Evaluating Nonexperimental Estimators for Multiple Treatments: Evidence from Experimental Data

Carlos A. Flores
University of Miami

Oscar A. Mitnik
University of Miami and IZA

Third Meeting of the Impact Evaluation Network at LACEA
December 3, 2009

Motivation

- There has been recent interest and increasing use of non-experimental estimators to evaluate programs with multiple, or multivalued, and continuous treatments
- There has been a focus on methodological advances and issues (Imbens, 2000; Lechner, 2001; Hirano & Imbens, 2004; Imai & van Dyk, 2004; Abadie, 2005; Flores, 2007; Cattaneo, 2009)
- And a great interest in evaluating such programs (Lechner, 2002a,b; Behrman et al., 2004; Frölich et al., 2004; Kluve et al., 2007; Plesca & Smith, 2007; Mitnik, 2008; Flores et al., 2009; etc.)

Objectives of this Paper

- The main question of this paper is: How well do different non-experimental estimators for multiple treatments work?
- We concentrate on estimators based on the unconfoundedness assumption (selection on observables)
- We study linear regression estimators, and partial mean and weighting estimators based on the generalized propensity score (GPS) –> probability of receiving a treatment given covariates
- We focus on estimators of the average outcome over all possible values of the treatment ("dose-response function")
- We analyze the key role of GPS in identifying individuals that are comparable (in observable characteristics) in each of the treatment groups

Previous Literature

- There is a long literature evaluating non-experimental estimators (Lalonde, 1986; Fraker & Maynard, 1987; Heckman & Hotz, 1989; Friedlander & Robins, 1995; Heckman et al., 1997/98; Dehejia & Wahba, 1999/02; Michalopoulos et al., 2004; Smith & Todd, 2005; Dehejia, 2005, Mueser et al., 2007)
- Virtually all focus has been on estimators for binary treatments
- Two approaches have been used in the literature to assess the value of methods based on unconfoundedness for estimation of average effects of binary treatments (Imbens, 2004):
  1. Uses data from experiment and non-experimental control groups –> aimed at assessing plausibility of unconfoundedness assumption and value of methods based on it
  2. Uses Monte Carlo simulations to evaluate the performance of alternative estimators under different scenarios –> helpful in identifying which particular estimators perform better in a given setting

Previous Literature (cont.)

- In this paper we will follow the first approach –> we want to assess the likely reliability of the methods based on the unconfoundedness assumption in a multiple treatment setting
- This approach in general uses data from a randomized experiment, and constructs a nonexperimental control group from additional data sets or locations
- Then, performance of nonexperimental estimators is evaluated by two alternative methods:
  1. Experiment results compared to results obtained from using experimental treatment group and nonexperimental control group
  2. Experimental and nonexperimental control groups are employed (implicit treatment effect=0)

Previous Literature (cont.)

What have we learned from previous literature?

- Basic message: We need to compare "comparable" individuals!
- Propensity score plays key role in identifying regions of data where treatment and control units are comparable
- Quality of data matters (we need good data)
- Comparing individuals in same local labor markets can be important

What Do We Do in this Paper?

- We have an experiment with control groups in multiple sites
- We use nonexperimental methods for multiple treatments to adjust for observable characteristics
- Our objective: Eliminate differences in outcomes across control groups in different sites simultaneously
- Why use data from experiment?
  1. Relatively "comparable" individuals (all welfare recipients)
  2. Same data for all individuals (and rich data)
  3. We use the experiment itself to develop benchmark measures to assess our nonexperimental results

What Do We Do in this Paper? (cont.)

- An "ideal" dataset to accomplish our objectives would have several nonexperimental control groups all belonging to the same labor market: we are not aware such dataset exists!
- However, having different geographic locations implies dealing with (potential) differences in local labor markets
  –> makes our exercise more difficult –> high yardstick
- Our approach is similar to that followed by Friedlander and Robins (1995), Michalopoulos, Bloom, and Hill (2004) and Hotz, Imbens and Mortimer (2005)
- Key difference: we focus on *simultaneously* comparing the individuals across *all* locations, not pairwise comparisons
  –> requires the use of nonexperimental methods for multiple treatments

Notation

- Each unit i, $i = 1, 2, \ldots, N$, comes from one of k sites
- Location indicator for unit i: $D_i \in \{1, 2, \ldots, k\}$
- Potential outcomes: $Y_i(t_d, d)$, t_d = treatment, d =location
- We focus only on control groups: $Y(0, d)$
- For each unit we observe: (Y_i, D_i, X_i),
  X_i = pre-treatment variables, $Y_i = Y(0, D_i)$
- Our parameters of interest in this paper are

$$\beta_d = E[Y(0, d)], \text{ for } d = 1, 2, \ldots, k$$

This gives average potential outcome under control treatment in location d for a unit randomly selected from the entire population (i.e., from any of the k sites)

Our Hypothesis

As we want to study whether our nonexperimental estimators can properly equalize average outcomes for control individuals across all sites, the hypothesis we test is

$$\beta_1 = \beta_2 = \ldots = \beta_k$$

Note:
This does not imply that the average potential outcome for controls in *each location* is the same across locations; i.e. this does not imply that

$$E[Y_i(0,d)\,|D_i = d] = E[Y_i(0,d)\,|D_i = f] \text{ for } d \neq f$$

Assessing Performance of Estimators

We assess estimators in two ways:

1. Perform a Wald test –> sensitive to estimators' variance
2. Use three measures of "distance":
   - Root mean square distance, $rmsd = \sqrt{\frac{1}{k}\sum_{d=1}^{k}(\widehat{\beta}_d - \overline{\beta})^2}$
   - Mean absolute distance, $mad = \frac{1}{k}\sum_{d=1}^{k}\left|\widehat{\beta}_d - \overline{\beta}\right|$
   - Maximum pair-wise distance among all estimates:

     $$Maximum\ Distance = \left|\max_{d=1,\ldots,k}\left\{\widehat{\beta}_d\right\} - \min_{d=1,\ldots,k}\left\{\widehat{\beta}_d\right\}\right|$$

     Where:
     - Outcomes standardized by their mean and S.D.–> comparability
     - $\widehat{\beta}_d$= an estimator of β applied to standardized data
     - $\overline{\beta}$ = mean value of $\widehat{\beta}$ among all sites

A successful estimator should make these distances "close" to zero

Assumptions

The estimators we study are based on two assumptions

Assumption 1 (Unconfounded site)

$$1\left(D_i = d\right) \perp Y_i\left(0, d\right) | X_i, \text{ for all } d \in \{1, 2, \ldots, k\}$$

- This assumption is similar to the one in Hotz, Imbens & Mortimer (2005) for the binary treatment case
- Referred as weak unconfoundedness by Imbens (1999, 2000)

Assumptions (cont.)

In addition, we impose a condition that guarantees that in infinite samples we are able to find individuals with the same values of the covariates across all k sites:

Assumption 2 (Simultaneous strict overlap) For all d and all x in the support of X

$$0 < \xi < \Pr(D_i = d|X = x), \text{ for some } \xi > 0$$

- Critical role for the asymptotic properties of semiparametric estimators of β_d
- Stronger than in binary case –> where is known as “strict overlap” (Busso et al. 2009a,b)
- Requires that for each individual in the population we are able to find comparable individuals in terms of covariates in each of the k sites

The Generalized Propensity Score (GPS)

Imbens (1999, 2000) defines the Generalized Propensity Score as:

$$r(d,x) = \Pr(D = d | X = x)$$

It defines several random variables:

- $R_i = r_i(D_i, X_i)$: cond. probability that i belongs to his own site
- $R_i^d = r_i(d, X_i)$: cond. probability that i belongs to site d

The GPS plays an important role in:

- Reducing dimentionality in estimation of β_d
- Identifying comparable individuals across sites

Estimators we Compare

Under Assumptions 1 and 2, and using iterated expectations, we can identify β_d as:

$$\beta_d = E[E[Y_i|D_i = d, X_i = x]]$$

This result suggests estimating β_d using a partial mean (Newey, 1994)

Thus we consider the following estimators

- "Raw" mean
- Linear regression-based partial mean (linear & flexible)
- GPS-based partial mean (parametric & non-parametric)
- Inverse Probability Weighting by GPS (w/o & w/ covariates)

We also compare regression-based estimators before & after overlap

GPS Estimation: parametric multinomial logit model
–> can be specified in a flexible way

Imposing Overlap Condition

We propose a rule that is less stringent than that previously used in the multiple treatment literature (e.g., Frölich et al., 2004)

- Let $R^d_{q,\{j \in A\}}$ denote the q-th quantile of the distribution of R^d over those individuals in subsample A
- Overlap region w/respect to particular site d given by subsample

$$Overlap_d = \left\{ i : R^d_i \geq \max \left\{ R^d_{q,\left\{j:D_j=d\right\}}, R^d_{q,\left\{j:D_j \neq d\right\}} \right\} \right\}$$

- Then, we define the overlap or common support region as

$$Overlap = \bigcap_{d=1}^{k} Overlap_d$$

- Compares R^d_i only among those in groups $D_i = d$ and $D_i \neq d$
- Exploits only lower tail of distributions of R^d_i
- We set $q = 0.002$ (also analyzed $q = 0$ to $q = 0.005$)

Our Data

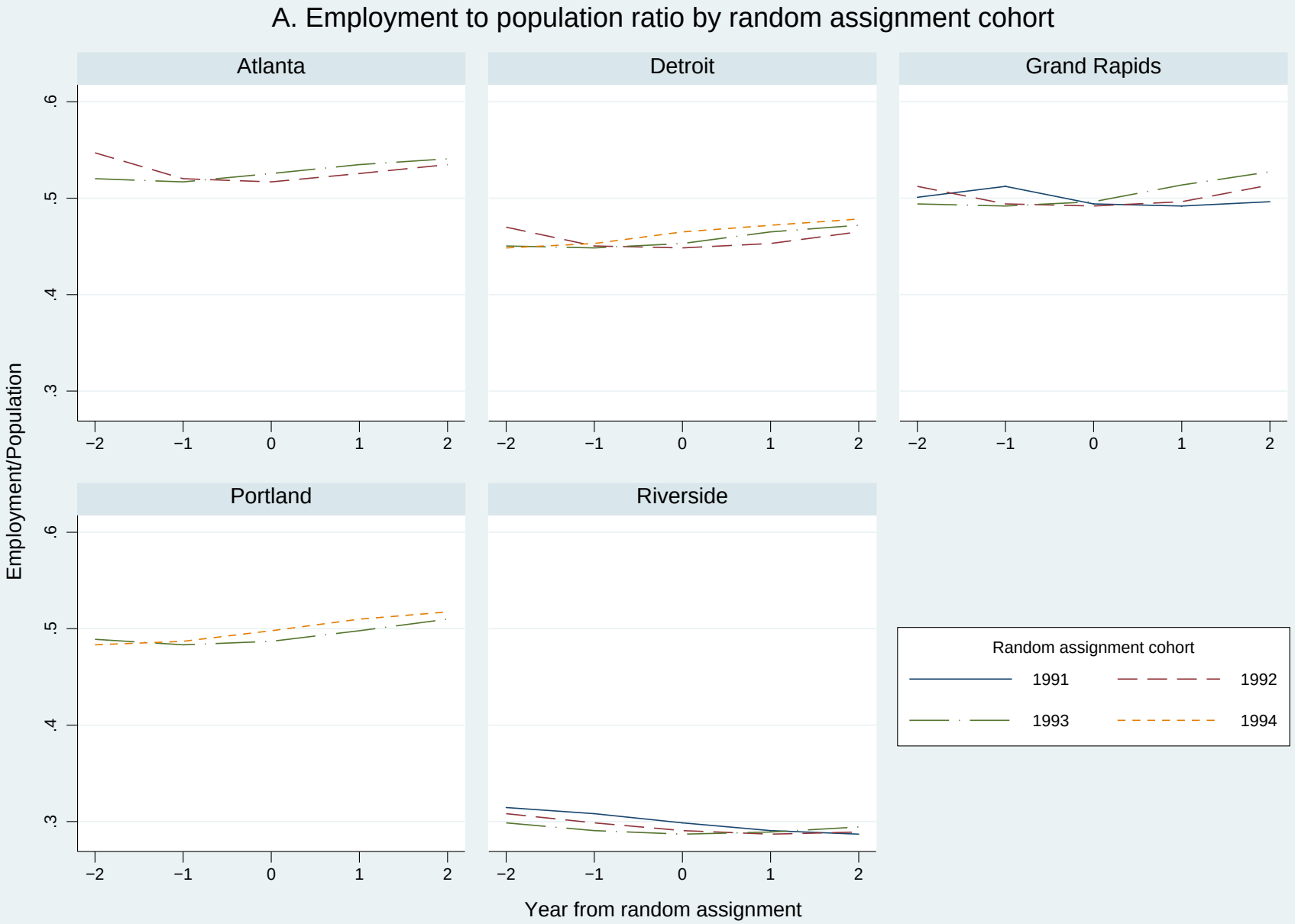
National Evaluation of Welfare-to-Work Strategies (NEWWS):

- U.S. experiment (randomization from 1991 to 1994) in 7 different sites
- Individuals randomly assigned to control group or LFA training or HCD training (in some sites to other types of programs)
- Because of treatment heterogeneity across sites, we concentrate only on comparing controls
- Only use women with non-missing data and drop 2 sites: Columbus, OH (not enough pre-RA information) & Oklahoma City, OK (randomization done at application, not on recipients)
- Analysis sample: 9,351 women in 5 sites: Atlanta, Detroit, Grand Rapids, Portland and Riverside
- Rich data before/after RA, both survey & administrative (some constraints because we use public-use version of the data)

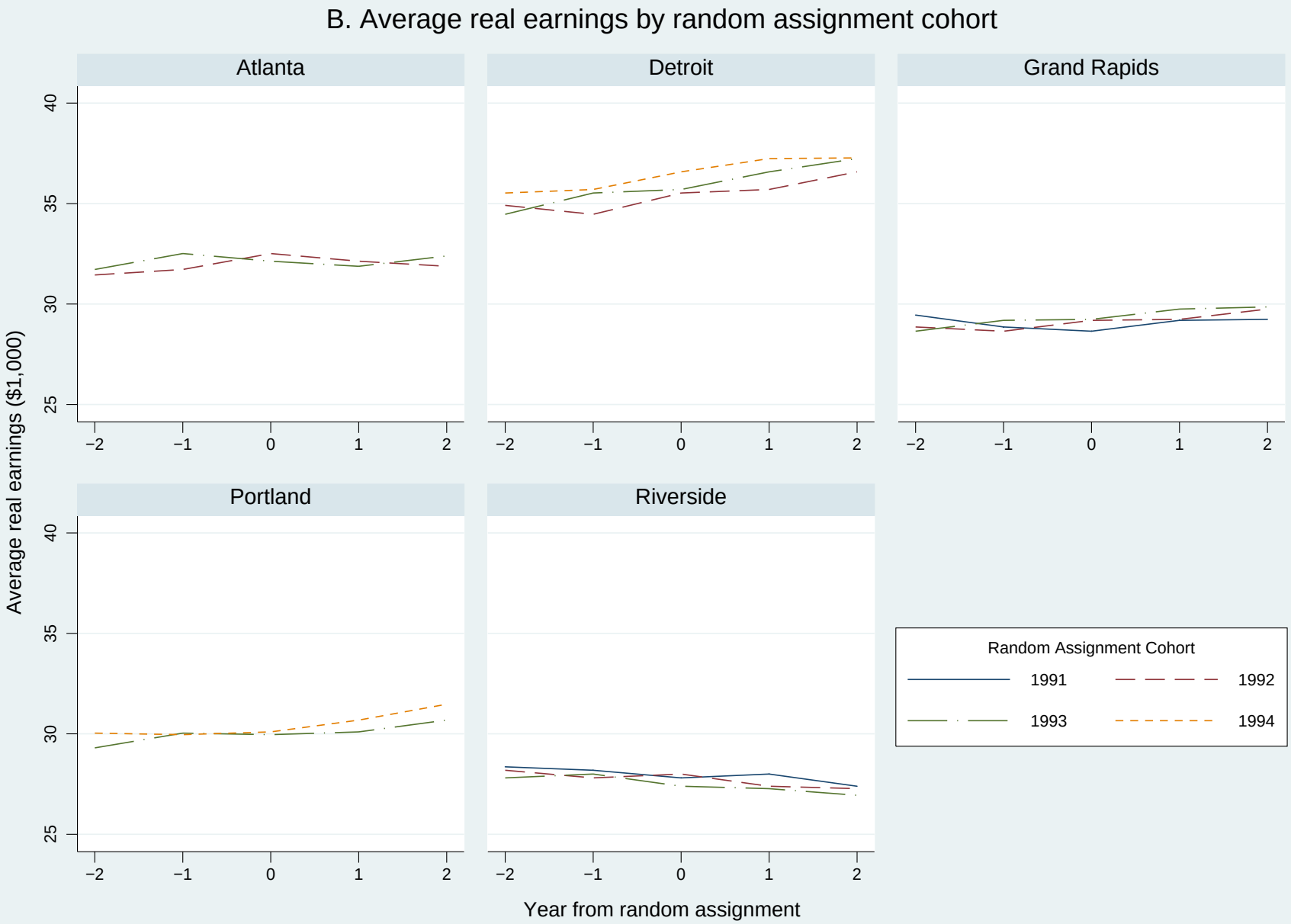
The Role of Local Economic Conditions

- The overlap assumption may fail even if individual characteristics are balanced across all sites, just because there are differences in local labor markets
- In our data we observe different cohorts for each site (determined by year of random assignment)
- This creates enough within-site variation to attempt controlling for pre-randomization differences in LECs across sites
  –> both in GPS estimation and in regression functions estimation
- We also explored adjusting by post-randomization variation (did not make much of a difference)

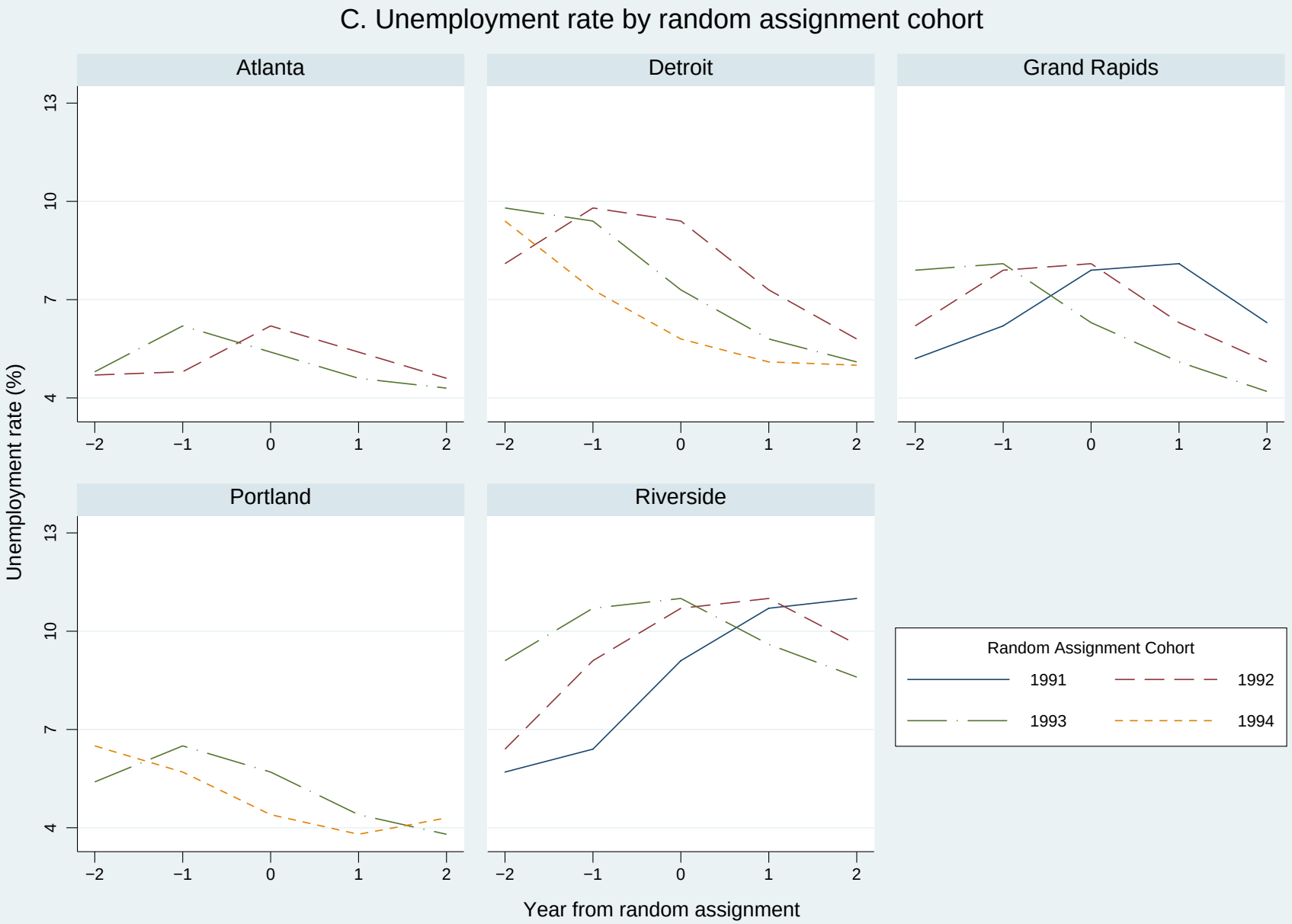
The Role of Local Economic Conditions (cont.) - Fig. 1



The Role of Local Economic Conditions (cont.) - Fig. 1



The Role of Local Economic Conditions (cont.) - Fig. 1



Outcomes

We analyze two outcomes

- Levels = 1{Ever employed during two years after RA}
- "Diff" = 1{Empl 2 yrs *after* RA} − 1{Empl 2 yrs *before* RA}

Balancing of Covariates Summary (Table 2 - 5 sites)

A. Joint equality of means tests for each covariate across all sites

Method	Number of covariates for which p-value ≤ 0.05 5 sites
Raw means before overlap	53
GPS-based Inverse Probability Weighting	11
Total number of covariates	53

B. Difference of means tests for each covariate - Each site versus all other sites pooled

Method	Number of covariates for which p-value ≤ 0.05 5 sites
Raw means before overlap	
Atlanta vs others	43
Detroit vs others	50
Grand Rapids vs others	35
Portland vs others	37
Riverside vs others	49
Blocking on GPS	
Atlanta vs others	4
Detroit vs others	6
Grand Rapids vs others	1
Portland vs others	4
Riverside vs others	2
Total number of covariates	53

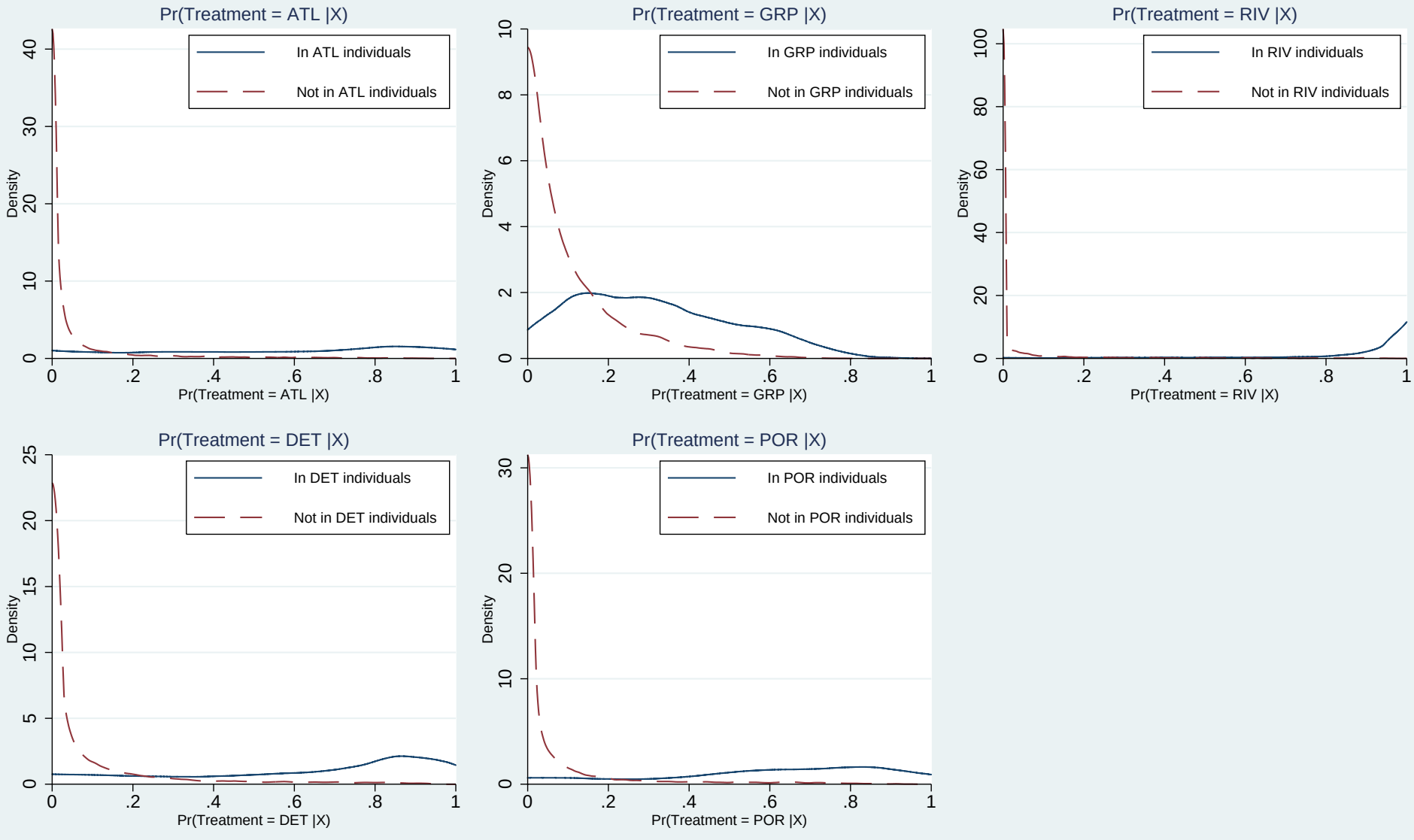
Note: GPS-based balancing tests are applied only to observations that satisfy the overlap condition.

Overlap Quality - Observations Dropped (Table 1 - 5 sites)

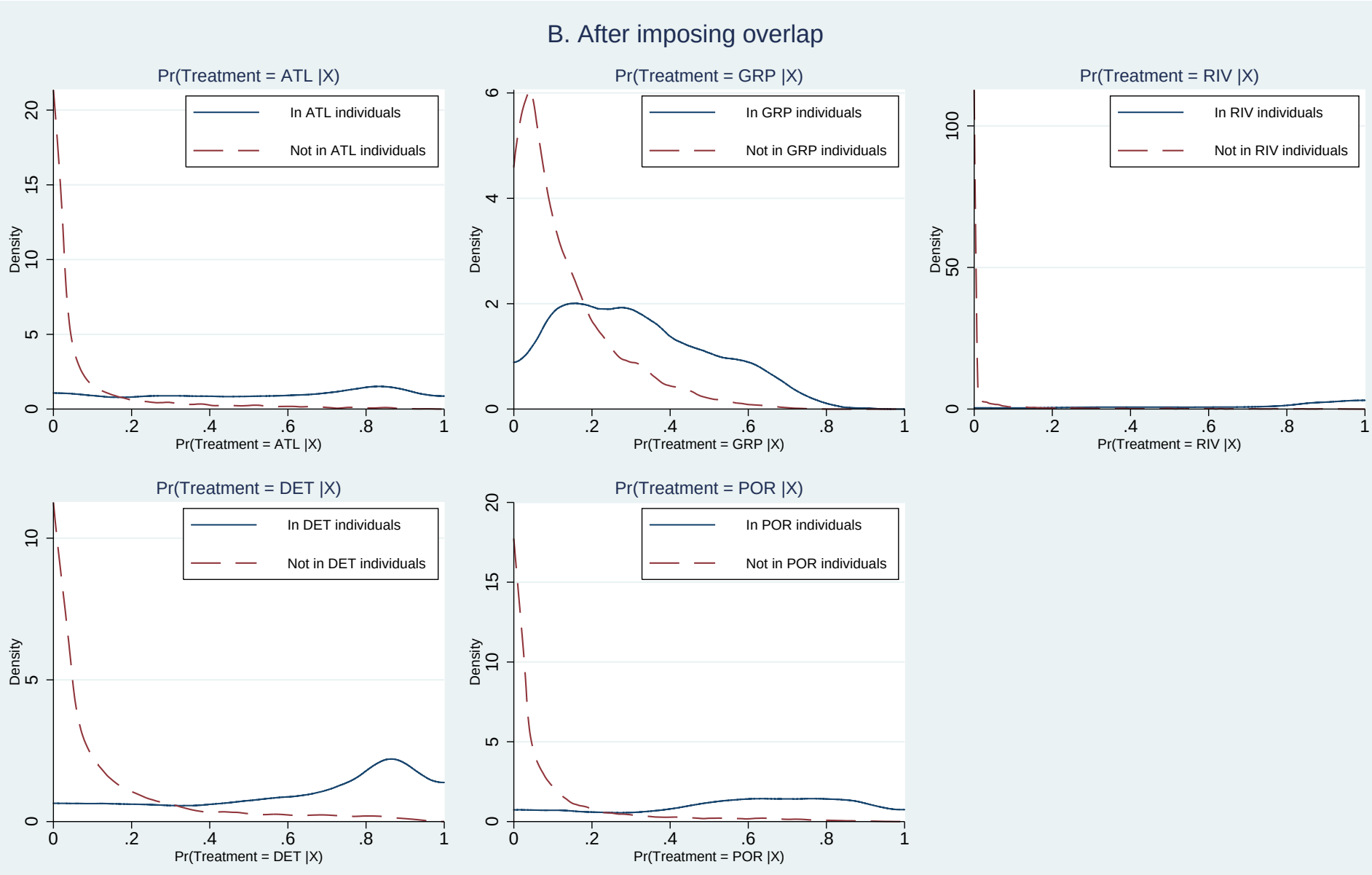
Site	Observations before overlap	Obs. after overlap	Obs. dropped due to ovlp (%)
Atlanta	1,372	1,184	13.7%
Detroit	2,037	1,943	4.6%
Grand Rapids	1,374	1,185	13.8%
Portland	1,740	1,432	17.7%
Riverside	2,828	1,107	60.9%
Total	9,351	6,851	26.7%

Overlap Quality - Kernel Densities - 5 sites (Fig. 2)

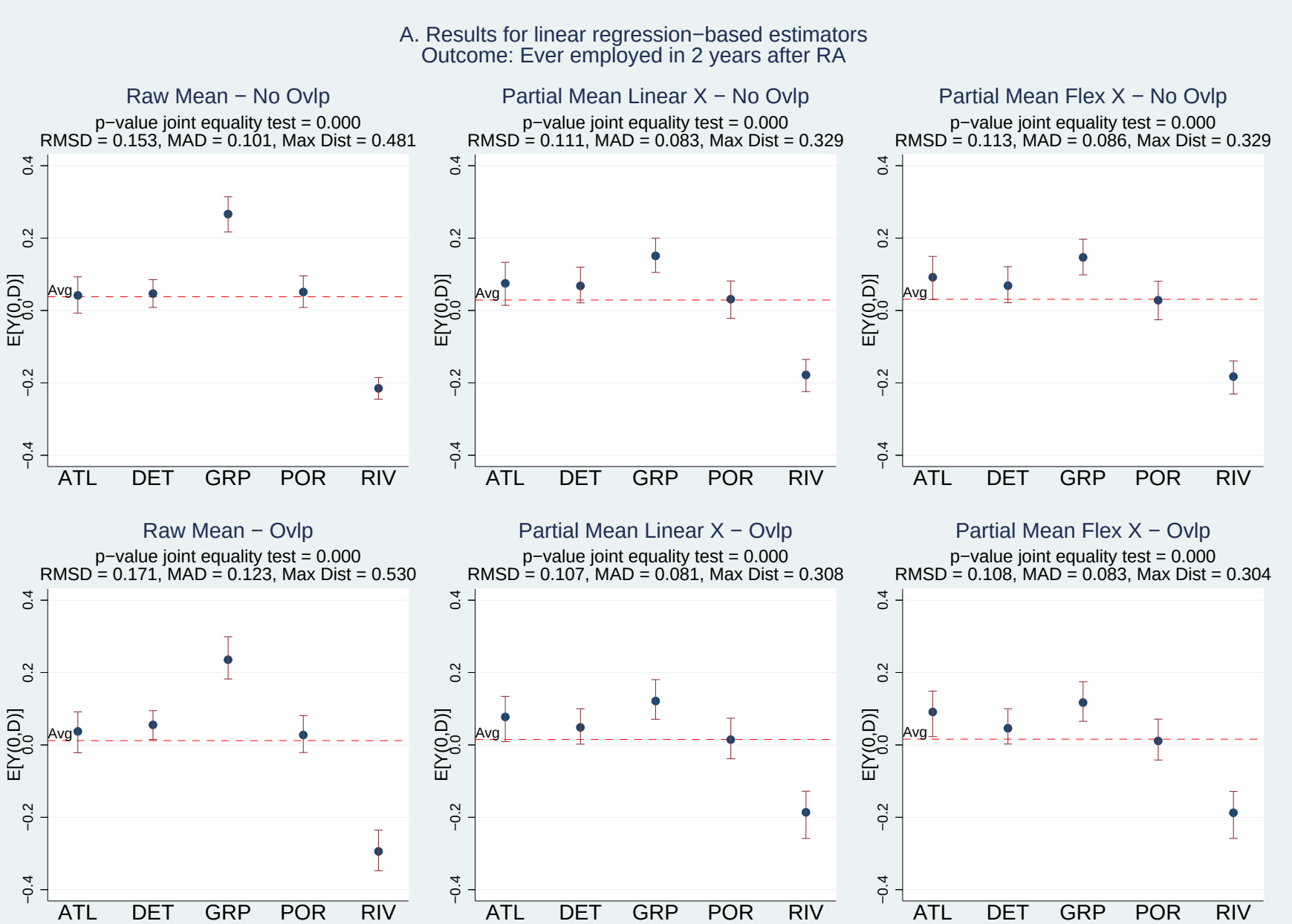
A. Before imposing overlap



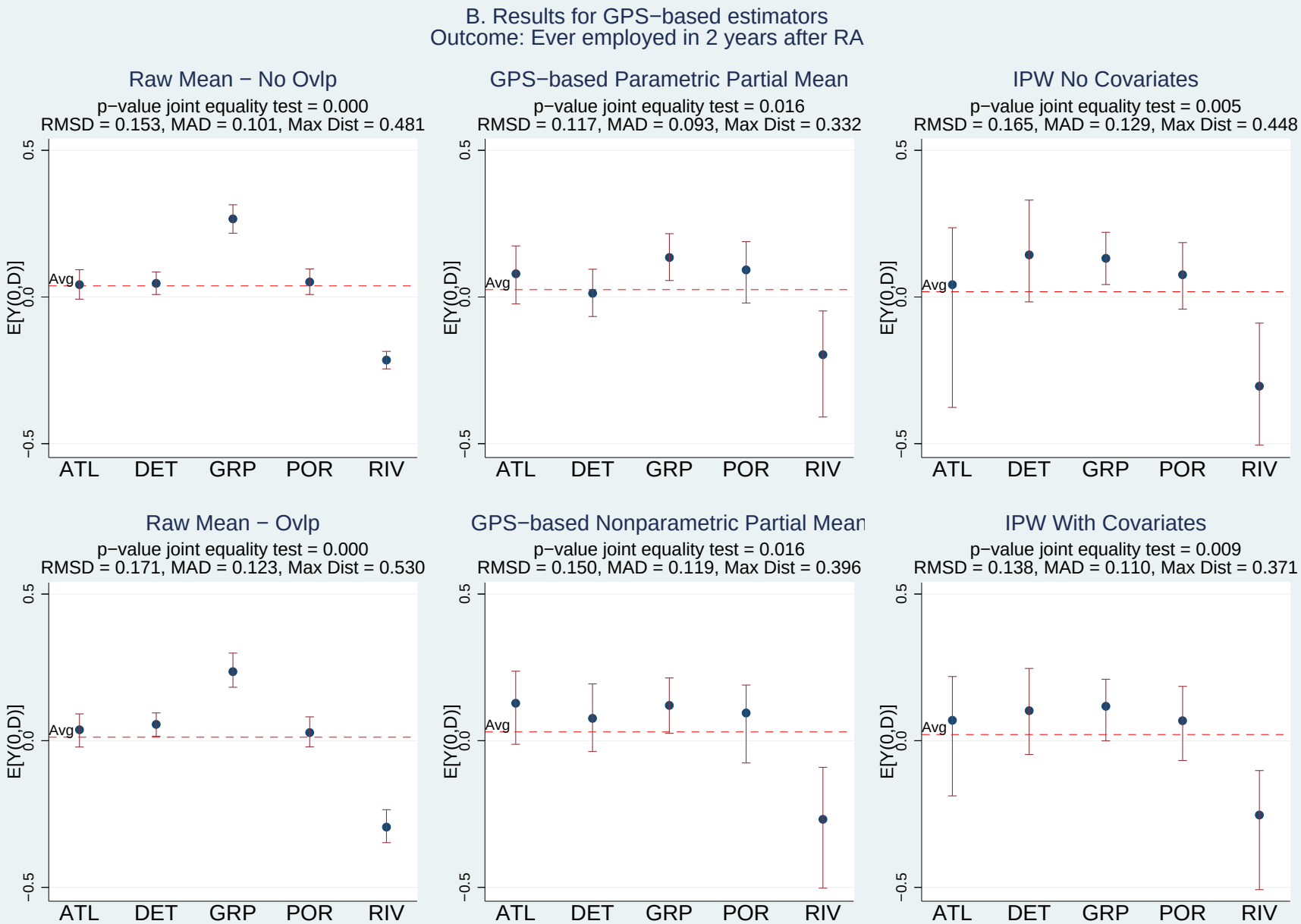
Overlap Quality - Kernel Densities - 5 sites (Fig. 2 - cont.)



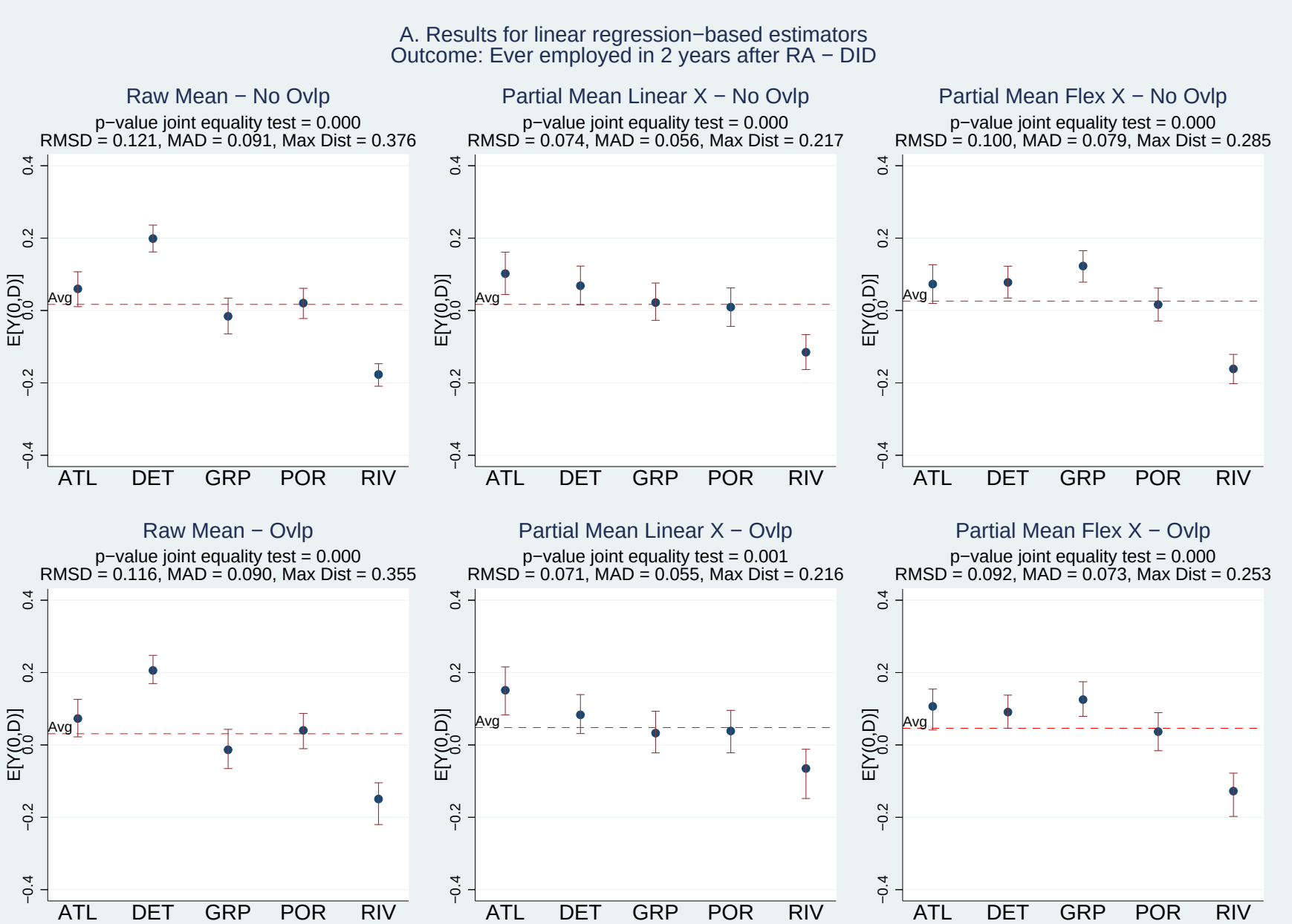
Outcome in Levels - 5 Sites (Figure 4.A)



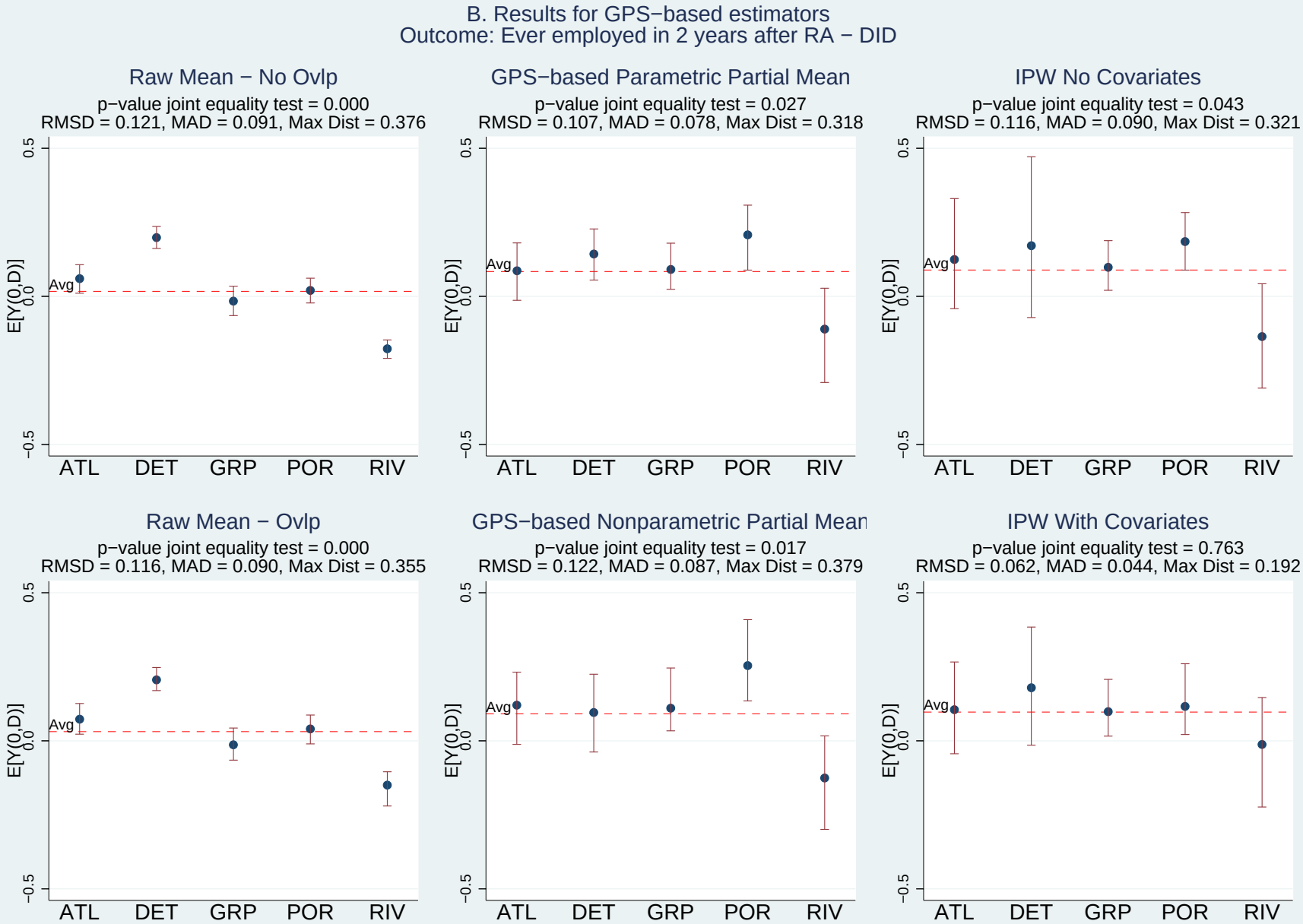
Outcome in Levels - 5 Sites (Figure 4.B)



Outcome in Differences - 5 Sites (Figure 5.A)



Outcome in Differences - 5 Sites (Figure 5.B)



Balancing of Covariates Summary (Table 2 - 5 vs. 4 sites)

A. Joint equality of means tests for each covariate across all sites

Method	Number of covariates for which p-value ≤ 0.05	
	5 sites	4 sites
Raw means before overlap	53	52
GPS-based Inverse Probability Weighting	11	5
Total number of covariates	53	53

B. Difference of means tests for each covariate - Each site versus all other sites pooled

Method	Number of covariates for which p-value ≤ 0.05	
	5 sites	4 sites
Raw means before overlap		
Atlanta vs others	43	36
Detroit vs others	50	47
Grand Rapids vs others	35	49
Portland vs others	37	34
Riverside vs others	49	-
Blocking on GPS		
Atlanta vs others	4	1
Detroit vs others	6	2
Grand Rapids vs others	1	1
Portland vs others	4	6
Riverside vs others	2	-
Total number of covariates	53	53

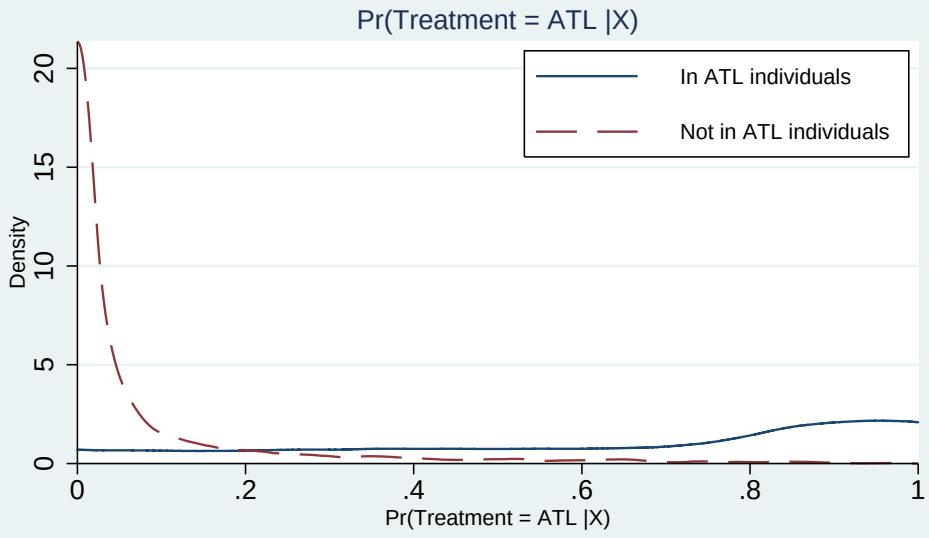
Note: GPS-based balancing tests are applied only to observations that satisfy the overlap condition.

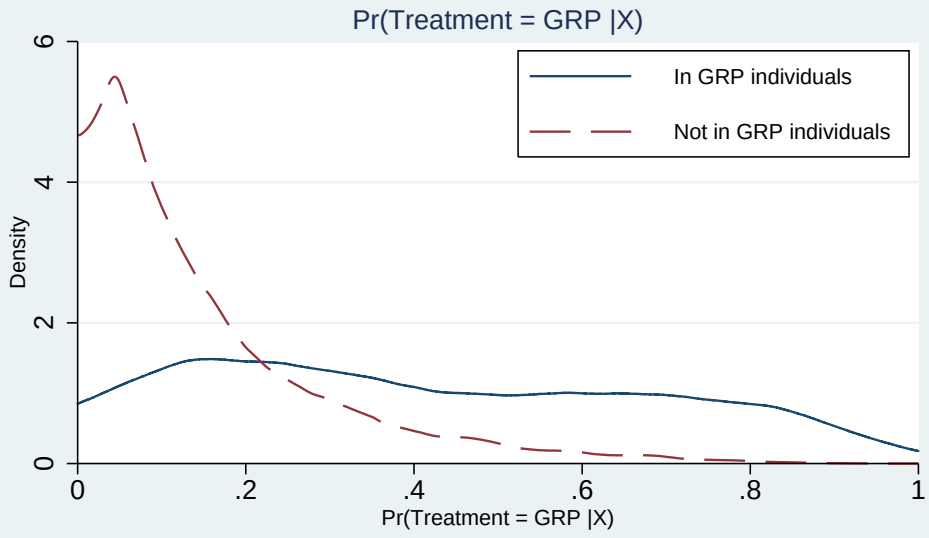
Overlap Quality - Obs. Dropped (Table 1 - 5 vs. 4 sites)

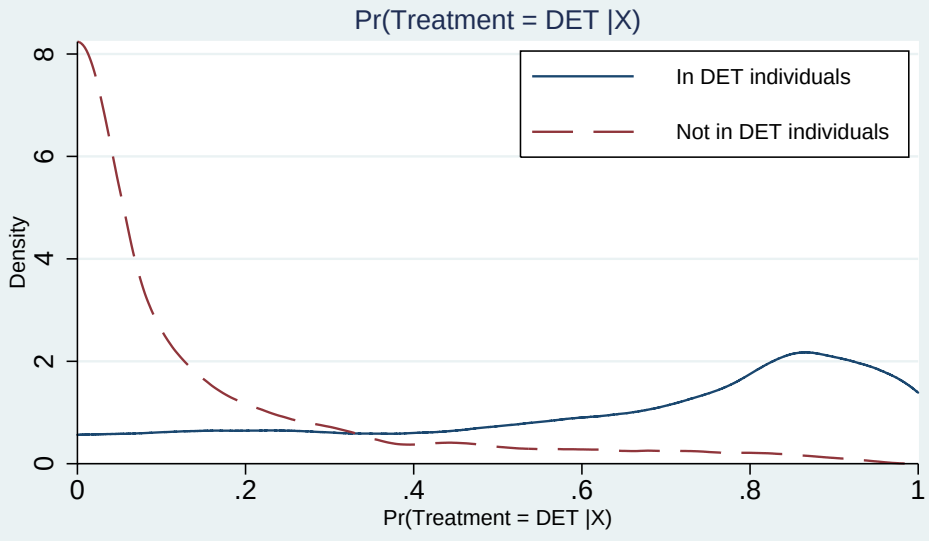
Site	Observations before overlap	Obs. after overlap		Obs. dropped due to ovlp (%)	
		5 sites	4 sites	5 sites	4 sites
Atlanta	1,372	1,184	1,245	13.7%	9.3%
Detroit	2,037	1,943	1,945	4.6%	4.5%
Grand Rapids	1,374	1,185	1,193	13.8%	13.2%
Portland	1,740	1,432	1,360	17.7%	21.8%
Riverside	2,828	1,107	-	60.9%	-
Total	9,351	6,851	5,743	26.7%	12.0%

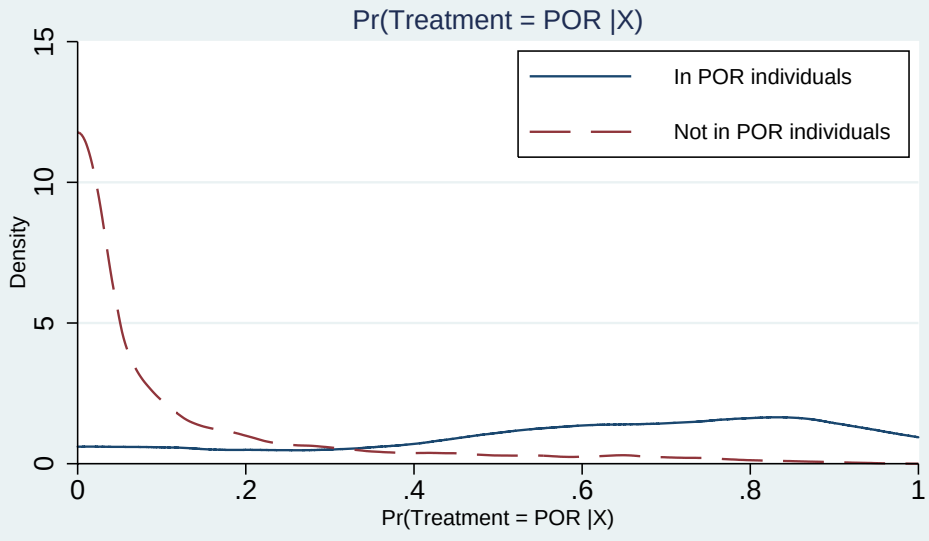
Overlap Quality - Kernel Densities - 4 sites (Fig. 3)

A. Before imposing overlap

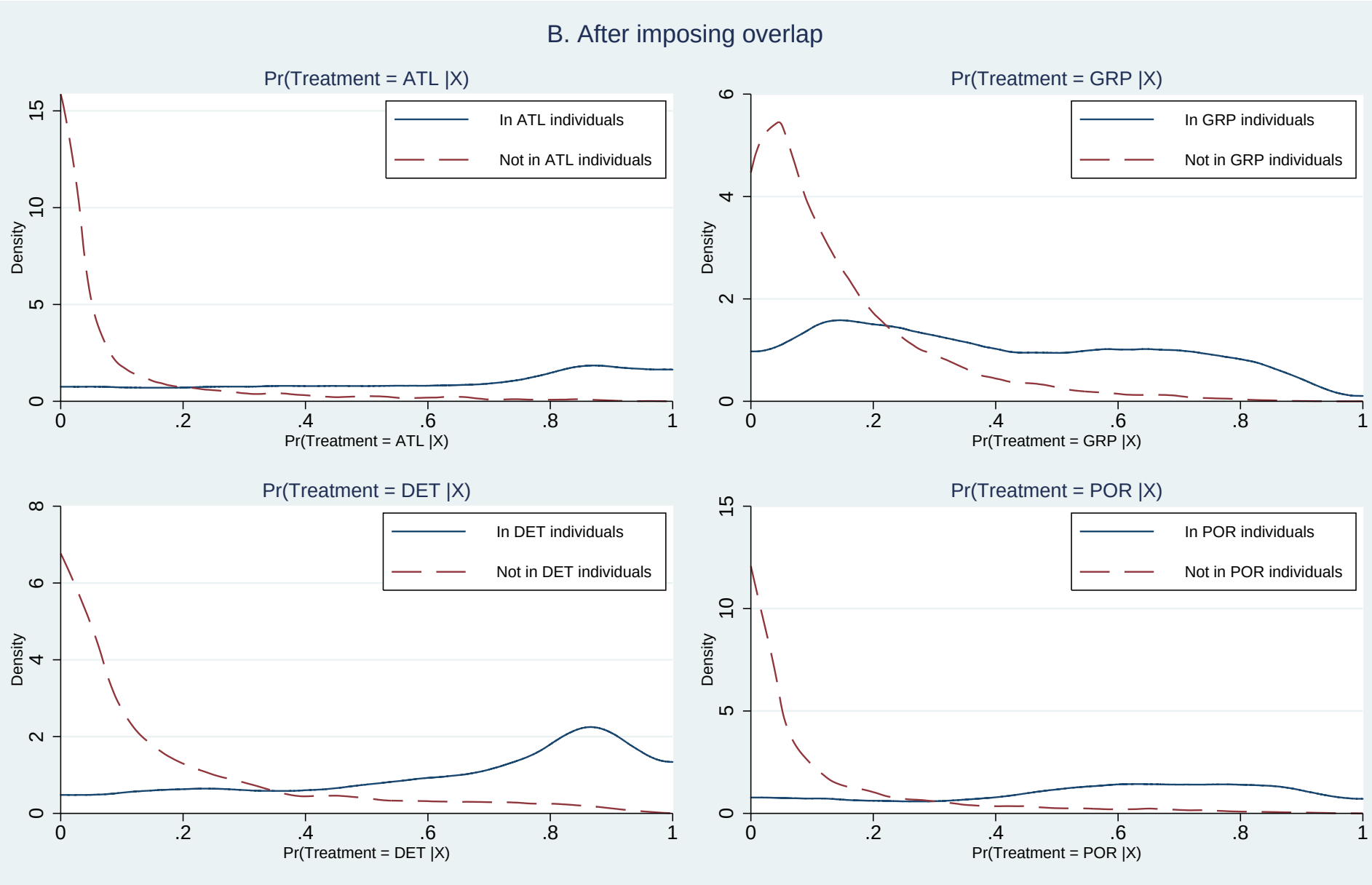




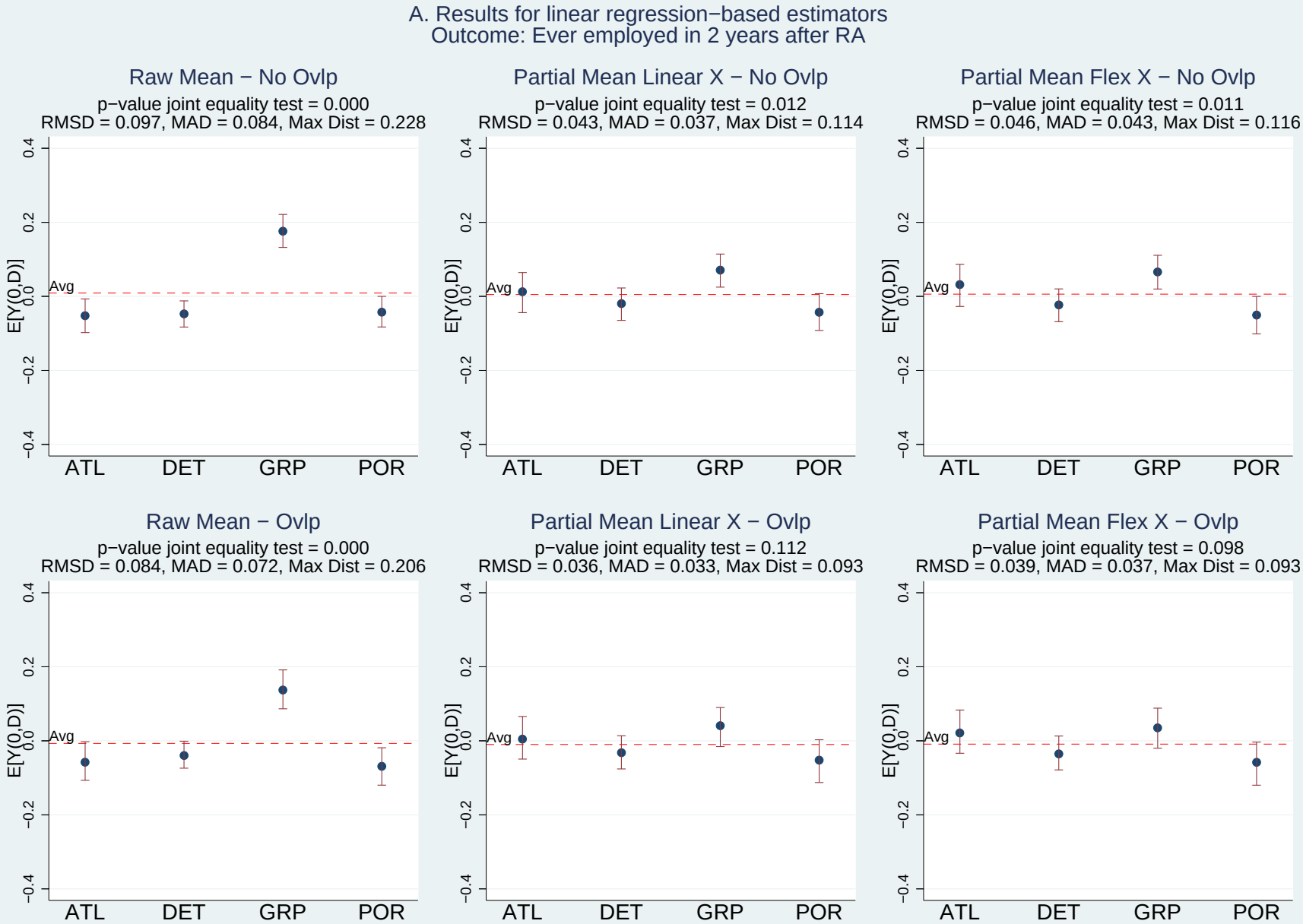




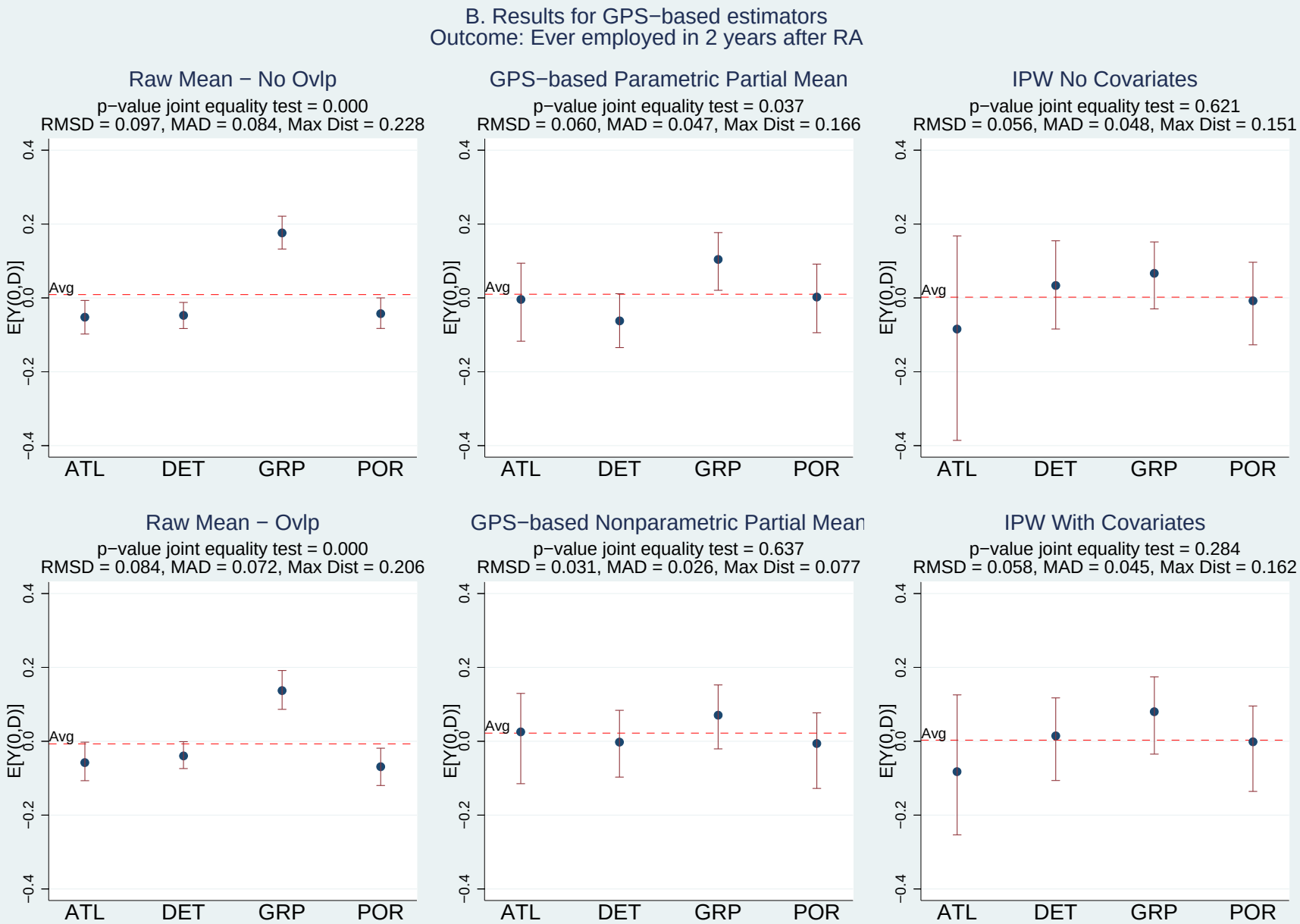
Overlap Quality - Kernel Densities - 4 sites (Fig. 3 - cont.)



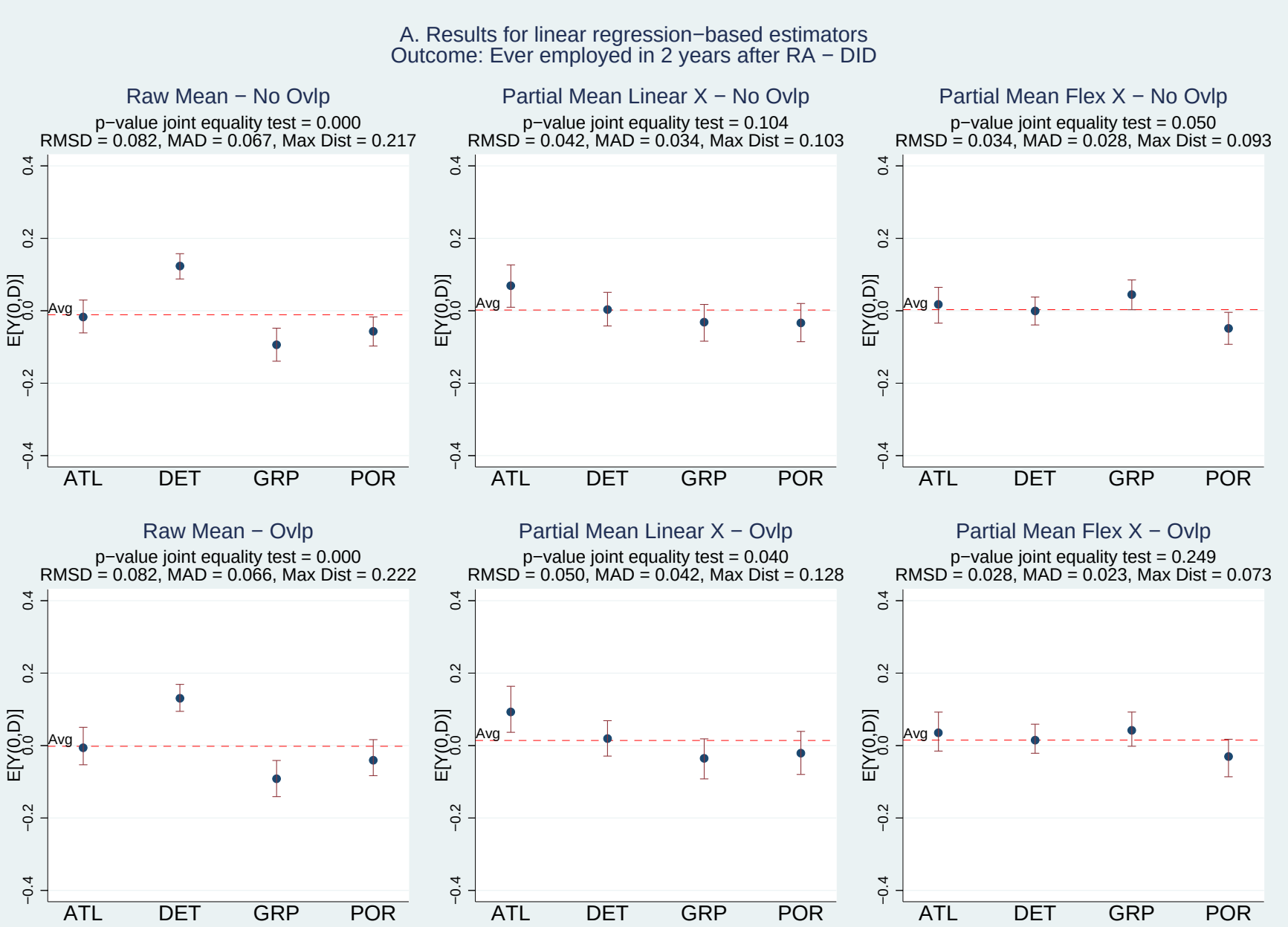
Outcome in Levels - 4 Sites (Figure 6.A)



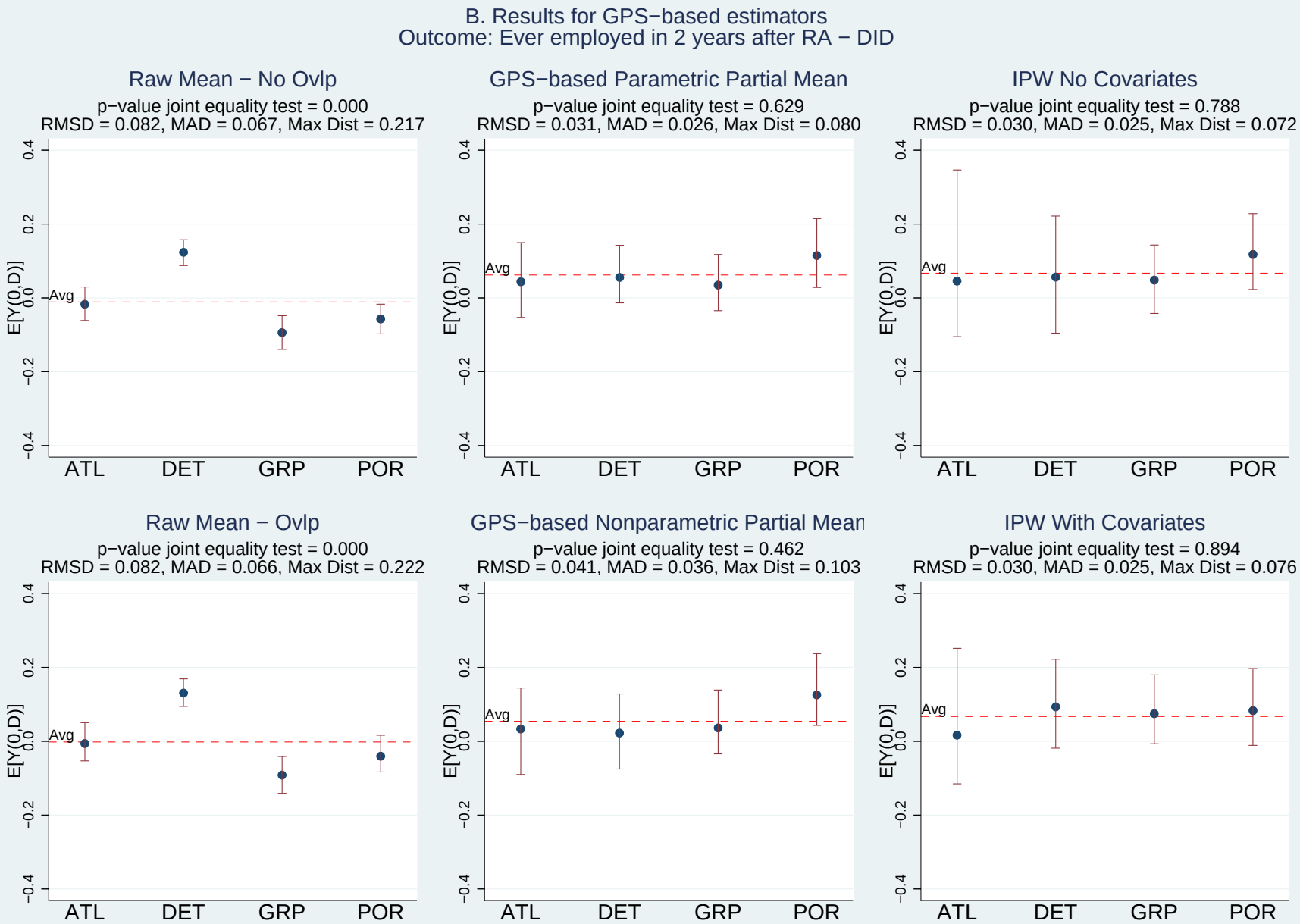
Outcome in Levels - 4 Sites (Figure 6.B)



Outcome in Differences - 4 Sites (Figure 7.A)



Outcome in Differences - 4 Sites (Figure 7.B)



How do we Evaluate if the Results are Good?

We use two ways to evaluate our results

1. Create a "placebo" treatment (randomized treatment assignment)–> calculate "benchmark" values of assessment measures
2. We exploit the fact that the NEWWS was an experiment
   - For three sites (ATL, GRP, RIV) individuals were randomly assigned to three treatments:
     - Control
     - Labor Force Attachment (LFA) training
     - Human Capital Development (HCD) training
   - Outcome: 1{Employment 2 yrs *before* RA}
   - Then, *within each site* we calculate "benchmark" assessment values

Placebo Experiment

Table 5. Benchmark values of the assessment measures for Raw Mean estimator from placebo experiments
Outcome: Employment rate in two years after random assignment

Outcome	P-value joint equality Wald test	Distance measures		
		Root Mean Square Distance	Mean Absolute Distance	Maximum Distance
A. 5 sites				
Levels	0.436	0.020	0.020	0.048
		[0.013,0.043]	[0.011,0.037]	[0.035,0.122]
DID	0.158	0.027	0.024	0.064
		[0.018,0.051]	[0.015,0.045]	[0.046,0.141]
B. 4 sites				
Levels	0.491	0.020	0.019	0.048
		[0.010,0.047]	[0.009,0.042]	[0.027,0.120]
DID	0.344	0.023	0.021	0.056
		[0.013,0.048]	[0.010,0.043]	[0.032,0.126]

Note: Bootstrap confidence intervals in brackets (based on 1,000 replications).

Pre-treatment Outcome for Experimental Group

Table 6. Benchmark values of the assessment measures for Raw Mean estimator from using within-site experimental treatment groups (3 treatments per site)
Outcome: Employment rate in two years *prior* to random assignment

Site	P-value joint equality Wald test	Distance measures		
		Root Mean Square Distance	Mean Absolute Distance	Maximum Distance
ATL	0.270	0.024	0.023	0.052
		[0.009,0.051]	[0.008,0.046]	[0.022,0.120]
GRP	0.250	0.024	0.021	0.057
		[0.009,0.051]	[0.008,0.045]	[0.021,0.120]
RIV	0.283	0.025	0.023	0.060
		[0.009,0.055]	[0.008,0.049]	[0.021,0.129]

Note: Bootstrap confidence intervals in brackets (based on 1,000 replications).

Robustness: Different Overlap Trimming Rules (Table 7)

Table 7. Assesment measures of estimators when applying different overlap trimming rules (quantile q) - 4 sites
Outcome: Employment Rate in Two Years after Random Assignment

	Overlap rule: q=0.000		Overlap rule: q=0.002		Overlap rule: q=0.005	
Estimator	**P-value jnt equality Wald test**	**Root Mean Square Distance**	**P-value jnt equality Wald test**	**Root Mean Square Distance**	**P-value jnt equality Wald test**	**Root Mean Square Distance**
A. Outcome in levels						
Linear regression-based						
Partial Mean Flex X - Ovlp	0.055	0.041 [0.025,0.072]	0.098	0.039 [0.024,0.072]	0.056	0.045 [0.023,0.072]
GPS-based (imposing Ovlp)						
IPW With Covariates	0.200	0.066 [0.031,0.125]	0.284	0.058 [0.028,0.113]	0.280	0.057 [0.024,0.108]
B. Outcome in differences (with respecto to years 1 and 2 before RA)						
Linear regression-based						
Partial Mean Flex X - Ovlp	0.131	0.032 [0.018,0.059]	0.249	0.028 [0.018,0.060]	0.127	0.036 [0.015,0.059]
GPS-based (imposing Ovlp)						
IPW With Covariates	0.991	0.012 [0.016,0.086]	0.894	0.030 [0.017,0.091]	0.708	0.041 [0.020,0.099]
Sample size after overlap	6,228		5,743		5,337	
Obs dropped due to overlap (%)	4.5%		12.0%		18.2%	

Notes: Results based on 1,000 bootstrap replications.

Conclusion

- Overlap condition stronger than for binary case
  –> harder to find comparable individuals for each treatment
- Crucial role of GPS in assessing comparability of treatment groups –> we propose a strategy for determining overlap region that is less stringent than previously used in literature
- GPS works well in balancing covariates (once site with poor overlap quality is removed)
- Estimators perform badly when there is poor overlap quality
- Estimators improve considerably with better overlap quality (and more similar LECs)
- DID estimators perform the best compared to benchmark measures based on experimental data
- Results very encouraging –> if satisfactory overlap quality
- Future work: simulations-based analysis