WP 22/18

How does a local Instrumental Variable Method perform across settings with instruments of differing strengths? A simulation study and an evaluation of emergency surgery

Silvia Moler-Zapata; Richard Grieve; Anirban Basu and Stephen O'Neill

July 2022

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

How does a local Instrumental Variable Method perform across settings with instruments of differing strengths? A simulation study and an evaluation of emergency surgery.

Authors: Silvia Moler-Zapata[1], Richard Grieve[1], Anirban Basu[2,3] and Stephen O'Neill[1].

Affiliations:

[1]Department of Health Services Research and Policy, London School of Hygiene & Tropical Medicine, London, UK
[2]The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, Department of Pharmacy, and Departments of Health Services and Economics, University of Washington, Seattle, USA.
[3]National Bureau of Economic Research, Cambridge, MA, USA.

Correspondence:

Silvia Moler-Zapata (address: Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK; Email: silvia.moler@lshtm.ac.uk; telf.: +44 (0) 207 927 2366).

Abstract:

Local instrumental variable (LIV) approaches use continuous/multi-valued instrumental variables (IV) to generate consistent estimates of average treatment effects (ATEs) and Conditional Average Treatment Effects (CATEs). However, there is little evidence on how LIV approaches perform with different sample sizes or according to the strength of the IV (as measured by the first-stage F-statistic). We examined the performance of an LIV approach and a two-stage least squares (2SLS) approach in settings with different sample sizes and IV strengths, and considered the implications for practice.

Our simulation study considered three sample sizes (n = 5000, 10000, 50000), six levels of IV strength (F-statistic = 10, 25, 50, 100, 500, 1000) under four 'heterogeneity' scenarios: effect homogeneity, overt heterogeneity (over measured covariates), essential heterogeneity (over unmeasured covariates), and overt and essential heterogeneity combined. Compared to 2SLS, the LIV approach provided estimates for ATE and CATE with lower levels of bias and RMSE, irrespective of the sample size or IV strength. With smaller sample sizes, both approaches required IVs with greater strength to ensure low (<5%) levels of bias. In the presence of overt and/or essential heterogeneity, the LIV approach reported estimates with low bias even when the sample size was smaller (n = 5000), provided that the instrument was moderately strong (F-statistic greater than 50, for the ATE estimand).

We considered both methods in evaluating emergency surgery across three different acute conditions with IVs of differing strengths (F-statistic ranging from 100 to 9000), and sample sizes (100000 to 300000). We found that 2SLS did not detect significant differences in effectiveness across subgroups, even with subgroup by treatment interactions included in the model. The LIV approach found there were substantive differences in the effectiveness of emergency surgery according to subgroups; for each of the three acute conditions, frail patients had worse outcomes following emergency surgery.

These findings indicate that when a continuous IV of a moderate strength is available, LIV approaches are better suited than 2SLS to estimate policy-relevant treatment effect parameters.

Keywords: Instrumental Variables, Instrument Strength, Tendency to Operate, Emergency Surgery.

1. Introduction

The personalisation of treatment choice can be informed by comparative effectiveness research that exploits the widespread availability of electronic health records (EHRs), but requires methods that address confounding and heterogeneity. For conventional linear Instrumental Variable (IV) methods, such as two-stage least squares (2SLS) to identify policy-relevant estimands such as the Average Treatment Effect (ATE) or Conditional Average Treatment Effects (CATEs), it is required that there is no essential heterogeneity. [1] Essential heterogeneity arises when treatment effects differ over levels of unmeasured confounders, in which case 2SLS no longer identifies the ATE, even if the instrument is strong and valid. [1] Essential heterogeneity, is a major concern in health care, as it is commonly the case that there are biological correlations between risk factors, some of which remain unobserved to the analyst.

In the presence of essential heterogeneity, Local Instrumental Variable (LIV) approaches can provide consistent estimates of the ATE and CATEs. [2] LIV methods draw on theory about individual's choices to identify 'marginal treatment effects' (MTEs) for individuals at the 'margin of treatment choice'.[3,4] These MTEs are identified for individuals for whom the level of the IV is such that observed characteristics encouraging treatment (including the IV) and unobserved characteristics discouraging treatment are balanced, so there is equipoise about the treatment decision. Here, a small change (or nudge) in the level of a valid, continuous IV 'tips the balance' for the treatment decision for these marginal patients, without changing the distribution of the underlying risk factors. Therefore, comparing mean outcomes between two groups of patients only separated by a small change in the IV, identifies MTEs for individuals who *comply* with the change in treatment, due to that small change in the IV. A continuous instrument with sufficient support allows all individuals to be defined as 'compliers' at some level of the IV. [4] Hence, given observed covariates, MTEs can be estimated along the continuum of the IV, and aggregated to provide CATEs and ATEs. [2,4,5]

The theoretical properties of these LIV methods in settings with essential heterogeneity have been discussed by Heckman et al. [1], Basu et al. [6] and Angrist et al. *inter alia.* [7] However, most simulation studies of IV methods only consider treatment effects that are homogeneous, or heterogenous according to measured factors (overt heterogeneity) [8–10]. Studies that have considered essential heterogeneity, have found that 2SLS provides inconsistent estimates of the ATE [11–13], whereas Basu [14] reports that a LIV method could provide consistent estimates of the ATE and CATE in finite samples. LIV methods have now been applied across a multitude of settings including cardiovascular and bariatric surgery, universal child care programs and transfers to intensive care units. [15–18]

A major barrier to wider use of IV approaches in general is that if the instrument is only weakly associated with treatment assignment, then IV estimators can provide very biased and imprecise estimates. [19–21] Weak IVs can also amplify the bias arising due to violations of the other assumptions. [20,22] While current practice tends to rely on the first-stage F-statistic exceeding the value of 10,[23] recent developments in the weak identification literature for IV models have revealed the shortcomings of an unequivocal decision rule for assessing weak identification.[24–28] For LIV to provide consistent, precise estimates of ATE or CATEs, requires a strong continuous/multi-valued IV with sufficient support to ensure that there is a level of the IV at which each unit 'complies' (i.e., is selected into treatment according to the level of the IV). However, no study has assessed the levels of IV strength that are required for an LIV estimator to perform well, nor how performance may differ according to the sample size available, in settings with essential heterogeneity.

This paper addresses this gap in the literature by contrasting LIV with the commonly used 2SLS estimator in Monte Carlo simulations, motivated by a case study which highlights typical issues pertaining to heterogeneity, sample size and IV strength. We simulate four scenarios: two of them under restrictive assumptions about heterogeneity (A: homogeneity; B: overt heterogeneity), one where treatment effects are allowed to be heterogenous according to an unmeasured confounder (C: essential heterogeneity), and one where both forms of heterogeneity are present (D: overt and essential heterogeneity). Across all scenarios, ATE and CATE are the parameters of interest.

This paper is structured as follows. In Section 2, we outline the motivating example. In Section 3, we define the estimands and identification assumptions for 2SLS and LIV, and present the methods for the simulation study. In Section 4, we present the results of the simulation study and the case study. In section 5, we discuss how this study adds to the literature and the implications for further research.

2. Motivating example: the ESORT study

The ESORT (Emergency Surgery OR noT) study evaluated the effectiveness of emergency surgery for acute gastrointestinal conditions. The primary outcome of the study was the number of 'days alive and out of hospital' (DAOH) at 90-days (see [29] for details), which encompasses mortality and total length of hospital stay (LOS). The study exemplifies the key issues that arise when applying IV methods to EHR data to provide policy-relevant estimates of comparative effectiveness. [29–31] Patients presented as emergency admissions and were selected for either emergency surgery (ES), or alternative interventions such as medical management or delayed surgery, according to unmeasured characteristics such as the severity of the disease, and hence unmeasured confounding and essential heterogeneity were major concerns.

The ESORT study followed Keele et al., [32] and developed a continuous preference-based IV for ES receipt to evaluate the effectiveness of ES for three acute gastrointestinal conditions: acute appendicitis, gallstone disease and abdominal wall hernia, using routine hospitalisation data from the hospital episode statistics (HES) inpatient database in England. The IV was the hospital's tendency to operate (TTO), a proxy measure of the hospital's latent preference for ES, defined as the proportion of eligible emergency admissions in each of 174 hospitals who had ES in the year preceding each admission. Given a relevant IV, two main assumptions need to hold: (i) conditional on the variables included in the models, the hospital's TTO was not correlated with the patient's outcome except through treatment assignment, (ii) it does not increase the probability of treatment for an individual at some value of the IV, but decrease it for higher values. The study design had some important features to support this assumption. First, in this emergency setting patients were unlikely to select the hospital according to quality of care. Second, the study only included direct admissions to hospital, so there was no scope to transfer the patient according to physician or patient choice. Third, information was collated on a rich set of proxies for the hospital's quality of acute care, including rates of mortality and emergency admissions in previous years, which were included in the models as fixed effects. Fourth, observed covariates, were balanced across all levels of the TTO, which helped support the requisite assumption that the IV also balanced unmeasured confounders.[29,33] The requisite assumption that the IV has a monotonic effect on treatment receipt could not be formally tested on the data. However, it was deemed plausible in this setting, as it seems unlikely that there are patients who would receive emergency surgery when admitted to hospitals with low TTO but receive NES when admitted into a hospital with high TTO.

The ESORT study highlighted several outstanding concerns pertaining to IV methods in general, and LIV approach in particular. While the study reported estimates of the ATE, from the outset, there was policy interest in estimating the CATEs, according to baseline covariates including age, number of comorbidities, and levels of frailty. While the sample sizes for each condition, were relatively large, they also differed across conditions, from 268,144 (appendicitis) and 240,977 (gallstone disease), to 106,432 (hernia) patients. There were also differences in the strength of the IV with F-statistics ranging from 141 (acute appendicitis), 739 (hernia) to 9,053 (gallstone disease). Hence, the ESORT study further motivated the interest in what strength of continuous IV was required to provide unbiased, efficient estimates of policy relevant estimands such as CATEs in settings with essential heterogeneity, and according to different sample sizes.

3. Methods

3.1. Instrumental variables methods

Throughout we use the Neyman-Rubin potential outcomes framework.[34,35] Let Y_D denote the observed outcome, D_Z denote the treatment received, and Z denote the instrumental variable, such that we observe (Y_D, D_Z, Z) for each individual. For each patient, let $Y_1 = \mu_1(X_O, X_U, \vartheta)$ and $Y_0 = \mu_0(X_O, X_U, \vartheta)$ denote the potential outcomes, where X_O is the vector of observed covariates, X_U is a vector of unmeasured confounders, and ϑ captures all the remaining unobserved random variables. Throughout, we assume exogeneity of the covariates (A1), so that the treatment assignment is the only source of endogeneity, such that $(X_O, X_U) \perp \vartheta$ and $X_O \perp X_U$.

3.1.1. Identification assumptions

Angrist et al.,[36] defined a series of structural assumptions for the identification of the LATE. Here, following Abadie [37] and Tan [38] we make the following assumptions which are the conditional version of the assumptions outlined by Angrist et al.[36]:

(A2)	Unconfoundedness of Z	$(Y_{d_z}, D_z) \perp Z \| X_O$
(A3)	Exclusion restriction	$Y_{d_z} = Y_d$ with probability 1
(A4)	Relevance	$0 < P(Z = z) < 1$, and
		$P(D = 1 \| X_O, Z) > P(D = 1 \| X_O)$
(A5)	Monotonicity	If $z' > z$ then $D_{z'} \geq D_z$ with probability 1
(A6)	Stable Unit Treatment Value Assumption	$D = D_Z$ and $Y = Y_D$

Assumption (A2) requires that Z is as good as randomly assigned within levels of X_O. Assumption (A3) rules out the possibility that Z has a direct effect on the outcome other than through D_z. Assumptions (A2) and (A3) ensure that the only effect of the Z on the outcome is through D_z. This is sometimes called the independence assumption. Assumption (A4) ensures that Z and D_z are correlated conditional on X_O. Assumption (A5) requires that an increase in Z always results in a higher or equal level of treatment assignment. Assumption (A6) requires that one individual's potential outcomes (Y_D) and treatments (D_z) are not influenced by other individuals' levels of Z (i.e., no interference),nor by how the instrument or treatment is delivered (i.e., no different versions of Z or D_z).

3.1.2. Estimands

Imbens and Angrist[39] and Angrist et al. [36] show that, under the assumptions outlined above, the LATE can be defined as $\Delta^{LATE}(x_o, z, z') = E[Y_1 - Y_0 | X_O = x_o, D_z < D_{z'}]$ and is identified by the IV estimand: [36,39]

$$\frac{E[Y|X_O = x_o,\ Z = z'] - E[Y|X_O = x_o,\ Z = z]}{E[D|X_O = x_o,\ Z = z'] - E[D|X_O = x_o,\ Z = z]}$$

Vytlacil [40] and Tan [38] showed that the independence (A2 and A3) and monotonicity assumptions (A5) of the LATE framework are equivalent to those imposed by a non-parametric selection model, where treatment assignment depends on whether a latent index ($\mu_D(X_O, Z)$) crosses a particular threshold (X_{U_D})[38,40]:

$$D_z = 1\{\mu_D(X_O, Z) \geq X_{U_D}\}$$

where X_{U_D} is a random variable that captures X_U and all other factors influencing treatment assignment but not the outcomes. As in Heckman and Vytlacil, [4,5] we can rewrite this equation as D_z=1$\{P(X_O, Z) > V\}$, where $V = F_{X_{U_D}}[X_{U_D}|X_O = x_O, Z = z]$ with $V \perp (Z, X_O)$ and $P(x_O, z) = F_{X_{U_D}|x_O,z}[\mu_D(X_O, Z)]$ is the propensity for treatment, and F represents a cumulative distribution function. Therefore, for any arbitrary distribution of X_{U_D} conditional on X_O and Z, by definition $V \sim Uniform[0,1]$ conditional on X_O and Z. Then, the MTE can be defined as, $\Delta^{MTE}(x_O, p) = E(Y_1 - Y_0|X_O = x_O, V = v)$ and Heckman and Vytlacil [4,5] showed that, under the standard IV assumptions, it can be identified by:

$$\frac{\partial E_\vartheta(Y|X_O = x_o, Z = z)}{\partial p} = E_\vartheta[(Y_1 - Y_0)|X_O = x_o, V = v]$$

MTEs can be aggregated directly to obtain estimates of the ATE as shown in [1]. Basu [14] showed that MTEs can be used to derive personalised treatment (PeT) effects for each individual that take into account the plausible range of values that V may take for each patient, in addition to their observed covariates, IV and actual treatment assignment (see Section 3.1.3). [14] The rationale for this approach is that the treatment assignment status provides some information on X_{U_D}. For patients in the treatment group ($D_z = 1$), the propensity to choose treatment based on X_O and Z must outweigh the propensity to choose the comparator strategy based on X_{U_D}, i.e. $P(x_O, z) > v$. For patients in the comparator strategy ($D_z = 0$), the opposite is true. The PeT effect for an individual is obtained by averaging the MTEs corresponding to that individual's level of X_O and Z over those values of unobserved variables that are compatible with that patient's treatment assignment. Hence, $\Delta^{PeT}(x_O, p, D) = E(Y_1 - Y_0|X_O = x_O, P(z, x_O) > v)$ for individuals with $D_z = 1$ and $\Delta^{PeT}(x_O, p, D) = E(Y_1 - Y_0|X_O = x_O, P(z, x_O) < v)$ for individuals with $D_z = 0$.

All of the treatment effect estimands, including ATE and CATEs, can be derived by appropriately aggregating the PeT effects since these are defined at the individual level (see section 3.1.3.)

3.1.3. Estimation methods

Two-stage Least Squares estimator

2SLS is a common approach to the implementation of IV methods that consistently estimates the ATE parameter under homogeneity, or the LATE parameter under essential heterogeneity given a binary IV. Under assumptions (A1)-(A6), the 2SLS (Wald) estimator involves: (i) estimating $E[D_Z|X_O, Z]$ by regressing D_z on X_O and Z, and (ii) estimating $E[Y_D|D_z, X_O, Z]$ by regressing on X_O and $\hat{E}[D_Z|X_O, Z]$. When the instrument is continuous, 2SLS reports a weighted average of LATEs, which requires careful interpretation.[41]

Local Instrumental Variables estimator: estimating PeT effects

Basu [14,42] describe in detail the series of steps required to estimate PeT effects using the LIV methodology. Briefly, D_z is regressed on Z and X_O, as above, using appropriate methods for binary outcomes and the propensity for treatment $p(x_O, z)$ is estimated. Next, Y is regressed on X_O and a function of $\hat{p}(x_O, z)$ including interactions with X_O. The approach outlined in Basu[14] involves differentiating the outcome model $g(Y)$ by $\hat{p}(x_O, z)$. Next, PeT effects for each individual can be obtained by performing numerical integration, with MTE ($\partial\hat{g}(Y)/\partial\hat{p}$) evaluated by replacing $\hat{p}$ using 1,000 random draws of $u{\sim}unif(\min(\hat{p}(x_O, z)), \max(\hat{p}(x_O, z)))$. Then, $D^* = \Phi^{-1}\{\hat{p}(x_O, z)\} + \Phi^{-1}(1-u)$ can be computed. PeT effects can be computed by averaging $\partial\hat{g}(Y)/\partial\hat{p}$ over values of u for which $D^* > 0$ if $D = 1$; or over values of $D^* \leq 0$ if $D = 0$. Finally, averaging PeT effects over all of the observations provides an estimate of the ATE for the population, and over strata of X_O gives the CATE for the subpopulation of interest. Standard errors can be computed using bootstrap methods.[42] We now consider the design of the simulation study to contrast the relative performance of the LIV and 2SLS approaches.

3.2. Simulation study

Motivated by the gaps in the extant literature, and the motivating example, this simulation study was designed to consider the relative performance of 2SLS and LIV approaches across settings that differed with respect to the form of heterogeneity, the sample size and the strength of the IV. We report the performance of the methods in a Monte Carlo Simulation study according to their mean bias (%) and Root Mean Squared Error (RMSE) for each estimand (ATE and CATE).

3.2.1. Data Generating process

We create 5,000 datasets each containing N= $\{5000,\ 10000,\ 50000\}$ units, of which 50% are assigned to the treated group. The data generating process (DGP) includes one observed (X_O) and one unmeasured (X_U) covariate. We draw X_O, X_U and the instrument, Z from normal distributions with mean 0, and standard deviation 3. Three subgroups of interest are defined by whether the individuals' values for X_O are more than 0.5 standard deviations below or above its mean.

Treatment model

The treatment assignment is determined by the latent variable D^*, defined as:

$$D^* = \delta_D + 3X_O - 3X_U + \delta_Z Z + (4 - \delta_Z)\epsilon_D$$

where ϵ_D has a normal distribution with mean 0 and standard deviation, 1. Treatment is then determined as $D = 1$ if $D^* > 0$ and $D = 0$ otherwise. The parameters δ_Z and δ_D are chosen to ensure the IV F-statistic, F_{IV}, equals the desired level $F_{Target} = \{10, 25, 50, 100, 500, 1000\}$ on average, with,

$$F_{IV} = \left(N - df_m - 1\right) * \frac{\sigma^2_{no\ IV} - \sigma^2_{IV}}{\sigma^2_{IV}}$$

where $\sigma^2_{no\ IV}$ and σ^2_{IV} indicate the residual variance from regressing D on X_O with or without including the IV respectively, and df_m is the number of parameters in the model excluding the IV (i.e. $df_m = 2$ here). For a given F-statistic, a larger sample size implies a lower compliance rate, which in turn will imply a weaker instrument. At low compliance rates, the MSE of IV estimates can increase substantially.[43] We estimate the compliance rate for each sample size and F-statistic, by contrasting treatment uptake at the 1st and 99th percentiles of the IV.

Outcome model

The outcome models under treatments (Y_1) and control (Y_0) can be written as:

$$Y_0 = \beta_0 + \beta_1 X_O + \beta_2 X_U + \epsilon_{Y_0}$$
$$Y_1 = (\beta_0 + \tau_0) + (\beta_1 + \tau_1)X_O + (\beta_2 + \tau_2)X_U + \epsilon_{Y_1}$$

Implying the treatment effect is $\tau = E(Y_1 - Y_0) = \tau_0 + \tau_1 X_O + \tau_2 X_U$. Specifically we define the outcome under control as follows:

$$Y_0 = -10 - 10X_O + 10X_U + N(0,1)$$

We consider 4 scenarios for the outcome under treatment, Y_1. In Scenario A, effects are homogeneous ($\tau = 50$). In Scenario B, effects are heterogeneous but depend only on observed confounders (overt heterogeneity) ($\tau = 40 + 20X_O$). In Scenario C, X_U influences both the treatment assignment and the gains from treatment ($\tau = 40 + 20X_U$). In this Scenario, there is essential heterogeneity but no overt effect heterogeneity. Finally in Scenario D there is both overt and essential heterogeneity ($\tau = 20 + 20X_O + 20X_U$). Table 1 displays the parameter values for each scenario. The parameter combinations of interest consist of combinations of $n = \{5000, 10000, 50000\}$ and $F_{Target.}$ =

$\{10, 25, 50, 100, 500, 1000\}$. For each parameter combination for each scenario, we create 5000 datasets using the DGP described above and estimate the treatment effects as described below.

3.2.2. Implementation of methods

For the 2SLS model, we control for X_O and instrument D by Z. To capture heterogeneity, we also include an interaction between X_O with D, and instrument this with interactions of Z and X_O. To obtain effect estimates, we use the recycled predictions approach, whereby the two potential outcomes (Y_0 and Y_1) are predicted from the second stage model after setting $D = 0$ or $D = 1$ and the interaction X_O*$D = 0$ or X_O.[44,45] The individual level effect is then estimated as $\hat{\tau} = \widehat{\mathrm{Y}}_1 - \widehat{\mathrm{Y}}_0$, allowing us to calculate the ATE, and CATEs for the three subgroups (CATE$_{1,}$ CATE$_{2,}$ and CATE$_3$).

For the LIV approach, we first estimate the propensity for treatment conditional on $X_O$and Z, and in the second stage outcome model we include X_O, D, the estimated propensity score ($\hat{p}$), $\hat{p}$*X_O and $\hat{p}$ ^{2}. We then estimate PeT effects for each individual as described in Basu [42] using the `petiv` command in Stata. The estimated PeT effects are then aggregated to obtain estimates of the ATE, CATE$_{1,}$ CATE$_{2,}$ and CATE$_{3.}$ Before applying either method, we remove observations at those levels of the estimated propensity score where there is insufficient overlap. [42]

4. Results

4.1. Simulation study

Figures 1-4 present mean (%) bias in the ATE and CATE estimates (Figures 1 and 2, respectively) and the corresponding plots for RMSE (Figures 3 and 4, respectively). The results for the three subgroups showed similar patterns, and hence, for brevity, we only report the results for one of them.

In settings with homogenous treatment effects, or with overt heterogeneity, both approaches reported relatively low levels of bias (<5%) in the ATE estimates, apart from 2SLS, which reported moderate levels of bias (5-10%) in settings with F-statistics below 100 or a smaller sample size (n = 5000) (Figure 1). In settings with essential heterogeneity, 2SLS reports relatively high (>10%) levels of mean bias across practically all combinations of IV strength and sample size. The mean (%) bias is quite variable with respect to the target F-statistic (Figure 1). Inspection of the distribution of percentage bias across the 5,000 simulations (not shown) suggests this is due to the fact that the tails of the distribution are fat, particularly at lower values of F. At very high (>100) levels of the target F, the mean and median % bias are similar however this is not the case at lower levels. LIV estimator reports low levels of bias in ATE estimates across all scenarios aside from those with both a smaller sample size (n = 5000) and a F-statistic of 10 or 25 (Figure 1). The distribution of bias across simulation runs (not shown) has thinner tails for the LIV method than seen for 2SLS, hence the mean bias is less volatile here.

The bias plots for the CATE estimates have a somewhat similar pattern, although for this estimand the 2SLS estimator reports high levels of mean bias even in settings with overt heterogeneity, unless the sample size is relatively large (n = 50000) and/or the F-statistic is above 100 (Figure 2). The LIV estimator reports lower levels of bias than 2SLS across the majority of scenarios.

In general for both methods, across most scenarios, for a given sample size, the levels of mean (%) bias decrease at higher levels of the F-statistic (Figure 2). The RMSE in the estimates of the ATE are substantially lower for the LIV than the 2SLS estimator, except for those settings with an F-statistic of 500 or 1000 (Figure 3). For the CATE, in general, the RMSE estimates mirror the bias results, in that they are substantially lower across all settings for LIV (Figure 4).

Compliance rates for a given F-statistic were sensitive to the sample size available. For a sample size of 5000, increasing the F-statistic from 10 to 1000 increases the compliance rate from 8% to 73%, while for a sample size of 50000, the compliance rate only increases from 3% to 29% (Table 2).

4.2. Case study

4.2.1. Case study: implementation of 2SLS and LIV approaches

LIV estimated PeT effects of ES versus NES on DAOH at 90 days, for each individual allowing for treatment effect heterogeneity and confounding.[27-30] These PeT effects were aggregated to report the effects of ES overall, and for each pre-specified subgroup of interest. Since DAOH at 90 days was left skewed due to the maximum being 90 days, we rescaled this to lie between 0 and 1 (90-DAOH)/90) and effects were then rescaled back to the original scale. Probit regression models were used to estimate the initial propensity score (first stage), while GLMs were applied to the outcome data, with the most appropriate family and link function chosen according to RMSE, with Hosmer-Lemeshow and Pregibon tests also used to check model fit and appropriateness. [46,47] The logit link and binomial family were selected for all three conditions. Models at both stages adjusted for baseline measures, time period, and proxies for hospital quality, defined by rates of emergency readmission and mortality in 2009-10 (time constant), and in the year prior to the specific admission concerned (time-varying).

Estimates of mean differences in DAOH between the comparison groups, overall and for pre-specified subgroups (CATEs) were reported with standard errors and confidence intervals (CI) obtained with the non-parametric bootstrap (300 replications), allowing for the clustering of individuals within hospitals. The 2SLS approach used the same model specification and selection (including covariates used for confounding adjustment) to report estimates overall and for subgroups.

4.2.2. Case study: results

The study reported somewhat similar that for both methods the 95% CIs surrounding the mean differences included zero (Figure 5). Beneath this overall result, the LIV approach reported evidence that the effectiveness of ES was heterogeneous according to pre-specified subgroups. In particular, for all three conditions, ES led to lower DAOH for patients who had severe levels of frailty, and for those with acute appendicitis, ES was less effective for older patients (aged 80-84) or those with three of more comorbidities. By contrast, the 2SLS approach, which failed to account for unobserved heterogeneity (e.g., disease severity), did not report any substantive differences in relative effectiveness according to patient subgroup (Figure 5).

5. Discussion

This paper formally assessed the performance of the LIV methodology developed by Heckman and Vytlacil [4,5] and further extended by Basu, [14] to provide policy relevant estimates of ATE and CATE in settings that differed according to the form of heterogeneity, the sample size, and level of IV strength. We contrasted the performance of LIV with that of the widely-used 2SLS approach. The scenarios considered in the simulation study were directly motivated by gaps in the literature and by a comparative effectiveness study that used LIV in evaluating emergency surgery for three acute gastrointestinal conditions for subgroups of prime policy relevance. In the case study, overt and essential heterogeneity were important concerns, amid differing levels of IV strength and sample sizes, and these issues motivated the scenario of prime interest for the simulation study (Scenario D). However, we also considered scenarios, which can, in principle provide accurate estimates of ATE and CATEs with conventional IV methods such as 2SLS (Scenarios A and B). We compared the performance of the two methods, according to bias and statistical efficiency (RMSE).

Four preliminary findings of the simulation study are worth emphasising. First, our results suggest that while LIV performs better according to increasing levels of IV strength and sample size, this estimator reports relatively low levels of bias in estimates of the ATE and CATEs across all scenarios including those with essential heterogeneity. These findings compliment those of Basu[14] in evaluating the reliance of the estimator on the relevance condition as well as the consistency of the estimator, but also by considering a wider range of assumptions about heterogeneity.

Second, our results suggest that 2SLS reports biased estimates of the ATE and CATEs in the presence of essential heterogeneity, except in those cases where the instrument is very strong (F-statistic above 500). These results are consistent with previous findings that 2SLS estimates cannot generally be extrapolated to broader populations beyond the compliers unless restrictive assumptions are made about the heterogeneity of treatment effects. [11,12] However, our results suggest that, even under homogenous treatment effects, 2SLS provides biased estimates of the ATE, in scenarios where the F-statistic is low, but the requisite magnitude of the F-statistic also depends on the sample size and the form of heterogeneity.

This finding further emphasises the inadequacy of guidance resting solely on a 'rule of thumb' for a single setting, the target F-statistic, and highlights the importance of these wider considerations when interpreting a study's results.

Thirdly, while 2SLS can reliably estimate CATEs in the presence of effect homogeneity or overt heterogeneity given a sufficiently strong IV or large enough sample, in the presence of essential heterogeneity, as theory would suggest, 2SLS can give extremely biased estimates of CATEs, and so in settings where essential heterogeneity is anticipated, 2SLS should not be used to estimate CATEs. In contrast, the LIV method provided estimates with low bias in the presence of overt and/or essential heterogeneity, provided the F-statistic was greater than 50. Interestingly, for the estimates of the CATEs, we find that as the sample size increases, an increase in the F-statistic is less beneficial in mitigating bias and reducing RMSE, in line with the observation that a given increase in the F-statistic has less impact on compliance rates at larger sample sizes.

Finally, LIV generally reported lower levels of RMSE than 2SLS, in particular for estimating the CATEs. However, it is important to note that here the propensity score and outcome models underlying the LIV method are correctly specified, and that performance may deteriorate where this is not the case. Data adaptive approaches could prove useful where model specification is not known.

The findings from the simulation study are informative in interpreting the CATE estimates in the ESORT study. The results offer reassurance that in such settings where essential heterogeneity would appear inevitable, that a LIV approach can provide unbiased estimate of policy-relevant estimands such as CATE, with sample sizes and F-statistics smaller than those of the ESORT study. Here, the LIV approach was able to report relative effectiveness according to subgroup, and the finding that for patients with high levels of frailty ES was not cost-effective (or cost-effective), provides important evidence to inform policy, and contributes to shared decision-making.[33]

This study has several strengths. First, it builds on insights and hypotheses raised by a large observational study using EHRs from England. The ESORT study illustrates the main challenges of using LIV methods for comparative effectiveness research and its findings in relation to IV strength, sample size requirements directly informed the scenarios considered in the simulation study. Second, while the uptake of LIV methods has been limited almost entirely to settings with essential heterogeneity, the simulation study considers different forms of heterogeneity of treatment effects as well as the scenario where treatment effects are assumed to be homogeneous in the study population. Future work will expand the simulation study to incorporate other well-known issues of IVs methods, including the challenges in applying IV estimation methods to non-linear outcome data. [48,49] Previous research has shown that the power of 2SLS conveyed by conventional F-statistic values is low. [25,26] In this future work, we will therefore consider the implications of sample size and instrument strength for the power of LIV analyses and confidence interval coverage. Future work will also formally assess whether imbalances in treatment assignment rates are detrimental to consistency and power of LIV inferences. This is an important concern for applied work using EHRs. For instance, the observed difference in the prevalence of ES and NES in ESORT (90/10 in the cohort with appendicitis) could reduce the power of the analysis. [50]

Acknowledgements

We would like to acknowledge Professor Luke Keele for helpful discussions, and other members of the ESORT study team. We would like to acknowledge the feedback from those who participated in the session and particularly the paper discussant at the HESG Summer 2022 conference, Dr Manuel Gomes.

References

1. Heckman JJ, Urzua S, Vytlacil E. Understanding instrumental variables in models with essential heterogeneity [Internet]. Vol. 88, NBER Working Paper No. 12574. Cambridge, MA; 2006. Available from: http://www.nber.org/papers/w12574
2. Heckman JJ, Vytlacil E. Structural equations, treatment effects and econometric policy evaluation. Econometrica. 2005;73(3):669–738.
3. Bjorklund A, Moffitt R. Estimation of Wage Gains and Welfare Gains from Self-Selection Models. IUI Working Paper, No. 105,. Stockholm; 1983.
4. Heckman JJ, Vytlacil EJ. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proc Natl Acad Sci U S A. 1999;96:4730–4.
5. Heckman JJ, Vytlacil EJ. Policy-Relevant Treatment Effects. Am Econ Rev. 2001;91(2):107–11.
6. Basu A, Heckman JJ, Navarro-Lozano S, Urzúa S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Econ. 2007;16(2007):1133–57.
7. Angrist JD, Fernández-Val I. ExtrapoLATE-ing: External validity and overidentification in the LATE framework. Adv Econ Econom Tenth World Congr Vol 3, Econom. 2011;401–34.
8. Terza J V., Bradford WD, Dismuke CE. The use of linear instrumental variables methods in health services research and health economics: A cautionary note. Health Serv Res [Internet]. 2008 Jun 1 [cited 2021 Jun 15];43(3):1102–20. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1475-6773.2007.00807.x
9. Terza J V., Basu A, Rathouz PJ. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. J Health Econ [Internet]. 2008 May [cited 2021 Jun 15];27(3):531–43. Available from: /pmc/articles/PMC2494557/
10. Martínez-Camblor P, MacKenzie TA, Staiger DO, Goodney PP, James O'Malley A. An instrumental variable procedure for estimating Cox models with non-proportional hazards in the presence of unmeasured confounding. J R Stat Soc Ser C Appl Stat. 2019;68(4):985–1005.
11. Chapman CG, Brooks JM. Treatment Effect Estimation Using Nonlinear Two-Stage Instrumental Variable Estimators: Another Cautionary Note. Health Serv Res [Internet]. 2016 Dec 1 [cited 2021 Jun 15];51(6):2375–94. Available from: https://pubmed.ncbi.nlm.nih.gov/26891780/

12. Brooks JM, Chapman CG, Schroeder MC. Understanding Treatment Effect Estimates When Treatment Effects Are Heterogeneous for More Than One Outcome. Appl Health Econ Health Policy. 2018;16(3):381–93.

13. Basu A, Coe NB, Chapman CG. 2SLS versus 2SRI: Appropriate methods for rare outcomes and/or rare exposures. Health Econ. 2018;27(6):937–55.

14. Basu A. Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. J Appl Econom. 2014;29:671–91.

15. Basu A, Jones AM, Dias PR. Heterogeneity in the impact of type of schooling on adult health and lifestyle. J Health Econ [Internet]. 2018;57:1–14. Available from: https://doi.org/10.1016/j.jhealeco.2017.10.007

16. Grieve R, O'Neill S, Basu A, Keele L, Rowan KM, Harris S. Analysis of Benefit of Intensive Care Unit Transfer for Deteriorating Ward Patients: A Patient-Centered Approach to Clinical Evaluation. JAMA Netw Open. 2019 Feb 1;2(2):1–13.

17. Reynolds K, Barton LJ, Basu A, Fischer H, Arterburn DE, Barthold D, et al. Comparative Effectiveness of Gastric Bypass and Vertical Sleeve Gastrectomy for Hypertension Remission and Relapse: The ENGAGE CVD Study. Hypertension. 2021;78(4):1116–25.

18. Cornelissen T, Dustmann C, Raute A, Schönberg U. Who benefits from universal child care? Estimating marginal returns to early child care attendance. J Polit Econ [Internet]. 2018;126(6):2356–409. Available from: http://www.christiandustmann.com/content/4-research/2-who-benefits-from-universal-childcare-estimating-marginal-returns-to-early-childcare-attendance/cornelissen_etal_2017_jpe_forthcoming.pdf

19. Nelson CR, Startz R. Some further results on the exact small sample properties of the instrumental variables estimator. Econometrica. 1990;967–76.

20. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J Am Stat Assoc. 1995;90(430):443–50.

21. Stock JH, Wright J, Yogo M. GMM , Weak Instruments , and Weak Identification. J Bus Econ Stat Symp. 2002;

22. Small DS, Rosenbaum PR. War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. Vol. 103, Journal of the American Statistical Association. 2008. 924–933 p.

23. Staiger D, Stock JH. Instrumental Variables Regression with Weak Instruments. Econometrica. 1997;65(3):557.

24. Andrews I, Stock JH, Sun L. Weak Instruments in Instrumental Variables Regression: Theory and Practice. Annu Rev Econom. 2019;11(Iv):727–53.

25. Lee D, McCrary J, Moreira MJ, Porter JR. Valid T-Ratio Inference for IV. arXiv. 2020;

26. Keane M, Neal T. A Practical Guide to Weak Instruments. 2021;16(1997).

27. Angrist J, Kolesár M. One Instrument to Rule Them All: The Bias and Coverage of Just-Id IV. SSRN Electron J. 2021;

28. Moffitt RA, Zahn M V. The Marginal Labor Supply Disincentives of Welfare : Evidence from Administrative Barriers to Participation. 2022;

29. Hutchings A, O'Neill S, Lugo-palacios DG, Moler-Zapata S, Silverwood R, Cromwell D, et al. Effectiveness of emergency surgery for five common acute conditions: an instrumental variable analysis of a national routine database. Anaesthesia. 2022;In Press.

30. ESORT Study Group. Emergency Surgery Or NoT (ESORT) study [Internet]. Study protocol. 2020. Available from: https://www.lshtm.ac.uk/media/38711

31. Hutchings A, Moler-Zapata S, O'Neill S, Smart N, Hinchliffe R, Cromwell D, et al. Variation in the rates of emergency surgery amongst emergency admissions to hospital for common acute conditions. BJS Open. 2021;00(0).

32. Keele L, Sharoky CE, Sellers MM, Wirtalla CJ, Kelz RR. An instrumental variables design for the effect of emergency general surgery. Epidemiol Method. 2018;7(1).

33. Moler-Zapata S, Grieve R, Lugo-Palacios D, Hutchings A, Silverwood R, Keele L, et al. Local instrumental variable methods to address confounding and heterogeneity when using electronic health records: an application to emergency surgery. Med Decis Mak. 2022;0(0):1–17.

34. Neyman J. On the application of probability theory to agricultural experiments. Stat Sci. 1990;5:463–480.

35. Rubin D. B. Estimating causal effects of treatment in randomized and nonrandomized studies. J Educ Psychol [Internet]. 1974;66(5):688–701. Available from: http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf

36. Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996;91(434):444–55.

37. Abadie A. Semiparametric instrumental variable estimation of treatment response models. Vol. 113, Journal of Econometrics. 2003. 231–263 p.

38. Tan Z. Regression and weighting methods for causal inference using instrumental variables. J Am Stat Assoc [Internet]. 2006 [cited 2021 Nov 22];101(476):1607–18. Available from: https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

39. Imbens GW, Angrist JD. Identification and Estimation of Local Average Treatment Effects. Econometrica. 1994 Mar;62(2):467.

40. Vytlacil E. Independence, monotonicity and latent index models: an equivalence result. Econometrica. 2002;70(1):331–41.

41. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. Stat Med. 2014;33(13):2297–340.

42. Basu A. Person-centered treatment (PeT) effects: Individualized treatment effects using instrumental variables. Stata J. 2015;15(2):397–410.

43. Little RJ, Long Q, Lin X. A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. Biometrics. 2009;65(2):640–9.

44. Basu A, Rathouz PJ. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. Biostatistics. 2005;6(1):93–109.

45. Stata Corp Lp. Stata Statistical Software: Release 7.0. College Station, TX. Vol. 2, p. 406. Stata Press Publication; 2001.

46. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. Wiley; 2000.

47. Pregibon D. Goodness of Link Tests for Generalized Linear Models. J R Stat Soc Ser C (Applied Stat. 1980;29(1):14–5.

48. Clarke P, Windmeijer F. Instrumental Variable Estimators for Binary Outcomes. C Work Pap Ser No 10/239 Instrum [Internet]. 2010; Available from: http://www.bristol.ac.uk/cmpo/Tel:

49. Vansteelandt S, Bowden J, Babanezhad M, Goetghebeur E. On instrumental variables estimation of causal odds ratios. Stat Sci. 2011;26(3):403–22.

50. Walker VM, Davies NM, Windmeijer F, Burgess S, Martin RM. Power calculator for instrumental variable analysis in pharmacoepidemiology. Int J Epidemiol. 2017;46(5):1627–32.

Tables and Figures

Table 1: Definition of the simulation scenarios

	Sample size	**F-statistic**	τ_0	τ_1	τ_2
Scenario A: Homogeneity	All sample sizes ($n = \{5000, 10000, 50000\}$)	All F-statistic values ($F_{Target} = \{10, 25, 50, 100, 500, 1000\}$)	50	0	0
Scenario B: Overt heterogeneity			40	20	0
Scenario C: Essential heterogeneity			40	0	20
Scenario D: Overt and essential heterogeneity			20	20	20

Table 2: Compliance rate by sample size (N) and F-statistic

F-statistic	**N = 5000**	**N = 10000**	**N = 50000**
10	8%	6%	3%
25	13%	9%	5%
50	18%	13%	6%
100	26%	20%	9%
500	56%	42%	21%
1000	73%	57%	29%

Figure 1: Bias plot for Average Treatment Effect (ATE) estimates across scenarios, with sample sizes of 5000 (left), 10000 (middle) and 50000 (right).

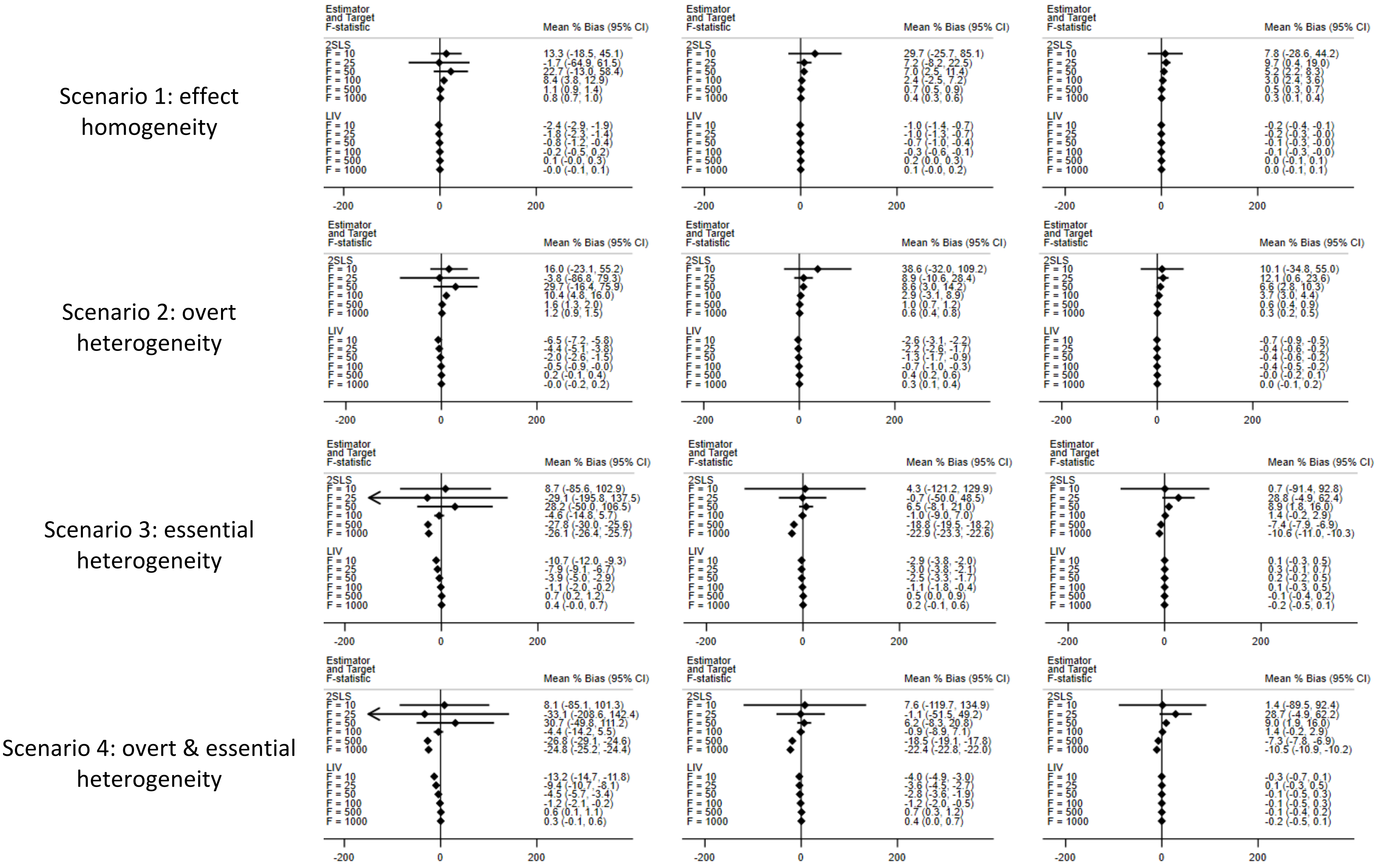


Figure 2: Bias plot for Conditional Average Treatment Effect (CATE) estimates across scenarios, with sample sizes of 5000 (left), 10000 (middle) and 50000 (right).

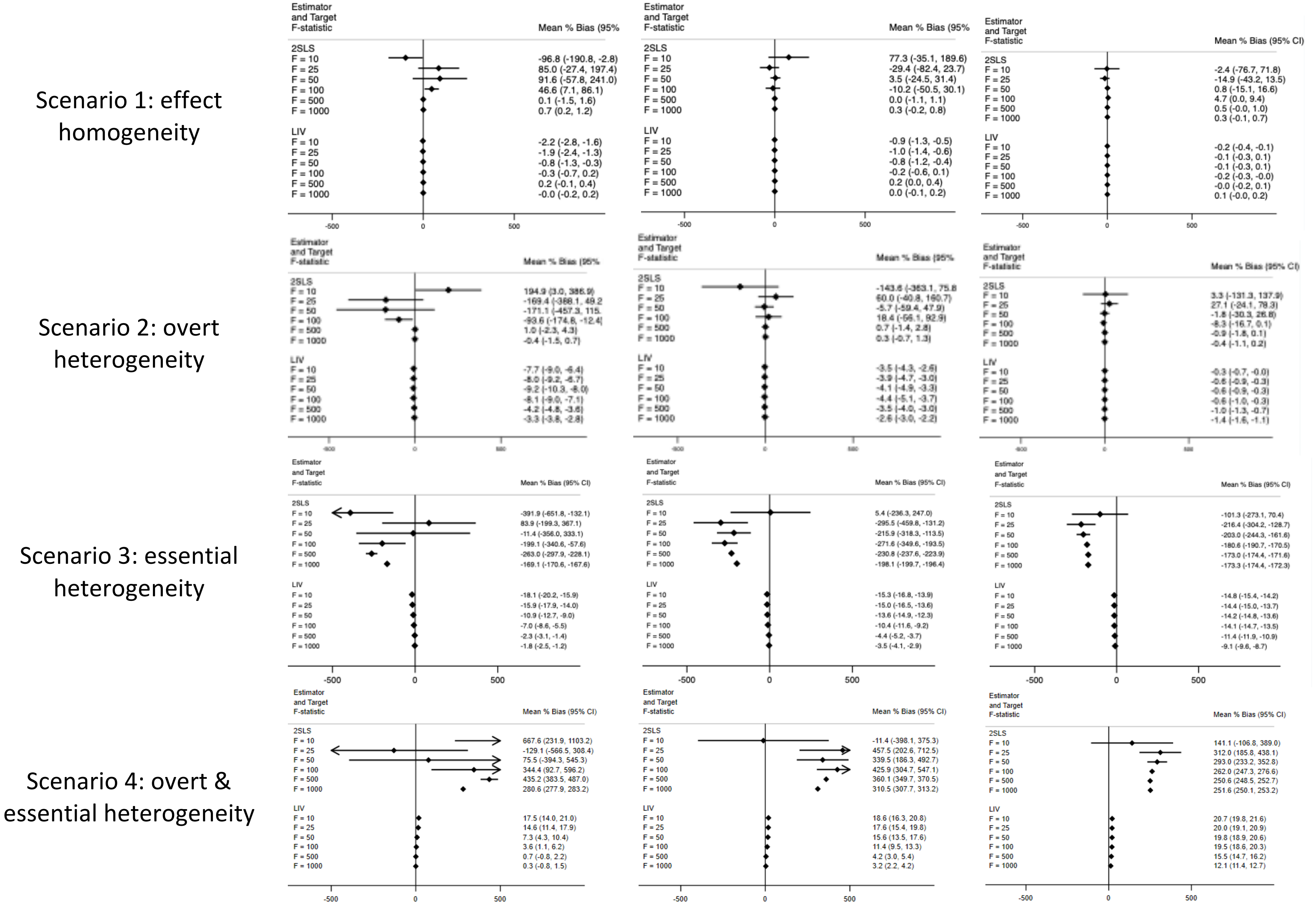


Figure 3: Root Mean Squared Error plots for Average Treatment Effect (ATE) estimates from 2SLS (dashed line) and LIV (solid line) across the scenarios

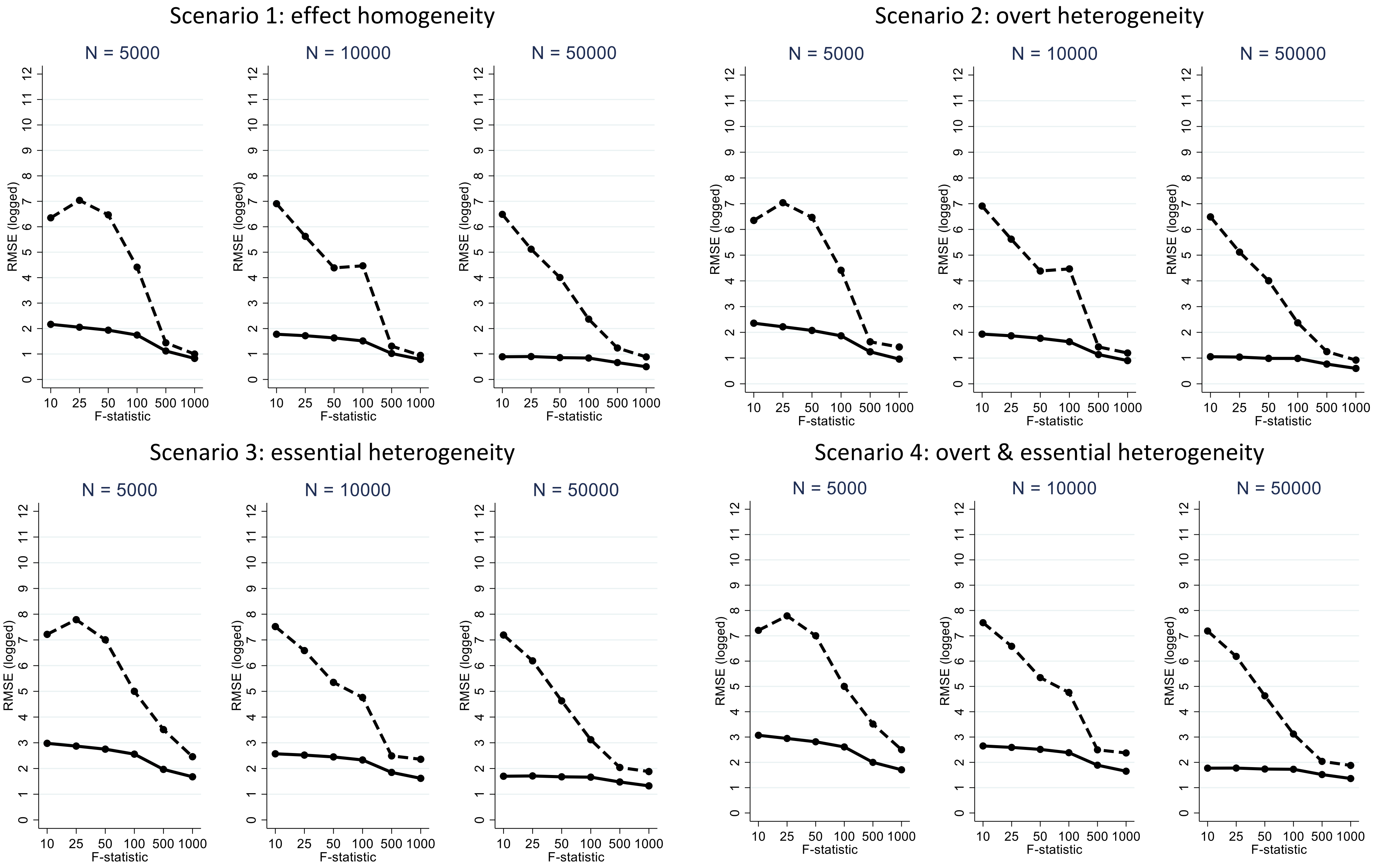
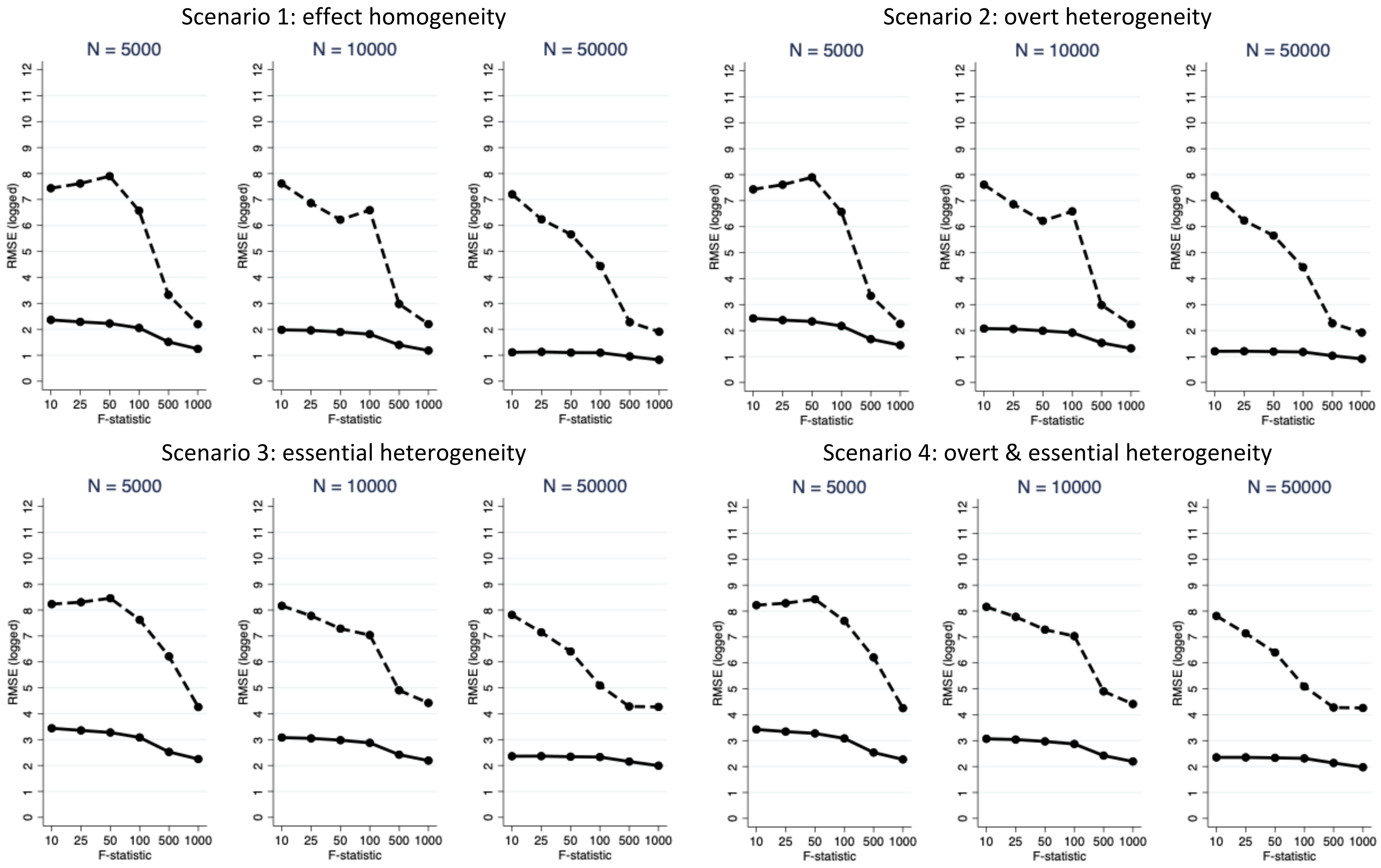


Figure 4: Root Mean Squared Error plots for Conditional Average Treatment Effect (CATE) estimates from 2SLS (dashed line) and LIV (solid line) across the scenarios



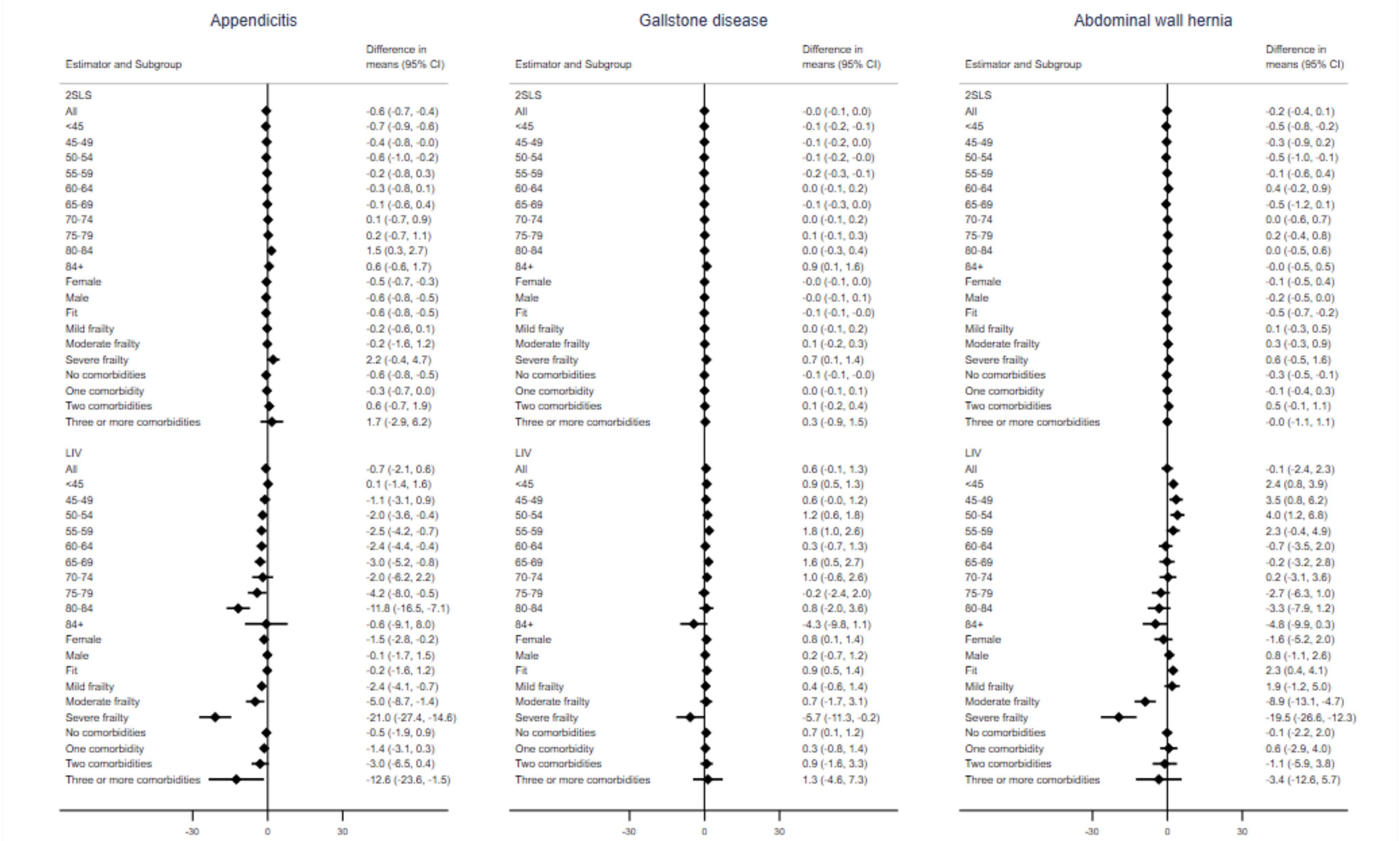
Scenario 1: effect homogeneity

Scenario 2: overt heterogeneity

Scenario 3: essential heterogeneity

Scenario 4: overt & essential heterogeneity

Figure 5: Mean differences in days alive and out of hospital (DAOH) between ES and NES for appendicitis (left), gallstone disease (centre) and hernia (right) subgroups



2SLS: two-stage least squares; CI: Confidence Interval; LIV: Local Instrumental variables.