *Statistics in Medicine*, 16(9): 1017–1029.

Dominitz JA, Robertson DJ. 2022. "Understanding the results of a randomized trial of screening

colonoscopy." *New England Journal of Medicine*, 387(17): 1609–1611.

El Sahly HM, et al. 2021. "Efficacy of the mRNA-1273 SARS-CoV-2 vaccine at completion of blinded phase." *New England Journal of Medicine*, 385(19): 1774–1785.

Glasziou P. 1992. "Meta-analysis adjusting for compliance: the example of screening for breast cancer." *Journal of Clinical Epidemiology*, 45(11): 1251–1256.

Gyawali B. 2022. "A controversial trial: exposing misunderstandings of NordICC." *Medscape*. https://www.medscape.com/viewarticle/982479.

Hayes JH, Barry MJ. 2014. "Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence." *JAMA*, 311(11): 1143–1149.

Hearst N, Newman TB, Hulley SB. 1986. "Delayed effects of the military draft on mortality." *New England Journal of Medicine*, 314(10): 620–624.

Hernán MA, Robins JM. 2017. "Per-protocol analyses of pragmatic trials." *New England Journal of Medicine*, 377(14): 1391–1398.

Holme et al. 2014. "Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: a randomized clinical trial." *JAMA*, 312(6): 606.

Holme et al. 2017. "Effectiveness of flexible sigmoidoscopy screening in men and women and different age groups: pooled analysis of randomised trials." *BMJ*, 356(i6673).

Holme et al. 2018. "Long-term effectiveness of sigmoidoscopy screening on colorectal cancer incidence and mortality in women and men: a randomized trial." *Annals of Internal Medicine*, 168(11): 775–782.

Imbens GW, Angrist JD. 1994. "Identification and estimation of local average treatment effects." *Econometrica*, 62(2): 467–475.

Imbens GW, Rubin DB. 1997. "Estimating outcome distributions for compliers in instrumental variables models." *The Review of Economic Studies*, 64(4): 555–574.

Imbens GW, Rubin DB. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge: Cambridge University Press.

Juul FE, et al. 2022. "15-Year benefits of sigmoidoscopy screening on colorectal cancer incidence and mortality: a pooled analysis of randomized trials." *Annals of Internal Medicine*, 175(11): 1525–1533.

Kowalski AE. 2021. "Mammograms and Mortality: How Has the Evidence Evolved?" *Journal of Economic Perspectives*, 35(2): 119–140.

Kowalski AE. 2023. "Behaviour within a clinical trial and implications for mammography guidelines." *The Review of Economic Studies*, 90(1): 432–462.

Lee Y, Kennedy EH, Mitra N. 2023. "Doubly robust nonparametric instrumental variable estimators for survival outcomes." *Biostatistics (Oxford, England)*, 24(2): 518–537.

Newcombe R. 1988. "Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur." *Statistics in Medicine*, 7(11): 1179–1186.

Newey WK. 1985. "Generalized method of moments specification testing." *Journal of Econometrics*, 29(3): 229–256.

Packer DL, et al. 2019. "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(13): 1261–1274.

Rembold CM. 1998. "Number needed to screen: development of a statistic for disease screening." *BMJ*, 317(7154): 307–312.

Roland M, Torgerson DJ. 1998. "Understanding controlled trials: What are pragmatic trials?" *BMJ*, 316(7127): 285–285.

Schoen RE, et al. 2012. "Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy." *New England Journal of Medicine*, 366(25): 2345–2357.

Segnan N, et al. 2011. "Once-only sigmoidoscopy in colorectal cancer screening: follow-up findings of the Italian randomized controlled trial–SCORE." *Journal of the National Cancer Institute*, 103(17): 1310–1322.

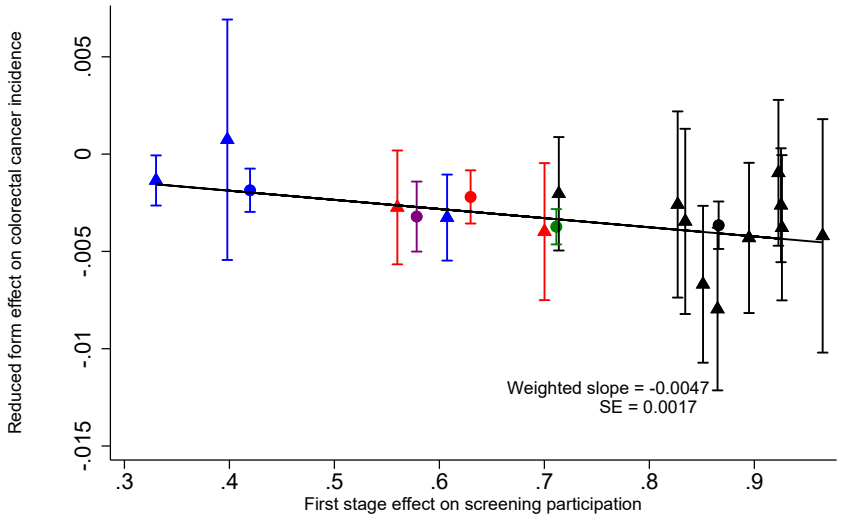
Senore C, et al. 2022. "Long-term follow-up of the Italian Flexible Sigmoidoscopy Screening Trial." *Annals of Internal Medicine*, 175(1): 36–45.

Swanson SA, et al. 2015. "Bounding the per-protocol effect in randomized trials: an application to colorectal cancer screening." *Trials*, 16: 541.

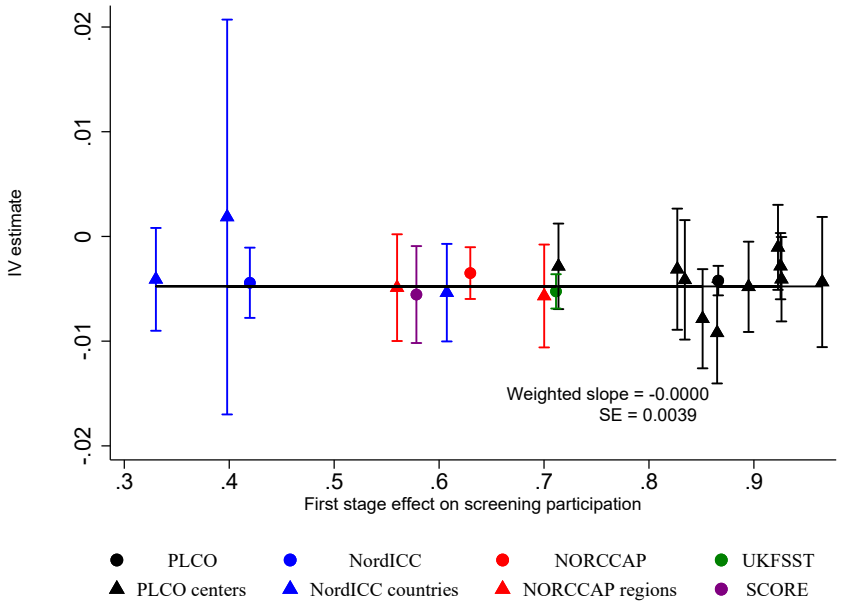
Winawer SJ. 2023. "Colonoscopy Screening and Colorectal Cancer Incidence and Mortality." *New England Journal of Medicine*, 388(4): 376–376. Letter.

Figure 1

Panel A. Visual Instrumental Variables

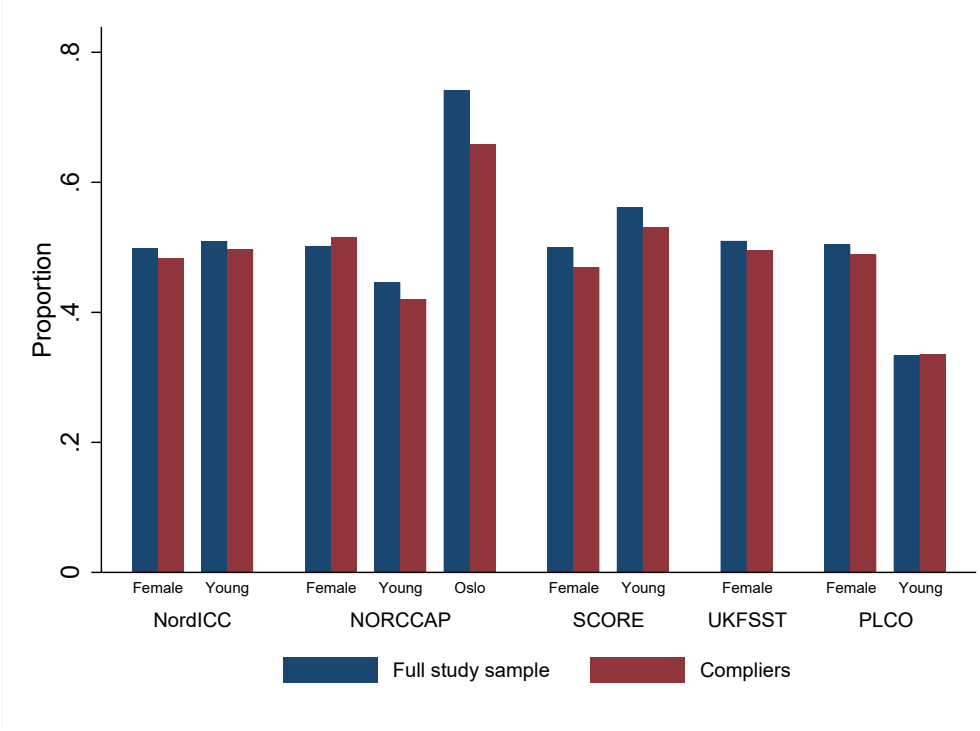


Panel B. IV Estimates Plotted Against First-stage Estimates



Notes: 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 10-12 years after random assignment. Regression lines plotted in the figures are weighted by $P_j p_j(1-p_j)$ where P_j is the sample size and p_j is the offer rate and computed omitting estimates for the full NordICC, NORCCAP, and PLCO samples to avoid duplication. The line in Panel A is fit without an intercept. The slope of this line is an estimate of λ, as described in the text. Whiskers mark 95% confidence intervals.

Figure 2: Complier Characteristics



Notes: 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 50-54 for NORCCAP and to 55-59 for NordICC, SCORE, and PLCO.

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	2009-2014	1999-2001	1995-1999	1994-1999	1993-2001
Initial Screening Type	Colonoscopy	Sigmoidoscopy + FOBT	Sigmoidoscopy	Sigmoidoscopy	Sigmoidoscopy
Follow-up Screening Type	Biopsy	Colonoscopy, Biopsy	Colonoscopy	Colonoscopy	Second sigmoidoscopy (after 3/5 year)
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	55-64	50-64	55-64	55-64	55-74
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 signoidoscopy or colonoscopy. Half of the NORCCAP treated group was invited for sigmoidoscopy, with the rest invited for both sigmoidoscopy and fecal occult blood test (FOBT). The PLCO trial offered a second screening 3 or 5 years after the initial screening. The 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 subjects with follow-up data. All participants who receive the first screening are invited to receive a second flexible sigmoidoscopy after 3 or 5 years. Follow-up screening invitation in NORCCAP, SCORE, and UKFSST are based on polyp detection in the initial screening. Second sigmoidoscopy screening invitations are to all invited group. NordICC does not provide follow-up screenings.

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-year colorectal cancer (CRC) incidence. The instrument is a randomly assigned invitation to undergo colonoscopy/sigmoidoscopy screening; treatment is participation in colonoscopy/sigmoidoscopy screening. The outcome variable indicates CRC diagnosis 10-12 years after random assignment. Column 1 reports the mean of CRC incidence diagnosis 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 screeening. Robust standard errors appear in parentheses.

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 Appendix A.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.

A Appendix

A.1 IV Estimates Contributing to Figure 1

Table A1: IV Estimates of Screening Effects Within Trials

	N	ITS	First Stage (Adherence)	IV/LATE
	(1)	(2)	(3)	(4)
		A. NordICC		
Poland	54,528	-0.0014	0.3301	-0.0041
		(0.0008)	(0.0035)	(0.0025)
Norway	26,411	-0.0033	0.6074	-0.0054
		(0.0014)	(0.0052)	(0.0023)
Sweden	3,646	0.0007	0.3980	0.0019
		(0.0038)	(0.0140)	(0.0095)
		B. NORCCAP		
Oslo	73,190	-0.0027	0.5600	-0.0049
		(0.0015)	(0.0049)	(0.0026)
Telemark	25,488	-0.0040	0.7002	-0.0057
		(0.0018)	(0.0045)	(0.0025)
		C. PLCO		
University of Colorado	13,164	-0.0067	0.8511	-0.0079
		(0.0021)	(0.0044)	(0.0024)
Georgetown University	8,105	-0.0035	0.8342	-0.0041
		(0.0024)	(0.0058)	(0.0029)
Pacific Health Research and Education Institute	10,842	-0.0026	0.8270	-0.0031
		(0.0024)	(0.0051)	(0.0030)
Henry Ford	24,662	-0.0020	0.7138	-0.0029
		(0.0015)	(0.0041)	(0.0021)
University of Minnesota	28,860	-0.0026	0.9254	-0.0028
		(0.0015)	(0.0022)	(0.0016)
Washington University in St Louis	15,041	-0.0080	0.8650	-0.0092
		(0.0021)	(0.0039)	(0.0025)
University of Pittsburgh	16,930	-0.0038	0.9263	-0.0041
		(0.0019)	(0.0028)	(0.0021)
University of Utah	14,363	-0.0010	0.9230	-0.0010
		(0.0019)	(0.0031)	(0.0021)
Marshfield	16,732	-0.0043	0.8949	-0.0048
		(0.0020)	(0.0034)	(0.0022)
University of Alabama at Birmingham	6,188	-0.0042	0.9651	-0.0044
		(0.0031)	(0.0033)	(0.0032)

Notes: This table reports IV estimates of the effect of colonoscopy screening on colorectal cancer (CRC) incidence for subgroups in NordICC, NORCCAP, and PLCO. The instrument is a randomly assigned screening offer. NORCCAP estimates in Panel B are based on a 15-year follow-up; other estimates are for 10-12 year follow-ups. Robust standard errors appear in parentheses.

A.2 Estimation Sources and Methods

This appendix gives sources and methods for the ITS, IV, and old-fashioned per-protocol estimates in Table 2 and Appendix Table A1.

NordICC

Estimates come from published results in Bretthauer et al. [2022]. Specifically, statistics used to compute reduced-form (ITS) and first-stage (adherence) estimates for the full trial are taken from Tables 1 and 2. IV estimates are computed from these as described in the text. ITS estimates reported in Bretthauer et al. [2022] appear to adjust for covariates; ours do not.

Our per-protocol estimate omitting never-takers is computed by dividing the number of CRC cases for screened and control groups reported in Table S4 by numbers of subjects in these groups as reported in Table 1 ($102/11843 - 622/56365 = -0.0024$). An as-treated estimate is computed by moving the invited but not screened participants to the control group ($102/11843 - (622 + (259 - 102))/(56365 + (28220 - 11843)) = -0.0021$). The number of CRC cases invited but not screened is obtained from Table S1 and S4, while the size of the group is obtained from Table 1.

Statistics used to compute country-specific adherence rates and ITS estimates are taken from Table 1 and Supplementary Table S5, respectively. Bretthauer et al. [2022] supplement Table S4 reports an "adjusted per-protocol estimate" of -0.0038; the supplement notes CRC risk gaps between invited non-screened subjects and controls. In contrast with estimates relying on covariate adjustment, IV estimates should be insensitive to control for (pre-treatment) covariates.

Standard errors are computed by applying conventional "robust" (i.e., heteroscedasticity-consistent) formulas for OLS and 2SLS. Our calculations exploit the fact that with dummy dependent variables, these formulas can be implemented using published data on sample sizes, screening rates, and CRC incidence since our outcomes, treatments, and instruments are dummies.

Specifically, robust standard errors for ITS estimates are given by:

$$\hat{\sigma}(\hat{\rho}_j) = \left[\frac{(\frac{1}{N}\sum_i (Z_i - \bar{Z})^2 (Y_i - \hat{Y}_i)^2)}{N\left(\frac{1}{N}\sum_i (Z_i - \bar{Z})^2\right)^2} \right]^{1/2}, \tag{11}$$

where N is trial sample size, $\bar{Z}$ is the sample mean of Z_i and $\hat{Y}_i$ is the fitted value from a regression of Y_i on Z_i. Since Z_i and Y_i are both dummies, the numerator of equation (11) is given by:

$$\begin{aligned}\frac{1}{N}\sum_i (Z_i - \bar{Z})^2 (Y_i - \hat{Y}_i)^2 =& p(Y=1, Z=0)(0-\bar{Z})^2(1-\bar{Y}^0)^2 + p(Y=0, Z=0)(0-\bar{Z})^2(0-\bar{Y}^0)^2 \\ &+ p(Y=1, Z=1)(1-\bar{Z})^2(1-\bar{Y}^1)^2 + p(Y=0, Z=1)(1-\bar{Z})^2(0-\bar{Y}^1)^2,\end{aligned}$$

where $p(Y = m, Z = n)$ denotes the sample proportion of observations with $Y_i = m$ and $Z_i = n$, and $\bar{Y}^n$ gives the sample average Y_i among observations where $Z_i = n$. The denominator of (11) is computed using:

$$\frac{1}{N}\left(\sum_i (Z_i - \bar{Z})^2\right) = p(Z=0)(0-\bar{Z})^2 + p(Z=1)(1-\bar{Z})^2,$$

where $p(Z = n)$ gives the sample proportion of observations where $Z_i = n$. Proportions are computed using $p(Y = m, Z = n)$ and $p(Z = n)$ numerator counts in Tables 1 and 2 in Bretthauer

et al. [2022]; the $\bar{Y}^n$ means are computed by dividing the number of CRC cases in a group by the size of the group. The mean offer rate, $\bar{Z}$, is computed by dividing the size of the group invited to screening by the overall sample size. Appropriately modified versions of these formulas yield robust standard errors for the old-fashioned per-protocol (OLS) estimates.

Robust standard errors for IV estimates are given by:

$$\hat{\sigma}(\hat{\lambda}_j) = \left[\frac{(\frac{1}{N}\sum_i (\hat{S}_i - \bar{S})^2 (Y_i - (\hat{\alpha}_{IV} + \hat{\beta}_{IV} S_i))^2)}{N\left(\frac{1}{N}\sum_i (\hat{S}_i - \bar{S})^2\right)^2} \right]^{1/2}, \tag{12}$$

where $\bar{S}$ is the sample mean of S_i, $\hat{S}_i$ gives the first-stage fitted value for observation i and $(\hat{\alpha}_{IV}, \hat{\beta}_{IV})$ are the IV (second-stage) intercept and slope estimate.[19] We compute $\hat{\beta}_{IV}$ by dividing the ITS estimate by the adherence rate. The intercept is $\hat{\alpha}_{IV} = \bar{Y} - \hat{\beta}_{IV}\bar{S}$, where $\bar{Y}$ is the sample mean of Y_i. Since Z_i, S_i, and Y_i are dummies, the numerator in equation (12) expands as:

$$\begin{aligned}
&\frac{1}{N}\sum_i (\hat{S}_i - \bar{S})^2 (Y_i - (\hat{\alpha}_{IV} + \hat{\beta}_{IV} S_i))^2 \\
=&p(Y=1,S=1,Z=1)(\bar{S}^1 - \bar{S})^2 (1 - (\hat{\alpha}_{IV} + \hat{\beta}_{IV}))^2 \\
&+ p(Y=0,S=1,Z=1)(\bar{S}^1 - \bar{S})^2 (0 - (\hat{\alpha}_{IV} + \hat{\beta}_{IV}))^2 \\
&+ p(Y=1,S=0,Z=1)(\bar{S}^1 - \bar{S})^2 (1 - \hat{\alpha}_{IV})^2 + p(Y=0,S=0,Z=1)(\bar{S}^1 - \bar{S})^2 (0 - \hat{\alpha}_{IV})^2 \\
&+ p(Y=1,S=0,Z=0)(\bar{S}^0 - \bar{S})^2 (1 - \hat{\alpha}_{IV})^2 + p(Y=0,S=0,Z=0)(\bar{S}^0 - \bar{S})^2 (0 - \hat{\alpha}_{IV})^2,
\end{aligned}$$

where $p(Y = m, S = q, Z = n)$ counts the sample proportion with $Y_i = m$, $S_i = q$, and $Z_i = n$ and where $\bar{S}^p$ gives the sample average S_i among observations where $Z_i = n$, with other sample proportions and sample means defined similarly. Here we use the fact that there are no always-takers, so $p(Y = 1, S = 1, Z = 0) = p(Y = 0, S = 1, Z = 0) = 0$. The denominator of equation (12) is computed using:

$$\frac{1}{N}\sum_i (\hat{S}_i - \bar{S})^2 = p(Z=1)(\bar{S}^1 - \bar{S})^2 + p(Z=0)(\bar{S}^0 - \bar{S})^2.$$

Note that the absence of always-takers implies $\bar{S}^1$ is the first-stage estimate and $\bar{S}^0 = 0$. As above, counts and averages are obtained from Tables 1 and 2 in Bretthauer et al. [2022].

NORCCAP

Reduced-form (ITS) and first-stage (adherence) estimates for the full trial are taken from Supplementary eTable2 in Holme et al. [2014]. A per-protocol estimate omitting never-takers is computed as the difference in CRC incidence in the adherent group and control group using numbers reported in Supplementary eTable2 in Holme et al. [2014], $(116/12955 - 890/78220)$. An as-treated estimate

[19] Note that $\bar{S}$ is also the sample mean of $\hat{S}_i$, since the fitted values are given by OLS.

is computed using the difference in CRC incidence in the adherent group and in the control and nonadherent grousp $(116/12955 - (890+92)/(78220+7617))$. We received region-specific adherence rates and ITS estimates from the NORCCAP study team. IV estimates and all standard errors are computed as for NordICC.

SCORE

Reduced-form (ITS) and first-stage (adherence) estimates for the full trial are computed using numbers from Tables 1 and 2 in Segnan et al. [2011]. As-treated and per-protocol estimates omitting never-takers are computed as for NORCCAP using numbers from Table 2 in Segnan et al. [2011]. IV estimates and all standard errors are computed as for NordICC.

UKFSST

Reduced-form (ITS) and first-stage (adherence) estimates for the full trial are computed using numbers from Table 1 and Figure 1 in Atkin et al. [2010]. As-treated and per-protocol estimates omitting never-takers are computed as for NORCCAP using numbers from Table 2 in Atkin et al. [2017]. IV estimates and all standard errors are computed as for NordICC.

PLCO

PLCO estimates and standard errors were computed using microdata obtained from the Cancer Data Access System (CDAS). The CDAS data set records trials subjects assignment, receipt of initial screening among those invited to screen, receipt of a later screening (in 3 years or in 5 years), the PLCO study center recruiting subjects, and a colorectal cancer (CRC) incidence diagnosis through 2009. Screening is coded according to whether treated subjects received any screening defined. The CRC outcome indicates whether participants were diagnosed with CRC any time from random assignment through 2009.

A.3 Visual IV and 2SLS

The no-constant best-fit line in Figure 1 weights trial sites by trial sample size times the within-trial variance of Z_i. We show here that the slope of this line is the two-stage least squares (2SLS) estimate of the common screening effect, λ.

2SLS gets its name from the fact that it's computed in two regression steps. The 2SLS estimator corresponding to Figure 1 can be described by a causal "second-stage" model of the form:

$$Y_i = \sum_j \beta_j D_{ij} + \lambda S_i + \varepsilon_i,$$

with corresponding first-stage equation:

$$S_i = \sum_j \alpha_j D_{ij} + \sum_j \pi_j D_{ij} Z_i + \nu_i.$$

In the context of multiple trials, the 2SLS estimation sample stacks data from J sites, while the D_{ij} are J dummies indicating participants in site j. 2SLS proceeds by estimating the first-stage equation, then replacing S_i in the causal model with first-stage fitted values. A key feature of 2SLS as implemented here is that the first-stage and reduced-form models are saturated, meaning there's a dummy for every possible combination of D_{ij} and Z_i). A regression therefore fits the corresponding sample conditional means perfectly.

2SLS and VIV equivalence follows from Theorem 3 in Angrist and Imbens [1995], which shows that the 2SLS estimator in models with a saturated first stage converges to a weighted average of one-at-a-time IV estimands:

$$\lambda_{2SLS} = \sum_j \left(\frac{P_j V_j \pi_j^2}{\sum_k P_k V_k \pi_k^2} \right) \frac{\rho_j}{\pi_j},$$

where $P_j = Pr(D_{ij} = 1)$ and $V_j = Var(Z_i \mid D_{ij} = 1)$. This formula can be rewritten:

$$\lambda_{2SLS} = \frac{\sum_j W_j \pi_j \rho_j}{\sum_j W_j \pi_j^2},$$

where $W_j = P_j V_j$. Thus, the 2SLS estimand is a weighted regression of ρ_j on π_j with no intercept and weights W_j. The 2SLS estimator is the sample analog of this formula.

A.4 Over-identification Test Statistics

The test statistics reported in Table 3 are computed using equation (9), where $\hat{\sigma}_j^2 = \hat{\sigma}(\hat{\lambda}_j)^2 \hat{\pi}_j^2$ and $\hat{\sigma}(\hat{\lambda}_j)$ is the robust IV standard error from equation (12). This variance estimate is derived from the asymptotic distribution of the IV estimates under the proportionality null hypothesis of $\rho_j/\pi_j = \lambda$:

$$\sqrt{N}(\hat{\lambda}_j - \lambda) \Rightarrow \mathcal{N}(0, N\sigma(\hat{\lambda}_j)^2),$$

where $N\sigma(\hat{\lambda}_j)$ is the probability limit of $N\hat{\sigma}(\hat{\lambda}_j)$. Note that $\hat{\pi}_j$ converges in probability to a constant. Application of Slutsky's theorem therefore gives:

$$\sqrt{N}(\hat{\rho}_j - \lambda\hat{\pi}_j) = \sqrt{N}(\hat{\lambda}_j - \lambda)\hat{\pi}_j \Rightarrow \mathcal{N}(0, N\sigma(\hat{\lambda}_j)^2\pi_j^2),$$

such that $\hat{\sigma}(\hat{\lambda}_j)^2\hat{\pi}_j^2$ estimates the asymptotic sampling variance of $\hat{\rho}_j - \lambda\hat{\pi}_j$.

An infeasible test statistic defined as

$$\tilde{T} = \sum\nolimits_j (1/\hat{\sigma}_j^2)(\hat{\rho}_j - \lambda\hat{\pi}_j)^2$$

is asymptotically a sum of J squared standard normals, and so is asymptotically distributed $\chi^2(J)$ under the null. Replacing λ with $\hat{\lambda}_{VIV}$ to obtain $\hat{T}$ changes the null distribution to $\chi^2(J-1)$; see Newey [1985] for details.

A.5 Complier Characteristics With Stratification

For the NORCCAP trial, which is stratified by region with different assignment rates in each, complier means are estimated separately in each region and then averaged using the share of compliers in each region. To see how this works, let categorical variable $W_i \in \{1, \ldots, K\}$ encode strata membership for subject i. Iterating expectations over strata yields:

$$Pr(X_i = 1 \mid C_i = 1) = \sum_k Pr(X_i = 1 \mid C_i = 1, W_i = k) Pr(W_i = k \mid C_i = 1). \tag{13}$$

The term $Pr(X_i = 1 \mid C_i = 1, W_i = k)$ can obtained by applying (10) within strata. The second term inside the sum, $Pr(W_i = k \mid C_i = 1)$, can be obtained using:

$$Pr(W_i = k \mid C_i = 1) = Pr(C_i = 1 \mid W_i = k) \frac{Pr(W_i = k)}{Pr(C_i = 1)},$$

where $Pr(C_i = 1 \mid W_i = k)$ is the strata-specific compliance (adherence) rate (equal to the first stage within strata). The overall compliance rate, $Pr(C_i = 1)$, is computed by averaging these strata-specific rates, weighting by strata size.

Even in trials with no always-takers (i.e. control-group crossovers), stratification may cause complier means to diverge from treated means. The latter are given by:

$$Pr(X_i = 1 \mid S_i = 1) = \sum_k Pr(X_i = 1 \mid C_i = 1, W_i = k) Pr(W_i = k \mid S_i = 1), \tag{14}$$

using $Pr(X_i = 1 \mid S_i = 1, W_i = k) = Pr(X_i = 1 \mid C_i = 1, W_i = k)$ when $S_{0i} = 0$. Moreover,

$$\begin{aligned} Pr(W_i = k \mid S_i = 1) &= \frac{Pr(C_i = 1 \mid W_i = k) Pr(Z_i = 1 \mid W_i = k) Pr(W_i = k)}{Pr(C_i = 1) Pr(Z_i = 1)} \\ &= Pr(W_i = k \mid C_i = 1) \frac{Pr(Z_i = 1 \mid W_i = k)}{Pr(Z_i = 1)}. \end{aligned}$$

The weights applied to $Pr(X_i = 1 \mid C_i = 1, W_i = k)$ in (13) and (14) differ unless assignment rates are constant within strata, in which case $Pr(Z_i = 1 \mid W_i = k) = Pr(Z_i = 1)$ for all k.