Impact of Roadside Mobile Monitoring Design on Epidemiologic Inference – A Case Study of Ultrafine Particles and Cognitive Function

Magali N. Blanco,[1] Adam A. Szpiro,[2] Paul K. Crane,[3] Lianne Sheppard[1,2]

[1]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA USA
[2]Department of Biostatistics, University of Washington, Seattle, WA USA
[3]Department of Medicine, University of Washington, Seattle, WA, USA

Contact Author:
Magali N. Blanco
magali@uw.edu
Dept of Environmental and Occupational Health Sciences
Box 351618
University of Washington
Seattle, WA 98195-1618

Conflicts of interest:
The authors declare they have no conflicts of interest related to this work to disclose.

Abstract

Background: Mobile monitoring campaigns are frequently used to develop air pollution exposure models to be used in health studies. Monitoring designs vary substantially, however, and it is unclear how design features impact exposure assessment models or health inferences.

Methods: We conducted a case study of the impact of mobile monitoring study design on ultrafine particle (UFP) exposure assessment and the estimated association between UFP and late-life cognitive function. We leveraged UFP measures from an extensive mobile monitoring campaign consisting of 309 temporary roadside stationary sites, each with ~29 temporally balanced visits over a year. We subsampled the data following common field designs: fewer visits per site (4-12); shorter campaign durations (1-4 seasons); business or rush hours (unadjusted and temporally adjusted); and an unbalanced number of visits where high variability sites received more or less visits than low variability sites. We developed annual average UFP exposure models with the resulting data and ran health analyses to estimate the adjusted association between five-year UFP exposure and baseline cognitive function (Cognitive Abilities Screening Instrument – Item Response Theory [CASI-IRT]) in the Adult Changes in Thought (ACT) cohort (N=5,409).

Results: The reference UFP all-data exposure model (R^2=0.65) estimated that the adjusted mean CASI-IRT was lower by 0.020 (95% CI: -0.036, -0.004) per each 1,900 pt/cm^3. More restricted designs generally produced poorer performing exposure models (median R^2: 0.40-0.61), with business hours (R^2: 0.40-0.45), one-season (R^2: 0.43), and unbalanced visits (R^2: 0.48) performing worst. Health inferences were broadly consistent with those from the all-data exposure model with just fewer visits per location, but they had more bias and/or were inconsistent across campaigns with fewer seasons, business or rush hours, or unbalanced visits. Business and rush hour designs had the most biased and attenuated health estimates.

Conclusions: Thoughtful monitoring design can improve exposure models and subsequent health inferences.

1 Introduction

Increasing evidence links traffic-related air pollution (TRAP), including ultrafine particles (UFP) to adverse health outcomes such as cognitive function (Brugge & Fuller, 2020; HEI, 2013; US EPA, 2019). UFPs are typically defined as particles ≤ 100 nm in diameter and measured as a particle number concentration (PNC). These are not routinely monitored by traditional long-term fixed site networks like other pollutants such as fine particulate matter ($PM_{2.5}$) or nitrogen dioxide (NO_2).

Mobile monitoring campaigns, the use of a mobile platform such as a vehicle to collect repeated short-term air samples from many locations, are commonly implemented to address this gap and capture the high spatial variability of UFPs (Kim et al., 2023). Over the past few decades, mobile monitoring campaigns have helped us gain important insights into pollutant sources and their spatiotemporal variability, pollution "hotspots," commuter exposures, and more (Apte et al., 2017; Austin et al., 2021; Karumanchi et al., 2021; Knibbs et al., 2011; Weichenthal et al., 2016). An increasing number of recent campaigns aim to develop long-term exposure assessment models to be used in subsequent epidemiologic inferences. Monitoring designs vary substantially, however, and no standard protocols exist. Most collect only a handful of measurements (visits) per site (median: ~4, range: ~1-40), last between a few weeks up to approximately 3 months, and sample exclusively during weekday business or rush hours (Kim et al., 2023, p. 202). Nearly all collect samples in an unbalanced fashion, with some locations receiving more visits than others for various reasons, including logistical constraints. Most campaigns produce poor to moderately performing exposure assessment models. Limited attention has been paid to the impact of monitoring design on exposure assessment models (Apte et al., 2017; Blanco, Doubleday, et al., 2022; Blanco et al., 2023; Blanco, Gassett, et al., 2022; Hatzopoulou et al., 2017; Messier et al., 2018; Saha et al., 2019), and no studies have assessed how these design features impact exposure assessment models or subsequent epidemiologic health inferences.

Many epidemiologic studies require long-term exposures, which is unique from other applications such as commuter or high exposure studies where shorter-term, on-road, and/or weekday business or rush hour exposures may be most relevant. We have previously shown that common, restricted sampling designs produce biased annual-average exposure predictions (Blanco, Doubleday, et al., 2022; Blanco et al., 2023). The objective of this study was to investigate the degree to which stationary (temporary roadside stop) monitoring design choices impact subsequent epidemiologic inferences. We conduct a case study of UFP exposures and late-life cognitive function by leveraging an extensive mobile monitoring campaign that was specifically designed to estimate unbiased annual average UFP exposures in the greater Seattle area (Blanco, Gassett, et al., 2022) and the Adult Changes in Thought (ACT) cohort, a large prospective cohort study investigating the aging brain (Kukull et al., 2002). We follow common mobile monitoring designs to sample our rich mobile monitoring dataset; develop design-specific PNC exposure assessment models; use these to assess participant exposures; and fit health models to estimate the association between PNC exposure and cognitive function in ACT. We evaluate these results to provide guidance on mobile monitoring study design features

that should be prioritized if the goal is to develop exposure assessment models for epidemiologic applications.

2 Methods

2.1 Cohort and Cognitive Assessments

ACT is a community-based, prospective cohort study in the greater Seattle area that has been investigating the aging brain since 1994 (Kukull et al., 2002). The study randomly invites elderly (65+ yr) members of the Kaiser Permanente Washington integrated healthcare delivery system (formerly Group Health Cooperative) to participate. Invitees are assessed for cognitive function at baseline using the Cognitive Abilities Screening Instrument (CASI), which combines common screening tests including the Mini-Mental State Examination (MMSE) and the Hasegawa Dementia Rating Scale to quantitatively assess attention, concentration, orientation, short- and long-term memory, language abilities, judgement, and other functions (Teng et al., 2004). People with high cognitive scores (scores of $\geq$ 86/100) are enrolled. People with low scores are evaluated with a comprehensive neuropsychological battery and focused neurological examination. Results of those assessments and medical records including imaging are reviewed at a consensus conference to identify cases of dementia and Alzheimer's disease using standardized research criteria. People who do not have dementia from this consensus process are also invited to enroll in the study. The final cognition scores from the CASI are derived using Item Response Theory (CASI-IRT), to improve score accuracy, measure cognitive change with less bias, and to account for missing test items (Crane et al., 2008; Ehlenbach et al., 2010; Li et al., 2017). Participants are prospectively followed until dementia incidence, drop-out, or death. Extensive health, lifestyle, biological, and demographic data are also collected.

As of March of 2020, the total ACT enrollment consisted of 5,763 participants. This analysis was restricted to ACT study baseline and included 5,409 (94%) participants with a valid CASI-IRT score and those who had lived in the exposure monitoring region (see below) during at least 95% of the prior five years (SI Figure S1 details participant retention). The ACT repository has excellent residential histories and air pollution coverage for ACT participants (Blanco, Gassett, et al., 2022; Shaffer et al., 2021). On average, this analytic cohort lived in the monitoring area >99% of the time, had exact geocoded residential addresses 98% of the time (e.g., vs. street level geocoding), and had imputed addresses 5% of the time (i.e., from residential gaps).

Study procedures were approved by the University of Washington and Kaiser Permanente institutional review boards. ACT participants signed informed consent forms.

2.2 Exposure Assessment from Mobile Monitoring Campaigns

We leveraged an extensive and novel mobile monitoring campaign that was designed to assess unbiased annual average TRAP exposures for ACT (Blanco, Doubleday, et al., 2022; Blanco, Gassett, et al., 2022). Monitoring was conducted within a 1,200 land km^2 area in the greater Seattle area and consisted of repeated 2-minute measurements at 309 roadside locations that were representative of ACT residential locations. Approximately 29 (IQR: 29-29, range: 26-35) measurements were collected from each location over the course of a year between March 2019 and March 2020 across all four seasons, all days of the week, and most

hours of the day (5 AM – 11 PM). Median pollutant concentrations were estimated for each site visit. These were winsorized at the site level such that values below the 5th and above the 95th quantile were set to those thresholds to reduce the influence of extreme observations. These data were used to develop what we refer in this study as "all-data" site annual averages which were treated as gold standard reference estimates, as described below. The campaign measured UFPs using multiple instruments. In this study, we use UFP measures from the TSI NanoScan 3910, which measured total and size-specific PNC for 10-420 nm particles, and the TSI P-TRAK 8525, which measured total PNC for 20-1,000 nm particles. We considered total PNC from the NanoScan our primary measure since it measured smaller particles than the P-TRAK and was more consistent with the World Health Organization's suggested air quality guidelines of measuring particles down to at least 10 nm (WHO, 2021). In sensitivity analyses, we specifically looked at 10-100 nm PNC from the NanoScan since UFPs are commonly defined as ≤ 100 nm, and at 20-1,000 nm PNC from the P-TRAK, a common UFP monitoring instrument.

We subsampled the all-data campaign (309 roadside locations x ~29 visits each) with replacement following four common restricted sampling designs (30 campaigns each) (Table 1). In the first design, we sampled fewer visits per site (n=4, 6, and 12) with no additional temporal restrictions. In the second design, we restricted sampling to fewer (1-3) seasons and collected 12 visits from each site balanced across sampling seasons (e.g., 4 samples per season for a three-season campaign).

In the third design, we sampled fewer visits per site (n=12) during weekday business (9 AM - 5 PM) or rush (7-10 AM & 3-6 PM) hours. The fewer visit design with 12 visits per site was a reference for these designs – it collected the same number of visits per site without temporal restrictions. We used business and rush hour visit samples as is (unadjusted) and temporally-adjusted – a common approach for addressing known biases resulting from restricted sampling campaigns that do not sample during the full exposure period of interest (e.g., weekends, night time, when the goal is to estimate an annual average) (Eeftens et al., 2012; Klompmaker et al., 2015; Montagne et al., 2015; van de Beek et al., 2021; van Nunen et al., 2017). This approach generally entails using an air monitoring site with continuous monitoring (typically a "background" or low-concentration site); calculating time-specific adjustment factors, based most commonly on the difference between a time-specific (e.g., hourly) measurement and the site's long-term average; and applying these adjustment factors to the measured concentrations. Our approach approximating this general strategy is detailed in Note S1. In summary, our temporal approach consisted of: 1) simulating a long-term UFP monitoring at an urban background site (Beacon Hill; continuous measures were unavailable for the entire mobile monitoring study period) from periodic PNC measures, collocated NO_2 measures, and temporal indicators; 2) generating adjustment factors, defined as the difference between the predicted hourly PNC and the long-term average PNC at Beacon Hill; and 3) applying these adjustment factors to the mobile monitoring data collected under business and rush hours designs.

In our fourth design, we evaluated a strategy characterized by unbalanced visits, a practice employed by nearly all field campaigns. We sampled based on predicted site variability, defined by partial least squares (PLS) regression where we regressed site-specific PNC

interquartile range (IQR; median [range]: 7,183 [2,834-22,625] pt/cm^3 based on ~29 visits per site) against the first two PLS components summarizing hundreds of geographic covariate predictors (see below for example covariates). The in-sample model R^2 was 0.46. We used this model to predict in-sample site-specific IQR and ordered these such that 129 (42%) sites were treated as medium variability sites, and visits continued to be fixed to 12. The remaining sites were split into high (H) or low (Low) variability (n=90 [29%] each). Figure S5 shows the distribution of IQRs used for variability group. We incorporated more visits for high-variability sites (14 to 22 visits) and fewer visits for low-variability sites (10 to 2), and vice versa.

The same sampling campaigns (i.e., exact visit samples) were used for sensitivity analyses of 10-100 nm and 20-1,000 nm particles (vs 10-420 nm) for all designs other than the business and rush hour designs, where a different set of 30 campaigns were randomly sampled. This should not be a source of bias since all campaigns were randomly selected.

We calculated annual average site concentrations from each sampling campaign. In total, there were 480 candidate sampling campaigns and subsequent exposure models for our primary analysis using 10-420 nm PNC from the NanoScan in addition to the all-data exposure model.

Table 1.Reduced Sampling Designs from an extensive, "all-data" mobile monitoring campaign (309 roadside sites).[a]

Design[a]	**Versions**	**No. of Versions**	**Total Visits**	**Visits per Site**	**Campaign Repetitions**
All-data	All-data	1	8,969	29[b]	1
Fewer Visits (no temporal restrictions)	4, 6, 12 visits per site	3	1,236, 1,854, 3,708	4, 6, 12	30
Fewer Seasons[c]	1-4 seasons	4	3,708	12	30
Fewer Hours	Weekday business or rush hours, unadjusted or temporally-adjusted	4	3,708	12	30
Unbalanced Visits	High (H) and low (L) variability sites receive the following visits: H2 L22, H6 L18, H12 L12 (all receive 12 visits), H18 L6, H22 L2	5	3,708	2-22 (avg: 12)	30

[a] The all-data design is a reference for all other designs, which have fewer site visits.
[b] mean and median: 29; IQR: 29-29; range: 26-35
[c] Samples were distributed evenly across the randomly selected seasons (e.g., 12 site visits/3 seasons = 4 site visits/season).

We used the annual average site PNCs from each sampling campaign to develop universal kriging – partial least squares (UK-PLS) exposure prediction models. PNC was log-transformed and regressed against the first two PLS components, which summarized 188 geographic

covariates predictive of TRAP (e.g., land use, roadway proximity, population density), as previously detailed (Blanco et al., 2023; Blanco, Gassett, et al., 2022). We evaluated each model by comparing the five-fold cross-validated site predictions against the annual averages from the all-data campaign (our best estimates). We and others have shown the importance of validating model predictions against unbiased estimates, and how comparisons against biased, unstable campaign measurements (e.g., from restricted sampling designs) produces noisy and misleading conclusions (Blanco, Doubleday, et al., 2022; Blanco et al., 2023; Kerckhoffs et al., 2016; Messier et al., 2018). We evaluated the performance of each model using mean-square error (MSE) -based R^2 (R^2_{MSE}), which evaluates whether pairs of predictions and observations are the same (i.e., along the one-to-one line) rather than simply linearly associated, like traditional regression-based R^2, and is thus better suited to evaluate predictive performance.

We used each campaign model to predict time-weighted average PNC exposures for each participant at baseline based on their prior five-year residential history.

2.3 Inferential Analyses

We assessed the association between the five-year average PNC exposure prior to baseline using each exposure model and baseline cognitive function (CASI-IRT) using linear regression. Each model was adjusted for age, calendar year (2 yr categories), sex, and education (no degree, high school equivalent, bachelor's, master's, doctorate, other). The model was:

$$CASI\ IRT_i = \alpha + \beta_{1,m} X_{i,m}^{PNC} + \beta_2 X_i^{age} + \beta_3 I_i^{year} + \beta_4 I_i^{sex} + \beta_5 I_i^{edu} + \varepsilon_i \quad (1)$$

Where the *i* index denotes participant *i* and *m* the PNC exposure prediction from a given exposure model *m*. We compared the health effects parameter ($\hat{\beta}_{1,m}$) estimated from each PNC exposure model to the health effect estimated from the all-data campaign.

All analyses were conducted in R (v. 4.2.2) (R Core Team, 2023).

3 Results

3.1 Cohort Characteristics

Table 2 describes the baseline analytic cohort characteristics. On average (SD), participants were 74 (6) years old, slightly more were female, about half had at least a college education, and an average (SD) CASI-IRT score of 0.33 (0.71).

Table 2. Baseline cohort characteristics.[1]

	Low PNC (N=1785)	Medium PNC (N=1785)	High PNC (N=1839)	Overall (N=5409)
Visit Age (Years)				
Mean (SD)	73.6 (6.03)	74.0 (6.40)	74.4 (6.48)	74.0 (6.31)
Median [Min, Max]	72.0 [65.0, 98.0]	73.0 [65.0, 96.0]	73.0 [65.0, 101]	73.0 [65.0, 101]
Sex				
Male	767 (43.0%)	742 (41.6%)	750 (40.8%)	2259 (41.8%)
Female	1018 (57.0%)	1043 (58.4%)	1089 (59.2%)	3150 (58.2%)
Degree				
None	128 (7.2%)	136 (7.6%)	158 (8.6%)	422 (7.8%)
GED/High School	657 (36.8%)	607 (34.0%)	733 (39.9%)	1997 (36.9%)
Bachelor's	423 (23.7%)	443 (24.8%)	408 (22.2%)	1274 (23.6%)
Master's	288 (16.1%)	319 (17.9%)	265 (14.4%)	872 (16.1%)
Doctorate	110 (6.2%)	122 (6.8%)	98 (5.3%)	330 (6.1%)
Other	179 (10.0%)	158 (8.9%)	177 (9.6%)	514 (9.5%)
CASI-IRT				
Mean (SD)	0.368 (0.689)	0.365 (0.720)	0.277 (0.710)	0.337 (0.708)
Median [Min, Max]	0.408 [-1.96, 1.75]	0.398 [-1.98, 1.75]	0.304 [-2.12, 1.75]	0.371 [-2.12, 1.75]
Residential PNC (pt/cm3) Exposure				
Mean (SD)	8,760 (647)	10,100 (311)	12,500 (2,080)	10,500 (2,020)
Median [Min, Max]	8,890 [5,930, 9,570]	10,100 [9,570, 10,700]	11,700 [10,700, 22,100]	10,100 [5,930, 22,100]

[1]Low, medium, and high PNC tertile is based on the predicted PNC from the all-data exposure model.

3.2 Exposure Assessment and Model Performances

The median (interquartile range [IQR]) site PNC for the primary analysis (10-420 nm) from the all-data campaign was 9,747 (8,412-11,199) pt/cm^3 (Figure S6). Sampling designs had similar but slightly more variable annual average site estimates. Sensitivity analyses resulted in lower site concentration estimates; 20-1,000 nm PNC from the P-TRAK had the smallest concentrations (Figure S6).

The all-data campaign PNC exposure models had a cross-validated R^2_{MSE} value of 0.65 (Figure 1). Almost all sampling designs with restricted sampling had lower performing exposure models. Performances were incrementally worse for campaigns with fewer visits per site; shorter campaign durations; more restricted sampling days and times (business and rush hours). Performance was even worse when temporal adjustments were applied (see Figure S7

for paired comparisons of adjusted and unadjusted campaign model performances); and when there were an unbalanced number of visits sampled across sites, particularly when high variability sites had very few visits. Despite collecting the same number of total visits (309 sites x 12 visits), campaigns with one season duration, those conducted during business hours (adjusted and unadjusted), and those with few visits to high variability sites (even when more visits were collected from lower variability sites) performed worse than otherwise unrestricted 12 visit designs. Sensitivity analyses for 10-100 nm and 20-1,000 nm PNC showed similar patterns (Figure S9).

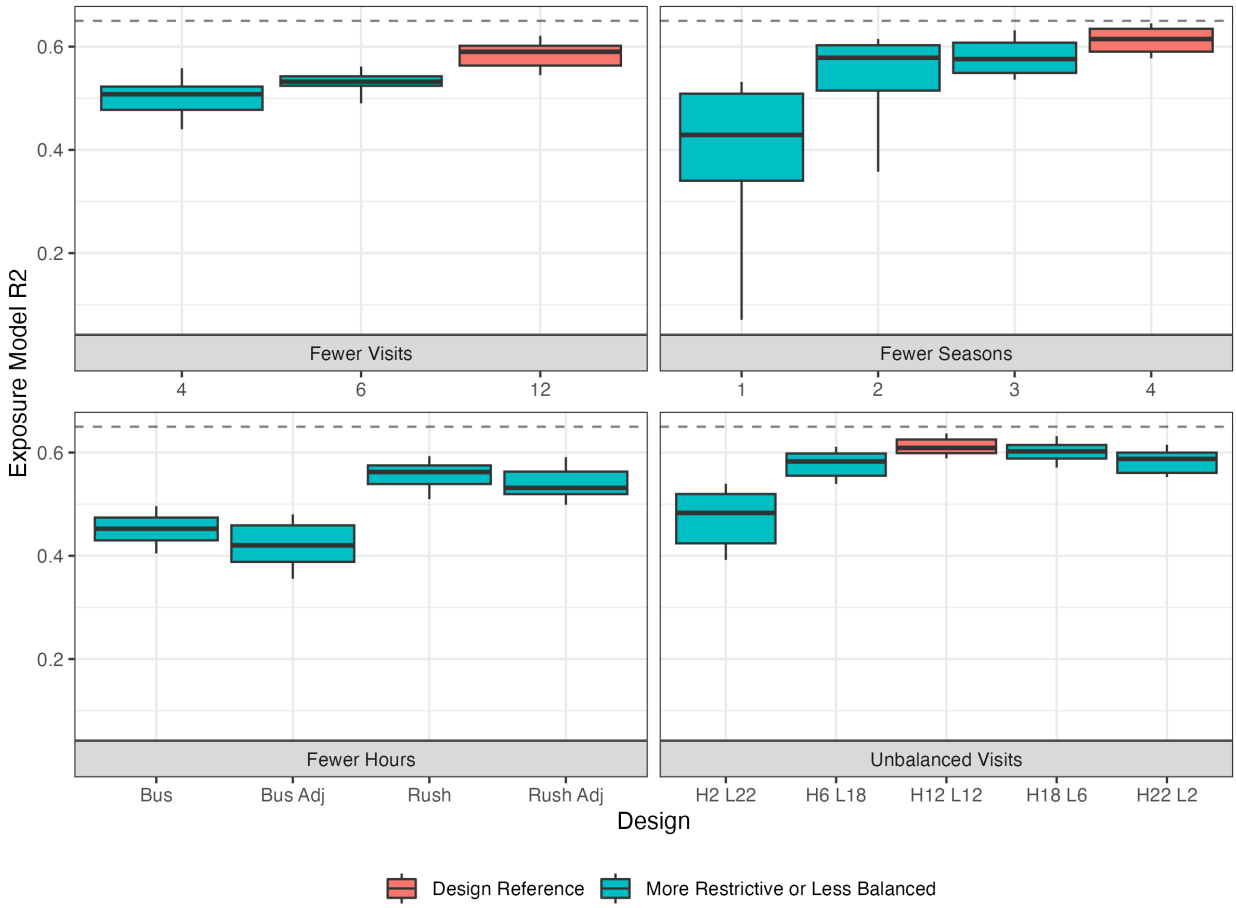


Figure 1. Cross-validated UFP model performances (N=30 campaigns per design). The dashed lines indidcate the all-data campaign performance. Red design reference boxplots indicate the least restrictive or most balanced campaigns; any of these can serve as a reference for the business and rush hours designs. Models are for total 10-420 nm PNC (pt/cm3) from the NanoScan instrument. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.

The median (IQR) predicted PNC for participants was 10,124 (9,293-11,100) pt/cm^3 and ranged from 5,930-22,134. Exposure predictions for sampling designs varied across campaigns (Figure S10). The business hour design tended to underpredict high exposures relative to the all-data exposure model, while the rush hour design overpredicted high exposures. Designs with few visits to high variability sites (and more visits to low variability sites, H2 L22) had highly

variable predictions across campaigns, particularly for high concentrations. We saw similar patterns in sensitivity analyses of 10-100 nm and 20-1,000 nm PNC.

Predictions from most designs were highly correlated with predictions from the all-data campaign (median Pearson correlations [R] > 0.85), although the business hour design was consistently lower than all other designs (R ~0.77-0.78; Figure S11). Lower correlations indicate differences in exposure surfaces (predictions) for mobile monitoring designs with fewer visits per site, those with shorter campaign durations, those limited to business hours, and those with fewer visits at high-variability sites. All designs had one or more atypical campaigns (i.e., outliers in the Figure S11 boxplots) that had a meaningfully lower correlation with the all-data campaign than the majority of other similarly designed campaigns, indicating potentially meaningful variability and lower exposure model performances across campaign iterations.

3.3 Inferential Analyses

Using the all-data campaign exposure model, the adjusted mean baseline CASI-IRT score was lower by -0.020 (95% confidence interval [CI]: -0.036, -0.004) for every increment of 1,900 pt/cm^3. Figure 2 summarizes the health effect point estimates across sampling campaigns and their percent difference relative to the health effect estimate obtained from the all-data exposure model. The health effect estimates for the fewer visit and season designs are similar, with campaigns with more visits and longer durations being most similar and associated with more consistent (less variable) results across campaigns. The fewer visit design with 4 visits per site and 1 season designs have the highest variability in the estimated health estimates across campaigns, indicating less consistent results. Business and rush hour designs, on the other hand, produce biased, attenuated health effect estimates that are about 50% and 40% different from the all-data estimate, respectively. Temporally adjusting these designs was associated with slightly more accurate health inferences. Campaigns with balanced designs where all sites receive the same number of visits (12) had health effects inferences that are the most consistent with the all-data exposure model. Figure S12 shows similar results for sensitivity analyses, with 20-1,000 nm PNC P-TRAK exposure models showing greater differences closer to 60% for business hour designs. Figure S13 further details the point and 95% CI for selected campaigns for primary and sensitivity analyses.

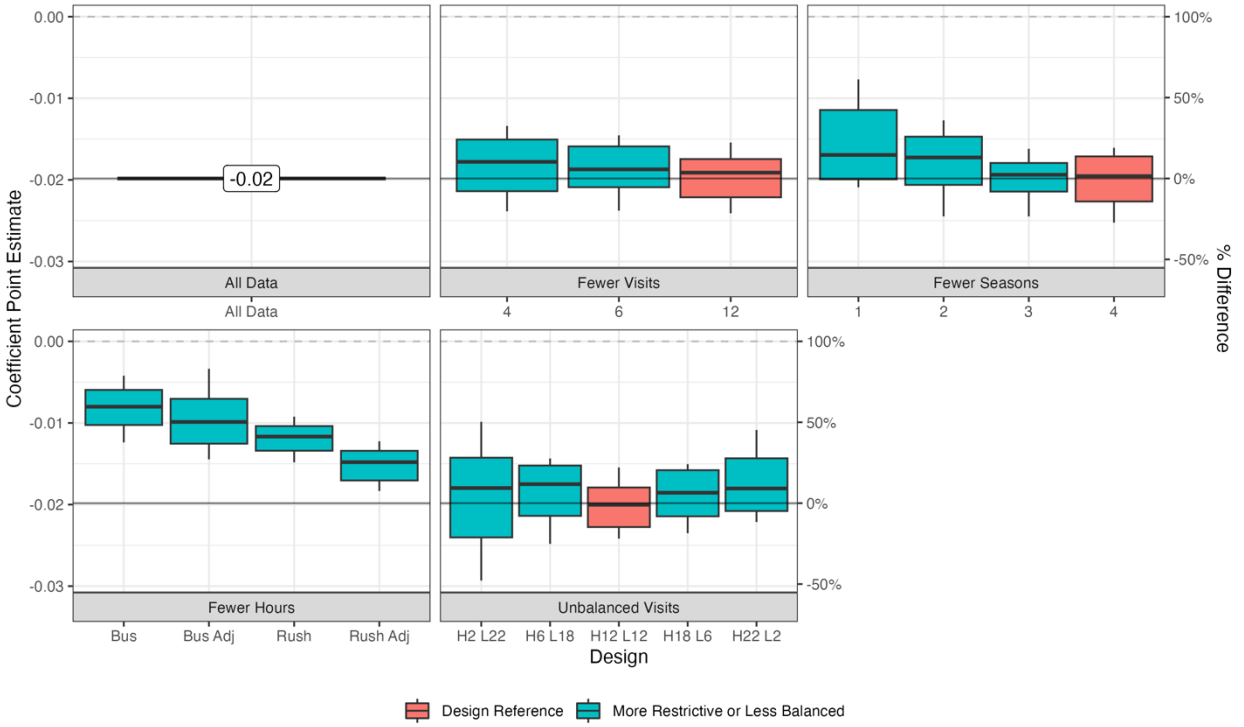


Figure 2. Health effect estimates from different sampling designs for the adjusted association between PNC (1,900 pt/cm^3) and cognitive function (CASI-IRT). Percent difference is relative to the health effect estimate from the all-data exposure model. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.

4 Discussion

Mobile monitoring campaigns to assess traffic pollutants, including UFPs, are being used around the globe to address monitoring gaps (Kim et al., 2023). Many campaigns now aim to develop exposure assessment models to be used in epidemiologic applications. This application generally necessitates capturing long-term, offroad (generally residential) exposures and is different from commuter exposures or hotspot identification studies, among others. Still, guidance on mobile monitoring study design for epidemiologic applications has been largely absent from the literature (Blanco, Doubleday, et al., 2022; Blanco et al., 2023; Blanco, Gassett, et al., 2022; Doubleday et al., 2023). As a result, there is substantial variability in how mobile monitoring campaigns are designed and implemented. We previously showed that monitoring design features like the number of sites, visits, campaign duration, sampling days, and sampling hours can greatly impact the predictive performance of exposure assessment models (Blanco, Doubleday, et al., 2022; Blanco et al., 2023). Here, we further assess additional monitoring approaches and how these differences in exposure model choices impact epidemiologic inferences.

We found that, when compared to an extensive mobile monitoring campaign intentionally designed for epidemiologic application (the all-data exposure model), campaigns with fewer visits (~4-12) but no temporal restrictions, shorter durations (~2-3 seasons), and a fixed number of visits across sites (12 in this case) had only slightly worse exposure model performances

(Figure 1); similar, highly correlated (mostly >~0.85) participant exposure assessments (Figure S11); and health inferences with only a small degree of bias (Figure 2). As expected, shorter campaigns (e.g., one season) and those with fewer repeat site visits generally had worse performing models with predictions that were less correlated (i.e., more different) to those from the all-data campaign. Rush hour and especially business hour designs, on the other hand, had much worse exposure model performances, more variable participant exposure assessments, and more biased, attenuated health effect estimates despite predicting exposures that were moderately (business) to highly (rush) correlated with those from the all-data exposure model. Moreover, the health effect estimates associated with these monitoring campaigns were noticeably different from all other designs, including much shorter (e.g., 1 season vs year-around) campaigns and those with fewer samples (e.g., 4 vs 12 visits per site), suggesting that capturing temporal variability through extended hours designs (e.g., sampling weekends and extending the sampling hours) is critical for capturing long-term annual average exposures. It's notable that these reduced day and hour campaigns are most common in the field since an operator is required to operate a vehicle and monitor instrumentation throughout the sampling period.

Interestingly, temporally adjusted business and rush hour designs were associated with worse exposure model performances (Figure 1, Figure S7, Figure S8) and had more biased health inferences than other designs, although temporal adjustment was able to reduce these health biases (Figure 2). Sites within a region can have different temporal patterns (Blanco, Doubleday, et al., 2022) related to major nearby sources (or their absence), for example, airports, highways, or industrial sites. Applying temporal adjustments from a single site may incorrectly or insufficiently adjust exposure estimates, depending on the heterogeneity of the monitoring sites. In our study, temporal adjustment did not improve predictions of overall exposure levels, but it was associated with less bias in health effect inference. Heuristically, the added complexity of the time adjustment introduced a form of classical-like measurement error that adversely impacted prediction accuracy, but the improved temporal alignment decreased the impact of Berkson-like error, which was responsible for the dominant health effect estimation bias from the unadjusted exposure estimates (Szpiro et al., 2011; Szpiro & Paciorek, 2013).

A feature of our temporal adjustment approach was that we based it on a simulated UFP monitoring site, as described in the Methods, from collocated PNC and highly temporally correlated NO_2 observations along with other temporal indicators. This approach was associated with good PNC predictions and captured much of the temporal variation in UFP (Figure S2 – Figure S4), suggesting it was a reliable source for estimating temporal adjustment factors. In the literature, various adjustment approaches (e.g., “difference” or “ratio” approaches) and sites have been used to temporally adjust mobile monitoring readings. These approaches have not been validated with data and themselves produce fluctuating adjustment factors. Finally, we used hourly adjustment factors to adjust noisier two-minute mobile monitoring site visits. We have previously shown that these are highly correlated (Blanco, Doubleday, et al., 2022).

There are different ways field sampling is conducted that lead to unbalanced sampling whereby some locations receive more visits than others. This may result from having non-fixed driving routes, logistical constraints that make it challenging to visit some sites while others along common driving routes are naturally oversampled, intentionally oversampling sites anticipated to have high variability while deprioritizing sites with low variability (e.g., suburban areas), etc. We present one approach whereby sampling is influenced by the anticipated site concentration variability across time. We did not see an appreciable benefit to exposure model performance from oversampling high variability sites although performances was lower when we dramatically undersampled high variability sites (Figure 1, H2 L22). Schemes with balanced samples across sites (12 visits each) had the least biased and least variable health inferences (Figure 2). These findings suggest using a balanced sampling design whenever feasible. If traveling to sites with low anticipated variability presents a significant logistical challenge, however, our results suggested that strategies characterized by somewhat fewer visits to these sites may be a reasonable choice.

We used predicted, in-sample IQR based on PLS regression analysis rather than "true" IQR to classify sites. This adds some error to site classifications (high, medium, low variability) despite being in-sample predictions (which can produce overfitted, optimistic results). Nonetheless, true site concentrations and variability are largely unknown prior to conducting in-field mobile monitoring, adding natural uncertainty to monitoring decisions. Moreover, defining target sites becomes more challenging for multiple pollutants since spatial and temporal patterns (e.g., pollutant variability) may vary across pollutants, such that a site could have high variability for one pollutant and low variability for another.

Overall, our findings suggest that strategic monitoring design can be implemented to optimize the likely accuracy of health inferences and the anticipated consistency of these results across campaigns (i.e., generally narrower boxplots in Figure 2) while keeping in mind the logistical constraints unique to mobile monitoring. We suggest prioritizing sampling during extended times beyond rush hours and business hours. Designs that focused on rush hour and especially on business hour monitoring were associated with attenuated health inferences that are only minimally made better with commonly used temporal adjustment approaches. Beyond that, collecting data over at least three seasons if the goal is to estimate an annual average, collecting a balanced (fixed) number of visits across locations, and collecting a higher number of visits per location were all associated with better health inference accuracy and lower variability across campaigns. One thing to note is that sampling design impacts some pollutants more than others (Blanco et al., 2023). This variability will likely translate to subsequent health inferences to varying degrees.

Spatial and temporal compatibility (i.e., similarity in distributions) between monitoring and cohort locations is an important feature for minimizing the impact of measurement error and consequently optimizing health inferences (Szpiro & Paciorek, 2013). Our study is inherently spatially aligned since our extensive mobile monitoring campaign was specifically designed to capture exposures for the ACT cohort (Blanco, Gassett, et al., 2022). It's notable that most campaigns select monitoring locations based on geographic features or sources (e.g., major roads, industry, airports) and do not explicitly set out to capture exposures based on the

geographical spread of study participants. Spatial compatibility is particularly relevant for air pollution epidemiology, where health effects of variable levels of exposure are associated with subtle health effects, and biases or lower levels of precision can easily obscure meaningful associations. In this case study of UFPs and cognitive function, we observed an association between our exposure and outcome of interest. While mobile monitoring inherently results in missing observations, the all-data and other similar designs (e.g., 3 seasons) estimated annual average exposure levels that are close to the true annual average (Blanco, Doubleday, et al., 2022). The day and time restricted designs, however, sample during times that are temporally misaligned with the longer-term exposures of interest. These designs contribute to bias from Berkson-like error, which is the difference between the true annual average exposure surface and the more limited part captured by the modeling process (Szpiro & Paciorek, 2013). The ideal way to eliminate the bias from temporal misalignment is by modifying the sampling design, but if this is not possible an alternative is to introduce a spatiotemporal model that fully captures the complexity of the underlying exposure surface. Our use of temporal adjustment can be viewed as a step in that direction, and as expected it did seem to reduce bias from Berkson-like error to some degree, but evidence of bias remains. Future research can explore the question of whether the available data are sufficiently rich to support a full spatiotemporal model that will more fully capture the underlying surface, and thus eliminate bias from Berkson-like error. Another possibility is to reweight the data to achieve temporal compatibility. However, it is not clear that either of these approaches will be successful with the rush-hour or business-hour designs since key information about what happens during the non-covered hours is completely unavailable.

More generally, our findings are conservative with respect to realized published studies. Many mobile monitoring campaigns incorporate multiple features that might limit their applicability to long-term population exposure studies, for example, campaigns that last less than a year, collect fewer repeat visits per site (median ~4), sample only during weekday business hours, and collect unbalanced numbers of visits per site (Kim et al., 2023). We anticipate that these designs will produce biased health effect estimates like those that we observed for the business hours design, if not more severe. Moreover, most mobile monitoring campaigns explicitly collect non-stationary, on-road data. While non-stationary designs achieve higher spatial coverage than stationary designs like the one used in this study, they measure on-road concentrations that are typically higher than those captured by stationary, offroad locations that are more similar to residential exposures; collect less data per location (seconds vs minutes) making for more unstable estimates; and the resulting exposure models are associated with poorer performance (Doubleday et al., 2023; Kim et al., 2023). Most on-road campaigns also do not adjust on-road data to minimize the influence of air pollution plumes (spike concentrations) common on roads but less so at residential locations. These design features are unique to on-road mobile monitoring, and their potential impact on health inferences should be investigated in future work.

We conducted this study using data from the long-standing, community-based ACT cohort. While we could have simulated health outcome data to conduct this analysis, we chose to use

this data source to reflect real-world impacts and incorporate aspects that might not be included in a simulation study. As such, this approach may be more illuminating of real world implications compared to a simulation study. ACT has consistently collected measures over time, including cognitive function, demographics, and lifestyle factors. ACT's extensive participant residential histories allowed us to assess UFP exposures for most participants. Since we used fixed annual average 2019 UFP exposure surfaces to assess exposures, there is inherent exposure assessment error in these analyses, including the all-data campaign, and this likely was higher for earlier time periods. Historical UFP data are rare and we assumed that the exposure surface was constant over time (Blanco, 2021; Kim et al., 2017; Levy et al., 2015; Meng et al., 2019; Molter et al., 2010; Wang et al., 2011). More generally, our inferential models in this analysis were not necessarily meant to characterize causal associations between UFP and cognitive function. Such analysis could consider more extensive confounding adjustment, and address potential selection biases that may have resulted, for example, from conducting complete case analyses. The goal of this analysis was to characterize how mobile monitoring design choices may impact estimated health effects of air pollution.

We investigate how mobile monitoring design impacts both air pollution exposure assessment and subsequent health outcomes and show that thoughtful monitoring design can be implemented to improve the accuracy of health inferences and consistency across campaigns. Critically, we recommend extending sampling beyond typical weekday business or rush hours. Health inferences can be further improved by collecting data over at least three seasons if the goal is to estimate an annual average, collecting a balanced (fixed) number of visits across locations, and collecting an increased number of visits. Our future work will investigate how monitoring design more specifically impacts non-stationary, on-road data exposure models and health inferences.

5 Acknowledgements

Research described in this article was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award No. CR-83998101) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers. This research was also supported by NIA/NIEHS R01ES026187 to LS. We thank the ACT participants for the data they have provided and the many ACT investigators and staff who steward that data (NIA U19AG066567). You can learn more about ACT at: https://actagingstudy.org/. All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the NIA or the NIH.

6 References

Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C. H., & Hamburg, S. P. (2017). High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. *Environmental Science & Technology*, *51*, 6999–7008.

Austin, E., Xiang, J., Gould, T. R., Shirai, J. H., Yun, S., Yost, M. G., Larson, T. V., & Seto, E. (2021). Distinct Ultrafine Particle Profiles Associated with Aircraft and Roadway Traffic. *Environmental Science & Technology*, *55*(5), 2847–2858. https://doi.org/10.1021/acs.est.0c05933

Blanco, M. N. (2021). Traffic-Related Air Pollution and Dementia Incidence in a Seattle-Based, Prospective Cohort Study. Dissertation. *University of Washington*.

Blanco, M. N., Bi, J., Austin, E., Larson, T. V., Marshall, J. D., & Sheppard, L. (2023). Impact of Mobile Monitoring Network Design on Air Pollution Exposure Assessment Models. *Environmental Science & Technology*, *57*(1), 440–450. https://doi.org/10.1021/acs.est.2c05338

Blanco, M. N., Doubleday, A., Austin, E., Marshall, J. D., Seto, E., Larson, T. V., & Sheppard, L. (2022). Design and evaluation of short-term monitoring campaigns for long-term air pollution exposure assessment. *Journal of Exposure Science & Environmental Epidemiology*, *33*(2), 465–473. https://doi.org/10.1038/s41370-022-00470-5

Blanco, M. N., Gassett, A., Gould, T., Doubleday, A., Slager, D. L., Austin, E., Seto, E., Larson, T. V., Marshall, J. D., & Sheppard, L. (2022). Characterization of Annual Average Traffic-Related Air Pollution Concentrations in the Greater Seattle Area from a Year-Long

Mobile Monitoring Campaign. *Environmental Science & Technology*, *56*(16), 11460–11472. https://doi.org/10.1021/acs.est.2c01077

Brugge, D., & Fuller, C. H. (Eds.). (2020). *Ambient combustion ultrafine particles and health*. Nova Science Publishers.

Crane, P. K., Narasimhalu, K., Gibbons, L. E., Mungas, D. M., Haneuse, S., Larson, E. B., Kuller, L., Hall, K., & van Belle, G. (2008). Item response theory facilitated cocalibrating cognitive tests and reduced bias in estimated rates of decline. *Journal of Clinical Epidemiology*, *61*(10), 1018–1027. https://doi.org/10.1016/j.jclinepi.2007.11.011

Doubleday, A., Blanco, M. N., Austin, E., Marshall, J. D., Larson, T. V., & Sheppard, L. (2023). Characterizing Ultrafine Particle Mobile Monitoring Data for Epidemiology. *Environmental Science & Technology*, *57*(26), 9538–9547. https://doi.org/10.1021/acs.est.3c00800

Eeftens, M., Tsai, M. Y., Ampe, C., Anwander, B., Beelen, R., Bellander, T., Cesaroni, G., Cirach, M., Cyrys, J., de Hoogh, K., De Nazelle, A., de Vocht, F., Declercq, C., Dedele, A., Eriksen, K., Galassi, C., Gražulevičiene, R., Grivas, G., Heinrich, J., … Hoek, G. (2012). Spatial variation of PM2.5, PM10, PM2.5 absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO2—Results of the ESCAPE project. *Atmospheric Environment*, *62*(2012), 303–317. https://doi.org/10.1016/j.atmosenv.2012.08.038

Ehlenbach, W. J., Hough, C. L., Crane, P. K., Haneuse, S. J. P. A., Carson, S. S., Curtis, J. R., & Larson, E. B. (2010). Association between acute care and critical illness hospitalization

and cognitive function in older adults. *JAMA*, *303*(8), 763–770. https://doi.org/10.1001/jama.2010.167

Hatzopoulou, M., Valois, M. F., Levy, I., Mihele, C., Lu, G., Bagg, S., Minet, L., & Brook, J. (2017). Robustness of Land-Use Regression Models Developed from Mobile Air Pollutant Measurements. *Environmental Science and Technology*, *51*(7), 3938–3947. https://doi.org/10.1021/acs.est.7b00366

HEI. (2013). *Understanding the Health Effects of Ambient Ultrafine Particles. Perspectives 3.* https://www.healtheffects.org/publication/understanding-health-effects-ambient-ultrafine-particles

Karumanchi, S., Siemiatycki, J., Richardson, L., Hatzopoulou, M., & Lequy, E. (2021). Spatial and temporal variability of airborne ultrafine particles in the Greater Montreal area: Results of monitoring campaigns in two seasons. *Science of The Total Environment*, *771*(2021), 144652. https://doi.org/10.1016/j.scitotenv.2020.144652

Kerckhoffs, J., Hoek, G., Messier, K. P., Brunekreef, B., Meliefste, K., Klompmaker, J. O., & Vermeulen, R. (2016). Comparison of Ultrafine Particle and Black Carbon Concentration Predictions from a Mobile and Short-Term Stationary Land-Use Regression Model. *Environmental Science & Technology*, *50*(23), 12894–12902. https://doi.org/10.1021/acs.est.6b03476

Kim, S.-Y., Blanco, M. N., Bi, J., Larson, T. V., & Sheppard, L. (2023). Exposure assessment for air pollution epidemiology: A scoping review of emerging monitoring platforms and designs. *Environmental Research*, *223*(2023), 115451. https://doi.org/10.1016/j.envres.2023.115451

Kim, S.-Y., Olives, C., Sheppard, L., Sampson, P. D., Larson, T. V., Keller, J. P., & Kaufman, J. D. (2017). Historical prediction modeling approach for estimating long-term concentrations of PM2.5 in cohort studies before the 1999 implementation of widespread monitoring. *Environmental Health Perspectives*, *125*(38–46). https://doi.org/10.1289/EHP131

Klompmaker, J. O., Montagne, D. R., Meliefste, K., Hoek, G., & Brunekreef, B. (2015). Spatial variation of ultrafine particles and black carbon in two cities: Results from a short-term measurement campaign. *Science of the Total Environment*, *508*(2015), 266–275. https://doi.org/10.1016/j.scitotenv.2014.11.088

Knibbs, L. D., Cole-Hunter, T., & Morawska, L. (2011). A review of commuter exposure to ultrafine particles and its health effects. *Atmospheric Environment*, *45*(16), 2611–2622. https://doi.org/10.1016/j.atmosenv.2011.02.065

Kukull, W. A., Higdon, R., Bowen, J. D., McCormick, W. C., Teri, L., Schellenberg, G. D., Van Belle, G., Jolley, L., & Larson, E. B. (2002). Dementia and Alzheimer disease incidence: A prospective cohort study. *Archives of Neurology*, *59*(11), 1737–1746. https://doi.org/10.1001/archneur.59.11.1737

Levy, I., Levin, N., Yuval, Y., Schwartz, J. D., & Kark, J. D. (2015). Back-extrapolating a land use regression model for estimating past exposures to traffic-related air pollution. *Environmental Science and Technology*, *49*(6), 3603–3610. https://doi.org/10.1021/es505707e

Li, G., Larson, E. B., Shofer, J. B., Crane, P. K., Gibbons, L. E., McCormick, W., Bowen, J. D., & Thompson, M. L. (2017). Cognitive Trajectory Changes Over 20 Years Before Dementia

579 Diagnosis: A Large Cohort Study. *Journal of the American Geriatrics Society*, *65*(12),
580 2627–2633. https://doi.org/10.1111/jgs.15077
581 Meng, J., Li, C., Martin, R. V., van Donkelaar, A., Hystad, P., & Brauer, M. (2019). Estimated
582 Long-Term (1981–2016) Concentrations of Ambient Fine Particulate Matter across
583 North America from Chemical Transport Modeling, Satellite Remote Sensing, and
584 Ground-Based Measurements. *Environmental Science & Technology*, *53*(9), 5071–5079.
585 https://doi.org/10.1021/acs.est.8b06875
586 Messier, K. P., Chambliss, S. E., Gani, S., Alvarez, R., Brauer, M., Choi, J. J., Hamburg, S. P.,
587 Kerckhoffs, J., Lafranchi, B., Lunden, M. M., Marshall, J. D., Portier, C. J., Roy, A., Szpiro,
588 A. A., Vermeulen, R. C. H., & Apte, J. S. (2018). Mapping Air Pollution with Google Street
589 View Cars: Efficient Approaches with Mobile Monitoring and Land Use Regression.
590 *Environmental Science and Technology*, *52*(21), 12563–12572.
591 https://doi.org/10.1021/acs.est.8b03395
592 Molter, A., Lindley, S., de Vocht, F., Simpson, A., & Agius, R. (2010). Modelling air pollution for
593 epidemiologic research – Part II: Predicting temporal variation through land use
594 regression. *Science of The Total Environment*, *409*(1), 211–217.
595 https://doi.org/10.1016/j.scitotenv.2010.10.005
596 Montagne, D. R., Hoek, G., Klompmaker, J. O., Wang, M., Meliefste, K., & Brunekreef, B. (2015).
597 Land Use Regression Models for Ultrafine Particles and Black Carbon Based on Short-
598 Term Monitoring Predict Past Spatial Variation. *Environmental Science and Technology*,
599 *49*(14), 8712–8720. https://doi.org/10.1021/es505791g

R Core Team. (2023). *R-A Language and Environment for Statistical Computing*. https://www.R-project.org

Saha, P. K., Li, H. Z., Apte, J. S., Robinson, A. L., & Presto, A. A. (2019). Urban Ultrafine Particle Exposure Assessment with Land-Use Regression: Influence of Sampling Strategy. *Environmental Science and Technology*, *53*(13), 7326–7336. https://doi.org/10.1021/acs.est.9b02086

Shaffer, R. M., Blanco, M. N., Li, G., Adar, S. D., Carone, M., Szpiro, A. A., Kaufman, J. D., Larson, T. V., Larson, E. B., Crane, P. K., & Sheppard, L. (2021). Fine Particulate Matter and Dementia Incidence in the Adult Changes in Thought Study. *Environmental Health Perspectives*, *129*(8), 087001. https://doi.org/10.1289/EHP9018

Szpiro, A. A., & Paciorek, C. J. (2013). Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*, *24*(8), 501–517. https://doi.org/10.1002/env.2233

Szpiro, A. A., Sheppard, L., & Lumley, T. (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics*, *12*(4), 610–623. https://doi.org/10.1093/biostatistics/kxq083

Teng, E. L., Hasegawa, K., Homma, A., Imai, Y., Larson, E., Graves, A., Sugimoto, K., Yamaguchi, T., Sasaki, H., Chiu, D., & White, L. R. (2004). The Cognitive Abilities Screening Instrument (CASI): A Practical Test for Cross-Cultural Epidemiological Studies of Dementia. *International Psychogeriatrics*, *6*(1), 45–58. https://doi.org/10.1017/s1041610294001602

US EPA. (2019). *Integrated science assessment (ISA) for particulate matter*. US Environmental Protection Agency Washington, DC. https://cfpub.epa.gov/ncea/isa/recordisplay.cfm?deid=347534

van de Beek, E., Kerckhoffs, J., Hoek, G., Sterk, G., Meliefste, K., Gehring, U., & Vermeulen, R. (2021). Spatial and Spatiotemporal Variability of Regional Background Ultrafine Particle Concentrations in the Netherlands. *Environmental Science & Technology*, *55*(2), 1067–1075. https://doi.org/10.1021/acs.est.0c06806

van Nunen, E., Vermeulen, R., Tsai, M.-Y., Probst-Hensch, N., Ineichen, A., Davey, M., Imboden, M., Ducret-Stich, R., Naccarati, A., & Raffaele, D. (2017). Land use regression models for ultrafine particles in six European areas. *Environmental Science & Technology*, *51*(6), 3336–3345.

Wang, Y., Hopke, P. K., Chalupa, D. C., & Utell, M. J. (2011). Long-term study of urban ultrafine particles and other pollutants. *Atmospheric Environment*, *45*(40), 7672–7680. https://doi.org/10.1016/j.atmosenv.2010.08.022

Weichenthal, S., Van Ryswyk, K., Goldstein, A., Shekarrizfard, M., & Hatzopoulou, M. (2016). Characterizing the spatial distribution of ambient ultrafine particles in Toronto, Canada: A land use regression model. *Environmental Pollution*, *208*(Pt A), 241–248. https://doi.org/10.1016/j.envpol.2015.04.011

WHO. (2021). *World Health Organization global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization. https://apps.who.int/iris/handle/10665/345329

Supplemental Material

Impact of Roadside Mobile Monitoring Design on Epidemiologic Inference – A Case Study of Ultrafine Particles and Cognitive Function

Magali N. Blanco,[1] Adam A. Szpiro,[2] Paul K. Crane,[3] Lianne Sheppard[1,2]

[1]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA USA
[2]Department of Biostatistics, University of Washington, Seattle, WA USA
[3]Department of Medicine, University of Washington, Seattle, WA, USA

Contact Author:
Magali N. Blanco
magali@uw.edu

Conflicts of interest:
The authors declare they have no conflicts of interest related to this work to disclose.

Table of Contents

1 METHODS 2

2 RESULTS 8

3 REFERENCES 18

List of Notes

NOTE S1. TEMPORAL ADJUSTMENT APPROACH 2

List of Figures

FIGURE S1. ANALYTIC COHORT USED IN THIS STUDY. 2

FIGURE S2. TIME SERIES OF HOURLY NO_2 (PPB) AND UFP (1,000 PT/CM3) AT BEACON HILL USED TO SIMULATE A CONTINUOUS, LONG-TERM PNC MONITORING SITE. 4

FIGURE S3. DISTRIBUTION OF HOURLY PNC LEVELS AT BEACON HILL, STRATIFIED BY DAY OF THE WEEK, HOUR OF THE DAY, AND SAMPLING MONTH. 5

FIGURE S4. TIME-SERIES OF SIMULATED PNC MODEL RESIDUALS. 6

FIGURE S5. DISTRIBUTION OF PNC IQR (PT/CM^3) FOR THE UNBALANCED VISITS DESIGN 7

FIGURE S6. DISTRIBUTION OF ESTIMATED ANNUAL AVERAGE SITE CONCENTRATIONS FOR THE ALL-DATA CAMPAIGN FOR PRIMARY AND SENSITIVITY ANALYSES. 8

FIGURE S7. COMPARISON OF UNADJUSTED AND TEMPORALLY ADJUSTED BUSINESS AND RUSH HOUR CAMPAIGN EXPOSURE MODELS 9

FIGURE S8. ABSOLUTE UFP PREDICTION ERRORS FROM THE BUSINESS AND RUSH HOUR DESIGNS FOR A SAMPLE OF 50 SITES 10

FIGURE S9.CROSS-VALIDATED MODEL PERFORMANCES 11

FIGURE S10. SMOOTH LINES COMPARING PREDICTED FIVE-YEAR AVERAGE PARTICIPANT PNC (PT/CM^3) EXPOSURE FROM THE ALL-DATA CAMPAIGN (N=1 CAMPAIGN) TO EXPOSURE PREDICTIONS FROM OTHER SAMPLING DESIGNS (N=30 CAMPAIGNS) FOR PRIMARY AND SENSITIVITY ANALYSES 13

FIGURE S11. PEARSON CORRELATIONS (R) COMPARING PREDICTED FIVE-YEAR AVERAGE PNC PARTICIPANT EXPOSURE FROM ALL-DATA CAMPAIGNS RELATIVE TO EACH SAMPLING DESIGN (N=30 CAMPAIGN CORRELATIONS PER DESIGN) FOR PRIMARY AND SENSITIVITY ANALYSES. 14

FIGURE S12. HEALTH EFFECT ESTIMATES PRODUCED FROM DIFFERENT SAMPLING DESIGNS FOR THE ADJUSTED ASSOCIATION BETWEEN PNC (1,900 PT/CM3) AND COGNITIVE FUNCTION (CASI-IRT) FOR PRIMARY AND SENSITIVITY ANALYSES 15

FIGURE S13. HEALTH EFFECT ESTIMATES FOR THE ASSOCIATION BETWEEN UFP (1,900 PT/CM3) AND CASI-IRT FOR SENSITIVITY ANALYSES AFTER ADJUSTING FOR AGE, CALENDAR YEAR, SEX, EDUCATION. 17

List of Tables

TABLE S1. CONTINUOUS OVERNIGHT SAMPLING PNC TIMES AT THE BEACON HILL MONITORING SITE 4

1 Methods

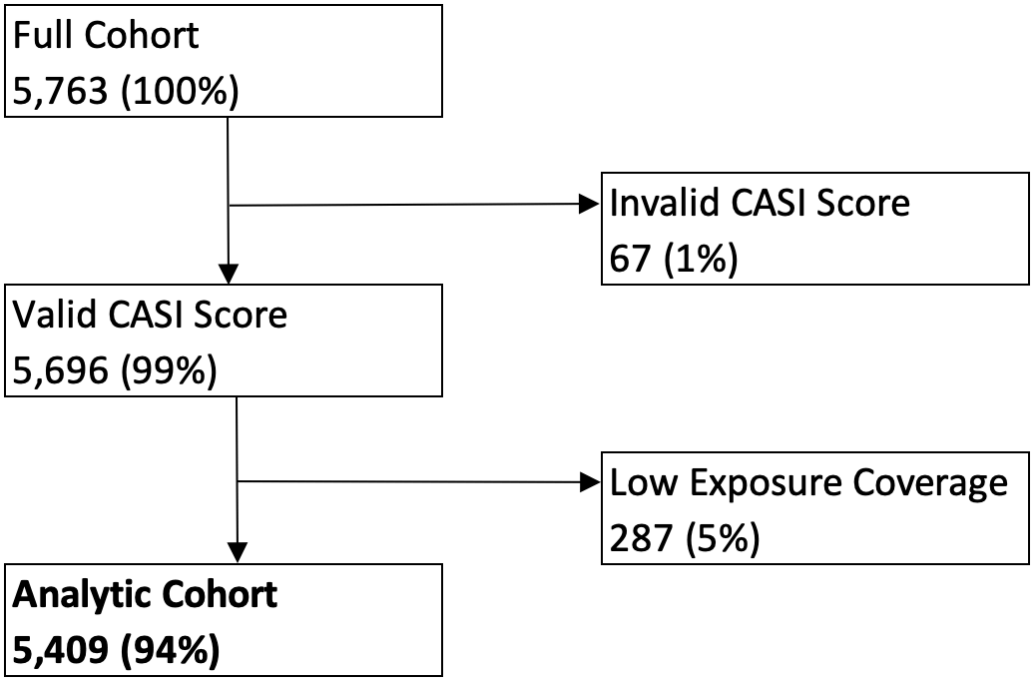


Figure S1. Analytic cohort used in this study.

Note S1. Temporal adjustment approach

In summary, our temporal approach consisted of the following:

1. Simulate a long-term UFP monitoring at an urban background site (Beacon Hill; continuous measures were unavailable for the entire mobile monitoring study period) from periodic PNC measures, collocated NO_2 measures, and temporal indicators.
2. Generate adjustment factors, defined as the difference between the predicted hourly PNC and the long-term average PNC at Beacon Hill.
3. Apply these adjustment factors to the mobile monitoring data collected under business and rush hours designs.

More specifically, we simulated one based on available continuous NO_2 measurements and collocated PNC measurements collected periodically throughout the study period at an urban background site in the study area (Beacon Hill) since a long-term UFP monitoring site was unavailable during the original mobile monitoring study period. Figure S2 depicts similar temporal trends between observed hourly PNC and NO_2 concentrations. Table S1 and Figure S3 summarize the available PNC measures (available from 31 sampling days across four months and three seasons). Hourly NO_2 observations were from the US EPA regulatory air network (US EPA, 2023). We imputed missing hourly NO_2 values (3%) based on a regression model fitting existing log-transformed NO_2 observations against an indicator for the observation month and a cubic cyclic spline for each day of the week (Mon-Sun) based on the hour associated with that observation (See Equation S1 for a similar layout). While most days with missing observations

had only one missing value, and we could have implemented simpler linear regression, we took this more flexible approach because 5 days had multiple (5+) missing values. We fit the following model to simulate a long-term PNC monitoring site:

$$\log(PNC_t) = \alpha + \beta \log(NO_{2,t}) + s(hour_t) * day\ of\ week_t + \epsilon_t \quad \text{(S1)}$$

where the log-transformed hourly average PNC at time *t* was regressed against the hourly average log-transformed NO_2 concentration at the same time *t* and a cubic cyclic spline for each day of the week (Mon-Sun) for the hour (0-23) associated with *t* (i.e., there are seven cyclic splines). The resulting model residuals had a minimal temporal variation, suggesting that the model captured important PNC temporal trends (Figure S4); and it had an in-sample model R^2 of 0.52.

We used this model to predict hourly PNC concentrations during the study period and simulate a long-term PNC monitoring site. We winsorized (i.e., set) extreme predictions above the 95th (16,380 pt/cm^3) and below the 5th quantile (2,963 pt/cm^3) to those quantiles, respectively, to reduce extreme temporal adjustments in the next step. We calculated hourly adjustment factors by taking the difference between the long-term (i.e., annual) and each hourly average PNC site concentration:

$$\hat{\delta}_t = PNC_{LTA} - \widehat{PNC}_{t,winsorized} \quad \text{(S2)}$$

Finally, PNC samples collected under the business or rush hours designs were adjusted:

$$\widehat{PNC}_{t,adj} = \widehat{PNC}_t + \hat{\delta}_t \quad \text{(S3)}$$

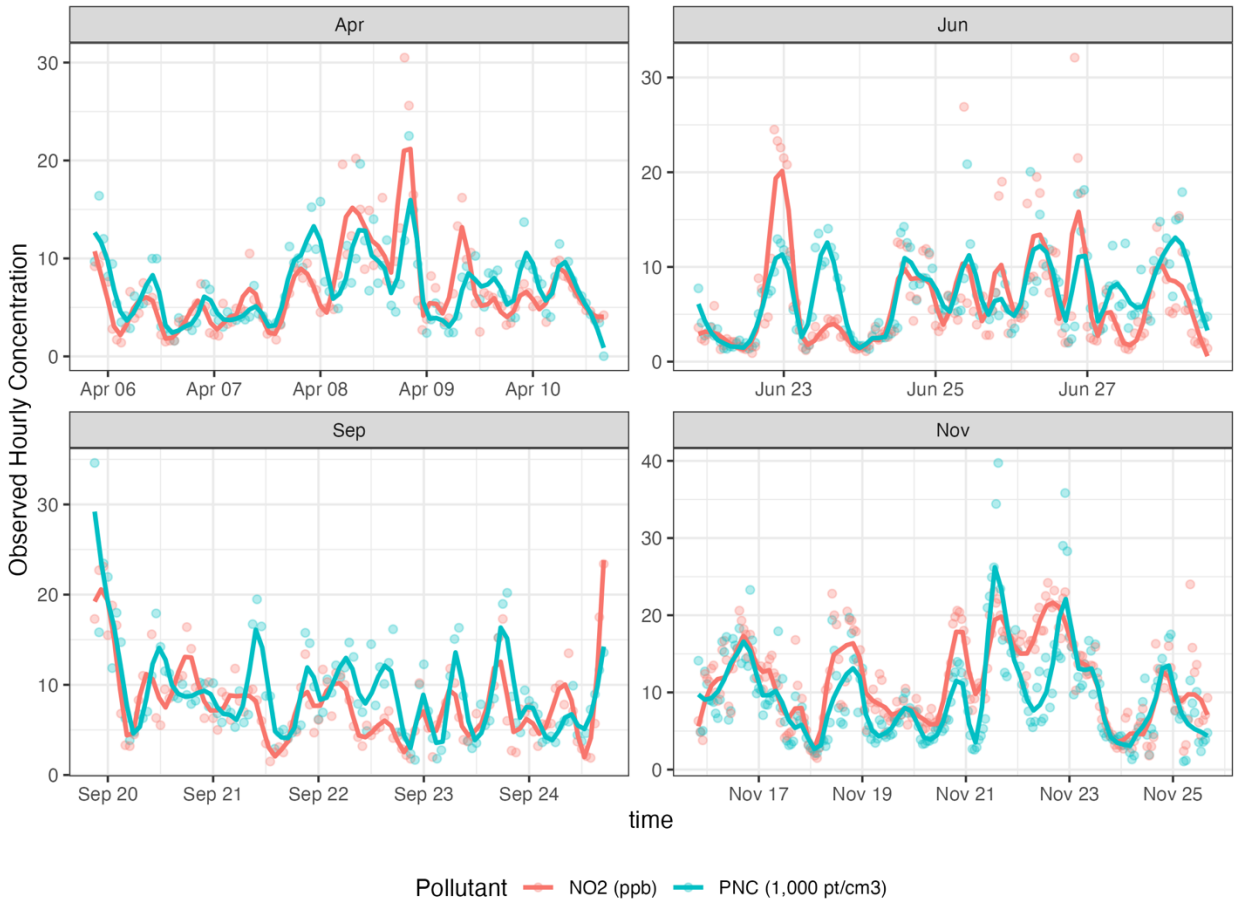


Figure S2. Time series of hourly NO_2 (ppb) and UFP (1,000 pt/cm3) at Beacon Hill used to simulate a continuous, long-term PNC monitoring site, as described in the Methods. There are 632 paired hourly observations with a Pearson correlation (R) of 0.64. Smooth lines describe the general pollutant trends.

Table S1. Continuous overnight sampling PNC times at the Beacon Hill monitoring site. Data were used to simulate a continuous, long-term PNC monitoring site, as described in the Methods.

Month	Dates	Start	End	Days
Apr	6	2019-04-05	2019-04-10	Fri, Sat, Sun, Mon, Tue, Wed
Jun	8	2019-06-21	2019-06-28	Fri, Sat, Sun, Mon, Tue, Wed, Thu
Sep	6	2019-09-19	2019-09-24	Thu, Fri, Sat, Sun, Mon, Tue
Nov	11	2019-11-15	2019-11-25	Fri, Sat, Sun, Mon, Tue, Wed, Thu

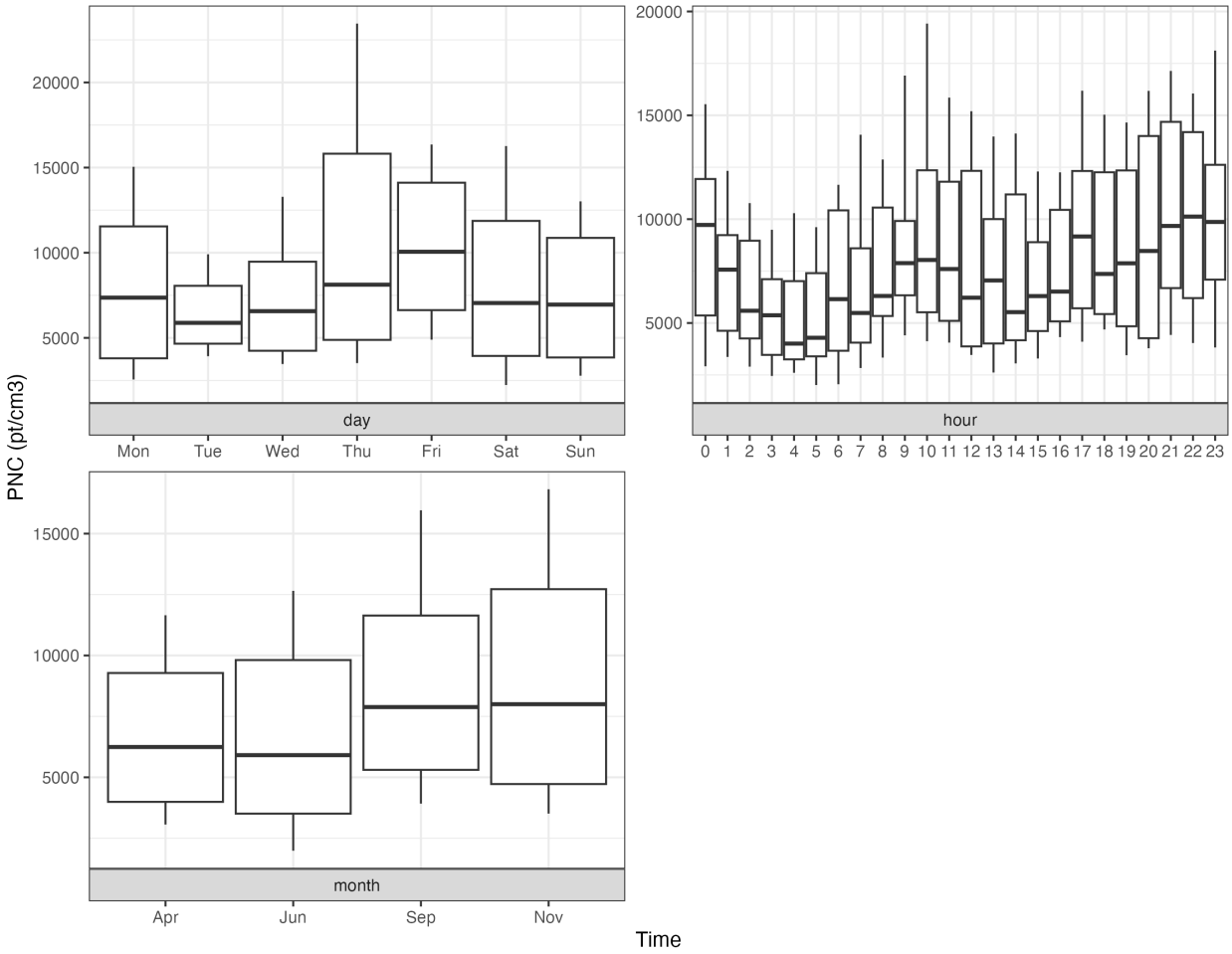


Figure S3. Distribution of hourly PNC levels at Beacon Hill, stratified by day of the week, hour of the day, and sampling month. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.

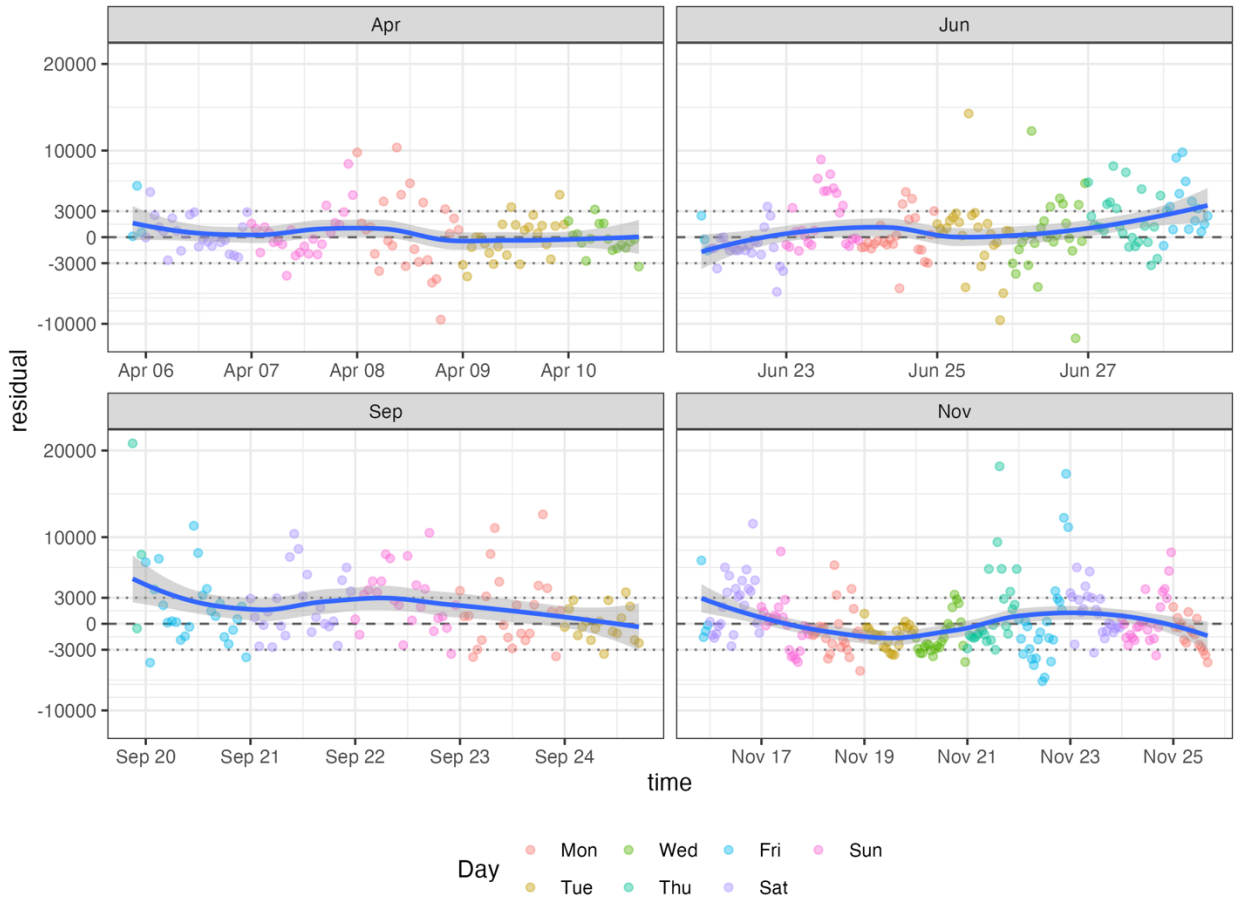


Figure S4. Time-series of simulated PNC model residuals. There is little remaining temporal trend in PNC.

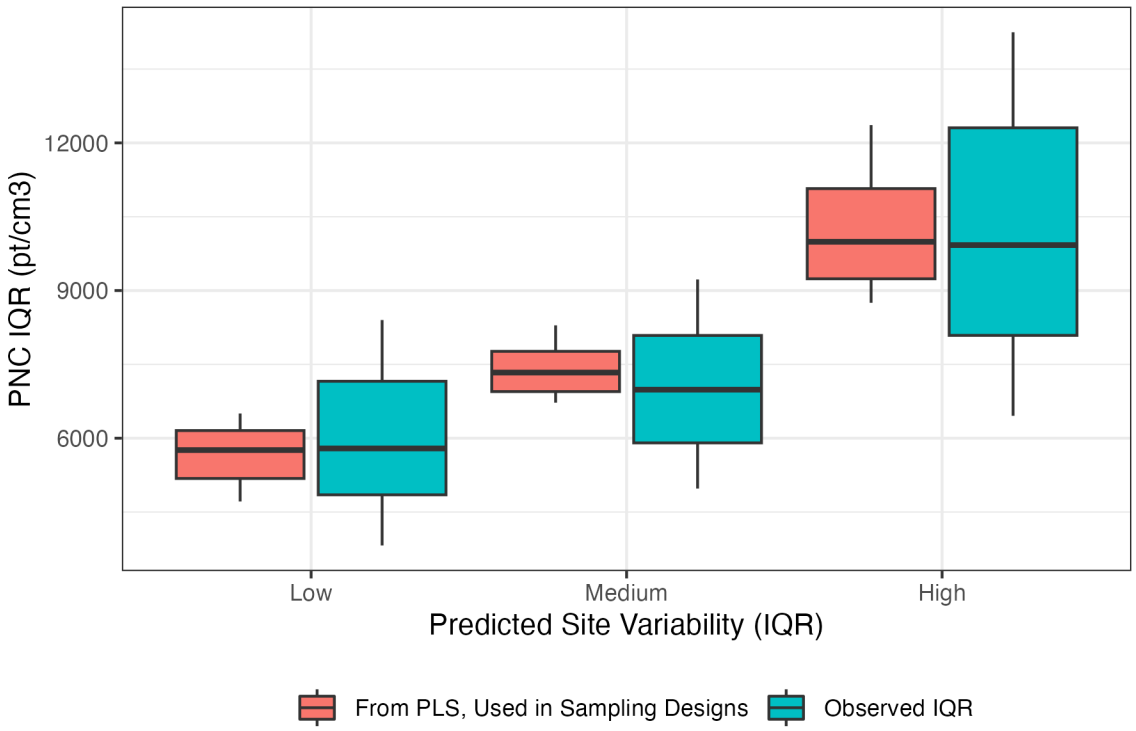


Figure S5. Distribution of PNC IQR (pt/cm³) for the unbalanced visits design, which categorizes sites as having low, medium, and high variability based on predicted IQR from PLS regression (see Methods for modeling details). The observed IQR used to fit the model is also shown. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.

2 Results

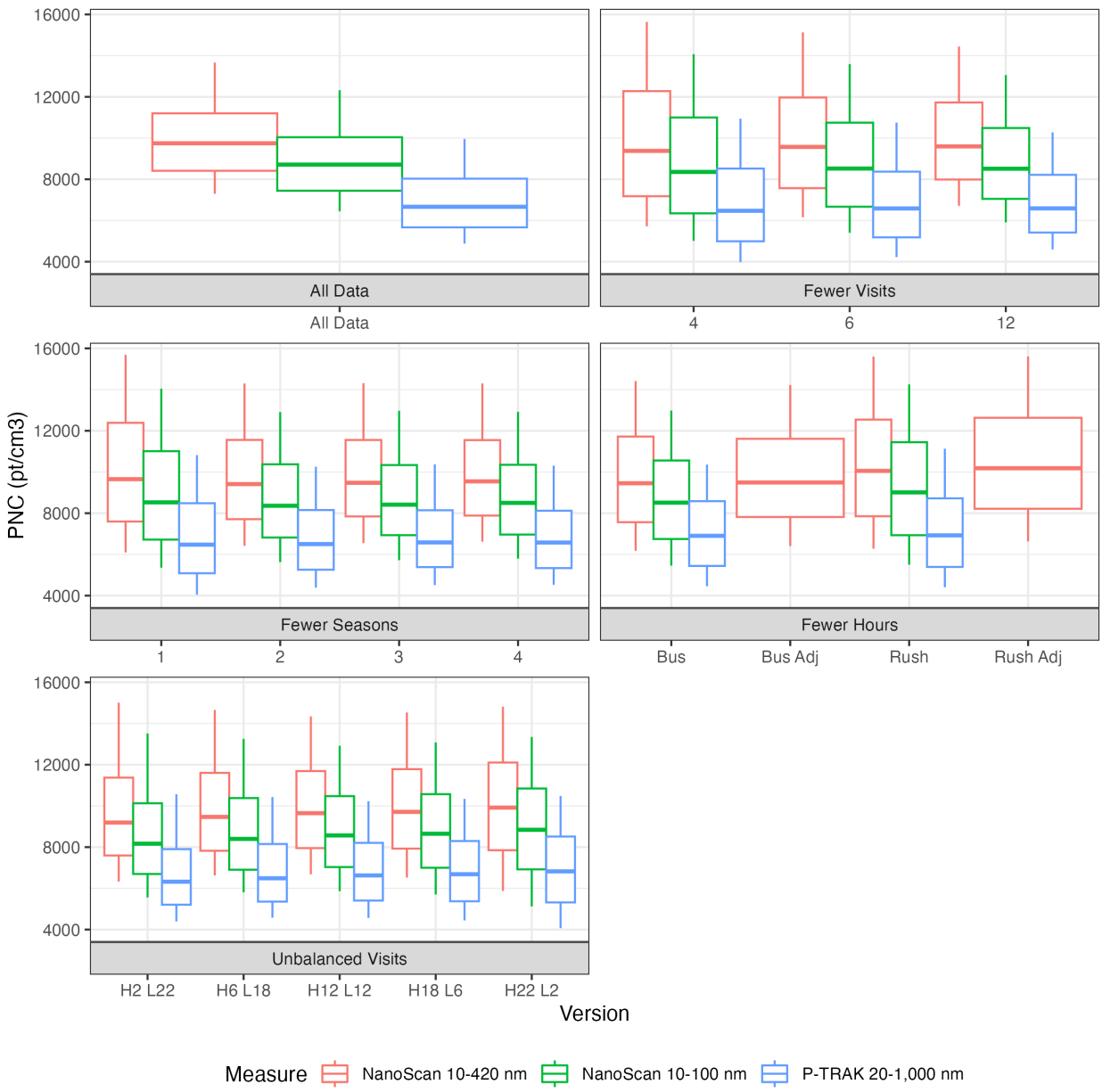


Figure S6. Distribution of estimated annual average site concentrations for the all-data campaign (N=1 campaign x 309 sites) and each sampling design (N=30 campaigns per design x 309 sites each) for primary (NanoScan 10-420 nm) and sensitivity analyses. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.

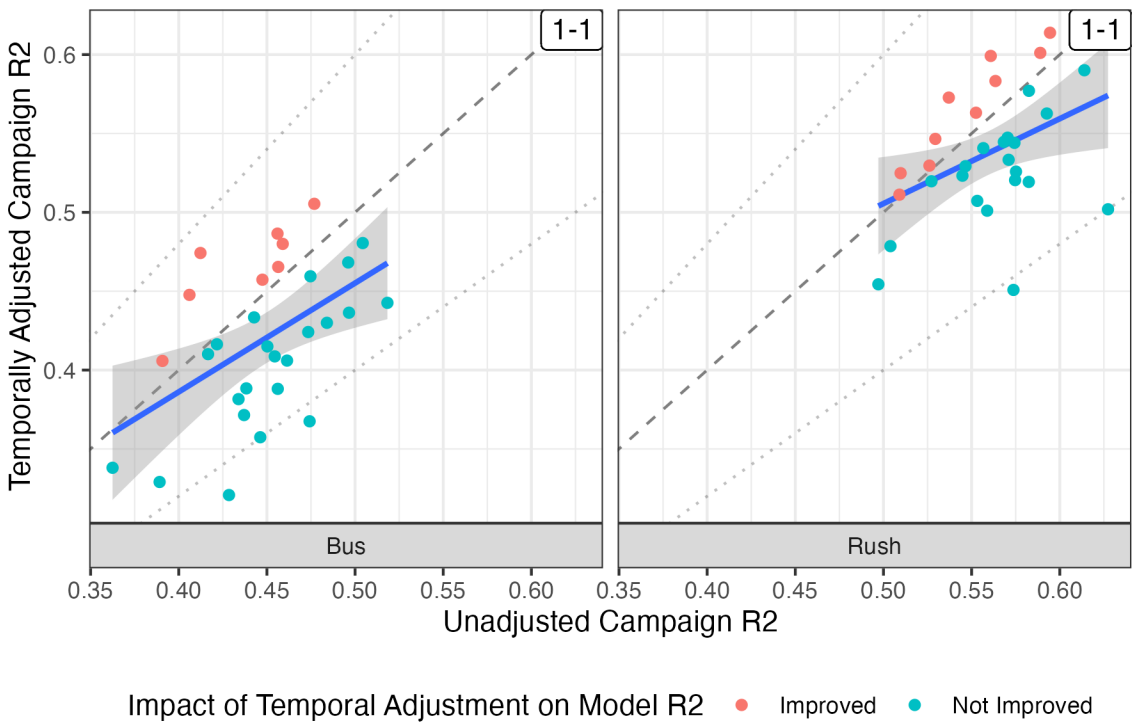


Figure S7. Comparison of unadjusted and temporally adjusted business and rush hour campaign exposure models. Temporal adjustment almost never improves business hour campaign models, and only sometimes improves rush hour campaign models. Dashed lines indicate the 1-1 and ±20% lines.

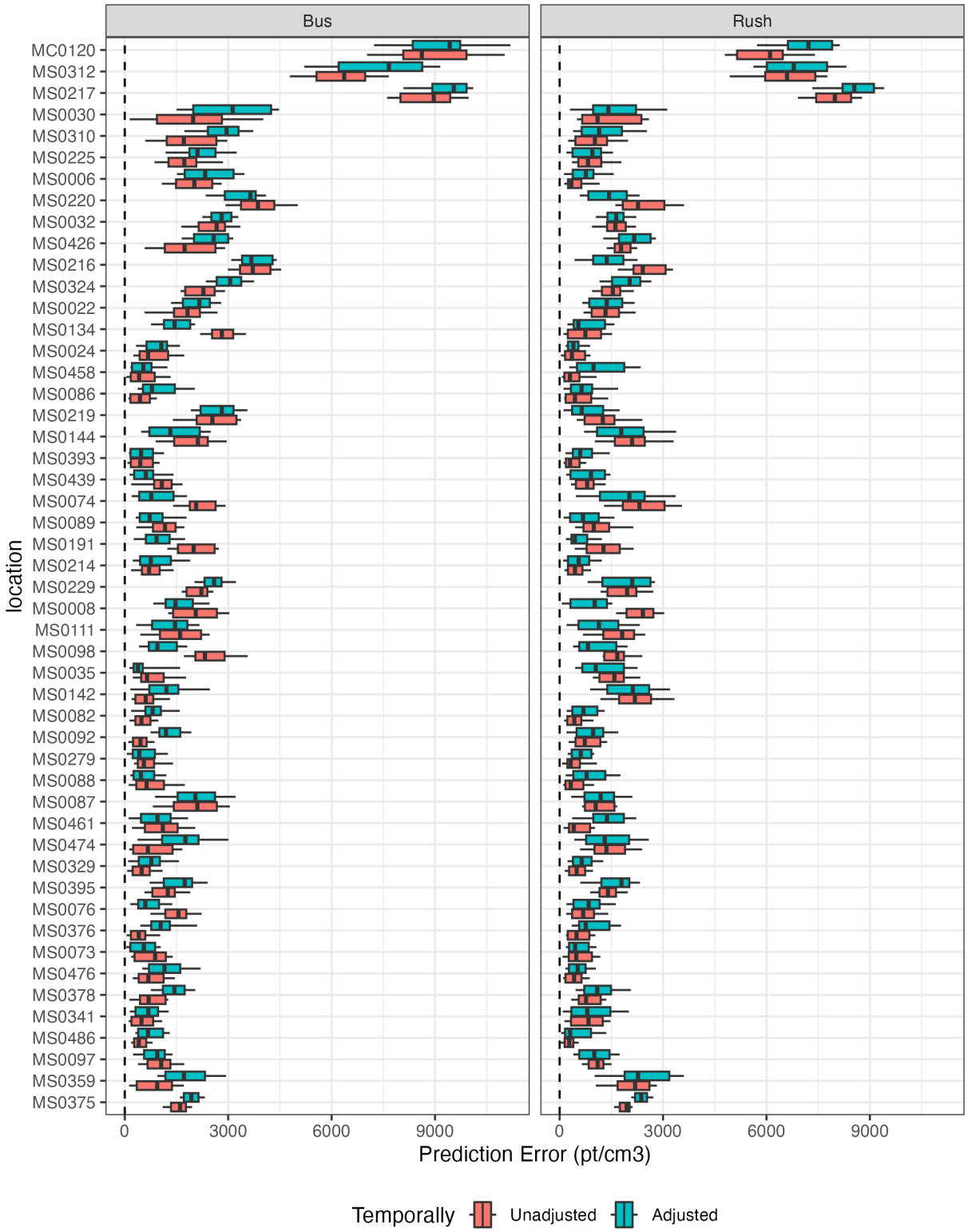


Figure S8. Absolute UFP prediction errors from the business and rush hour designs for a sample of 50 sites (n=30 prediction errors per site and adjustment approach, i.e., per boxplot). Prediction errors are calculated by comparing the cross-validated site prediction from each campaign to the all-data site observation. Sites are arranged by their all-data annual average concentration, with higher concentration sites near the top. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.

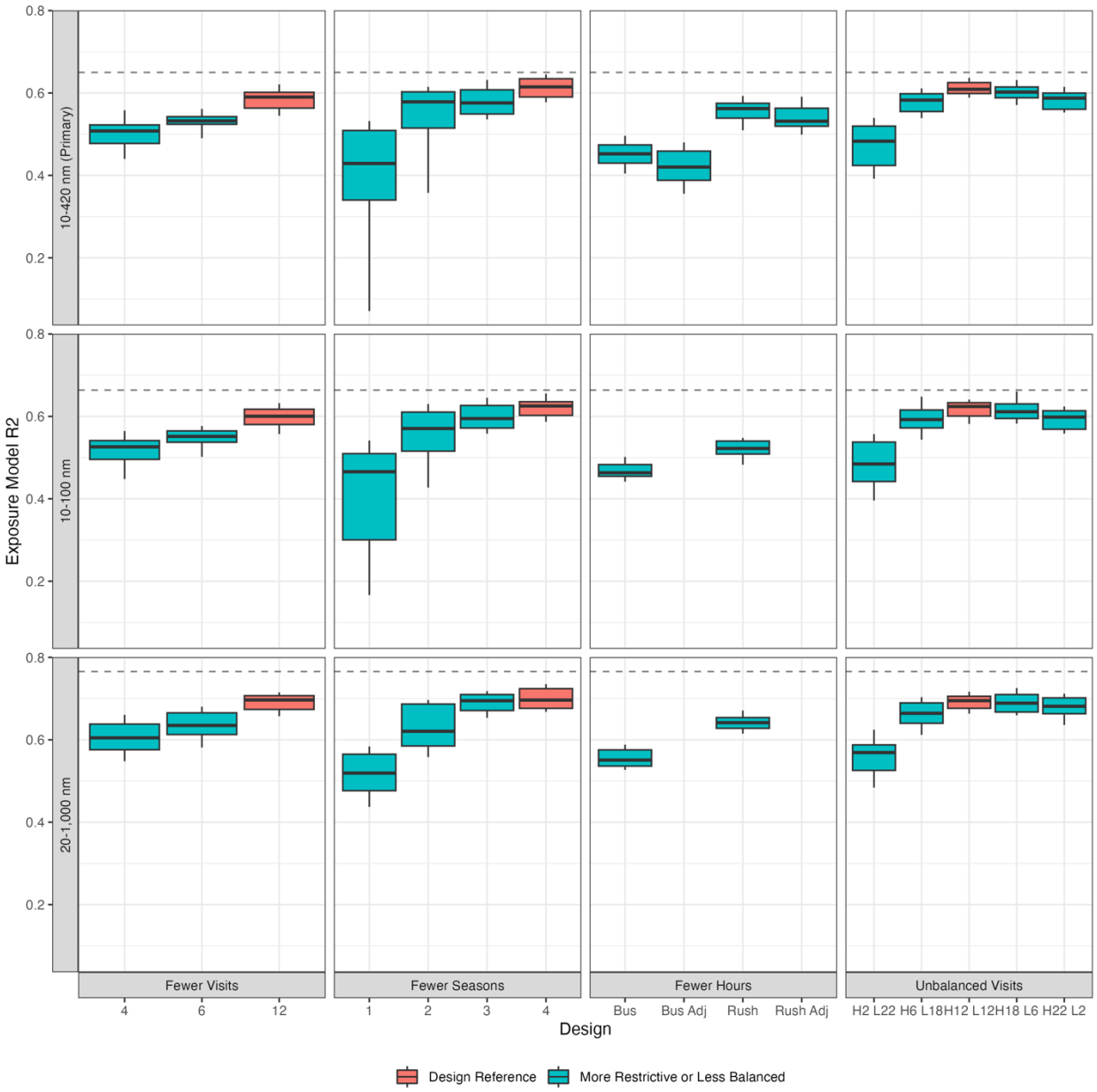
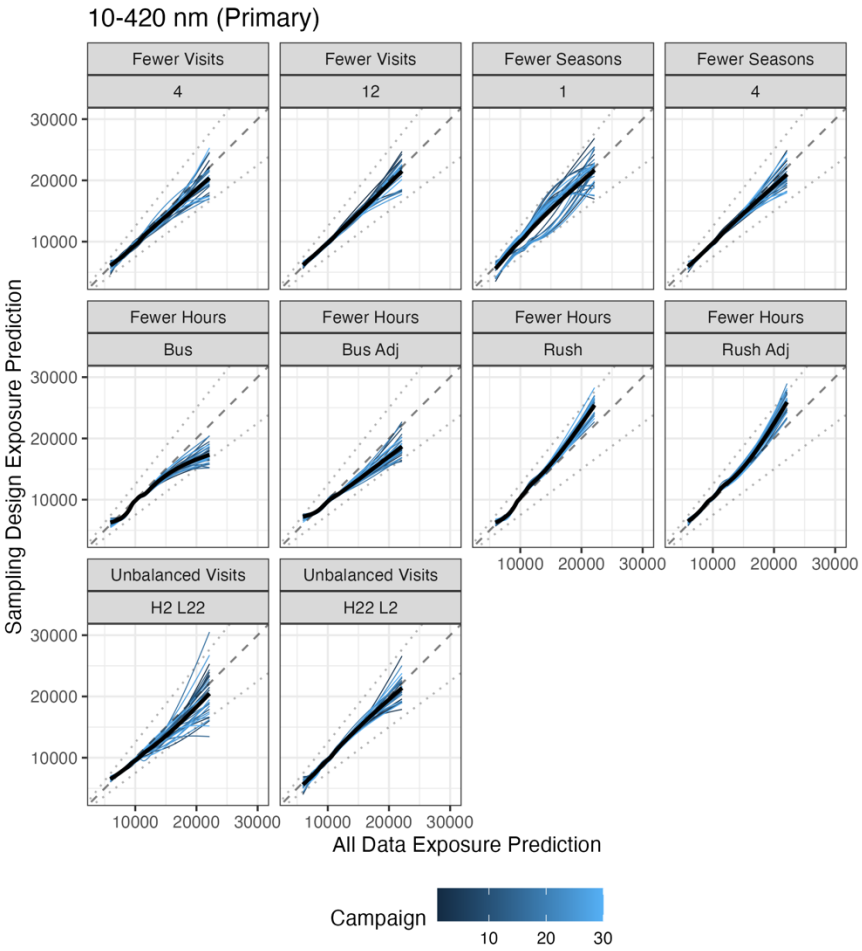


Figure S9.Cross-validated model performances (N=30 campaigns per design). The dashed lines indicate the all-data campaign performance for primary and sensitivity analyses. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.



(see below for caption)

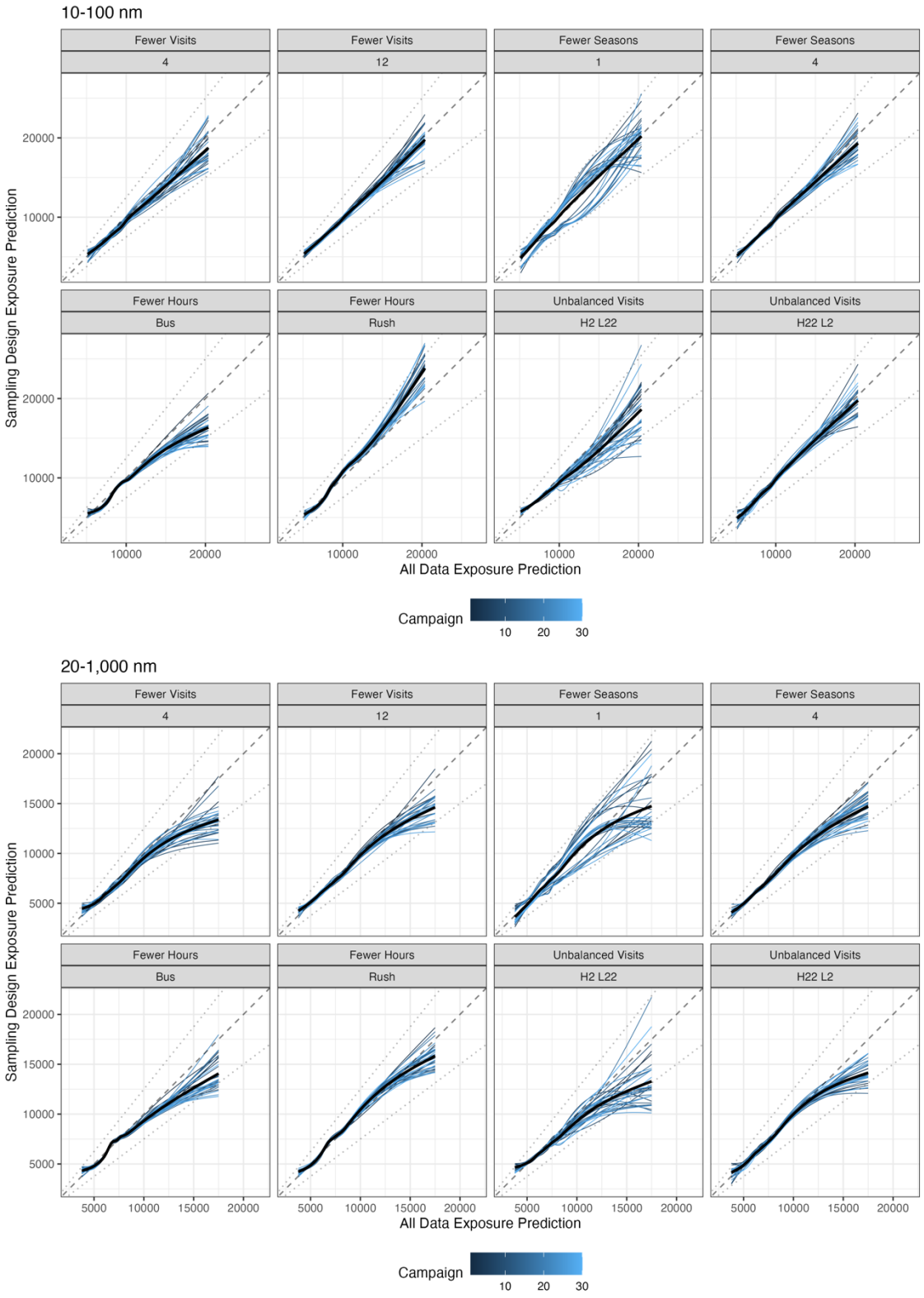


Figure S10. Smooth lines comparing predicted five-year average participant PNC (pt/cm³) exposure from the all-data campaign (N=1 campaign) to exposure predictions from other sampling designs (N=30 campaigns) for primary and sensitivity analyses. The black smooth line is the average trend for each design. The dashed lines indicate the 1-1 line as well as 25% above and below.

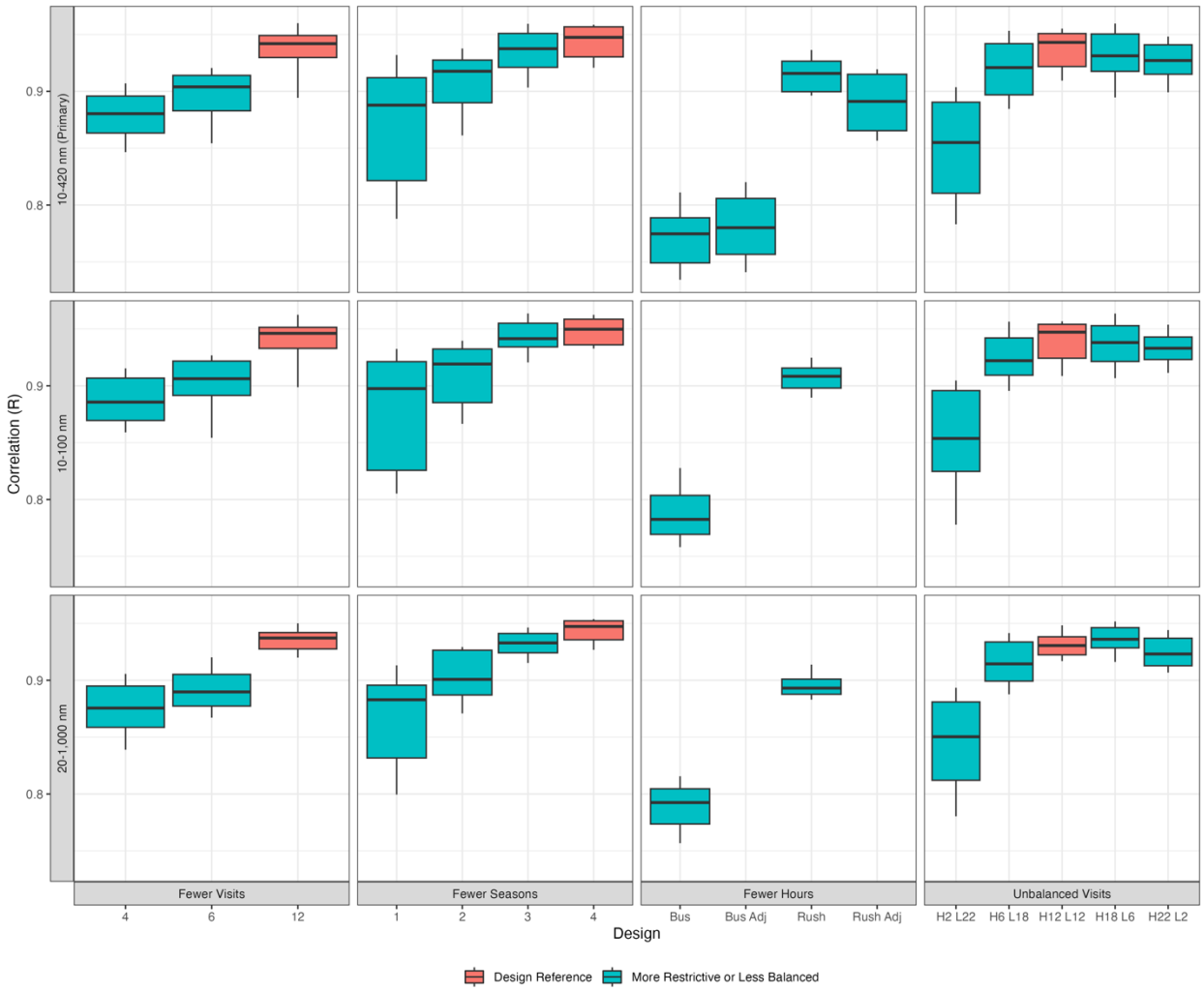


Figure S11. Pearson correlations (R) comparing predicted five-year average PNC participant exposure from all-data campaigns relative to each sampling design (N=30 campaign correlations per design) for primary and sensitivity analyses. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.

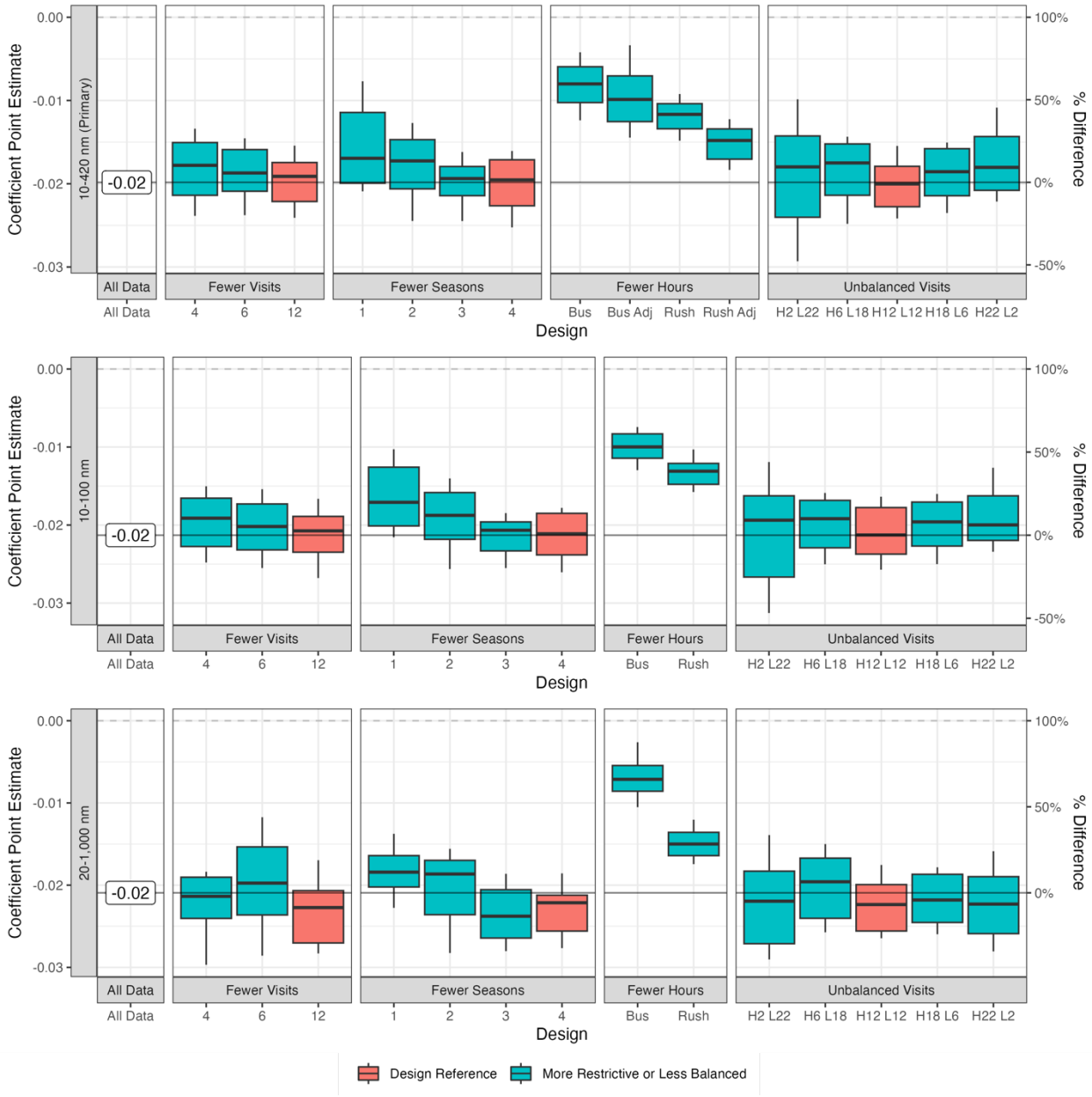
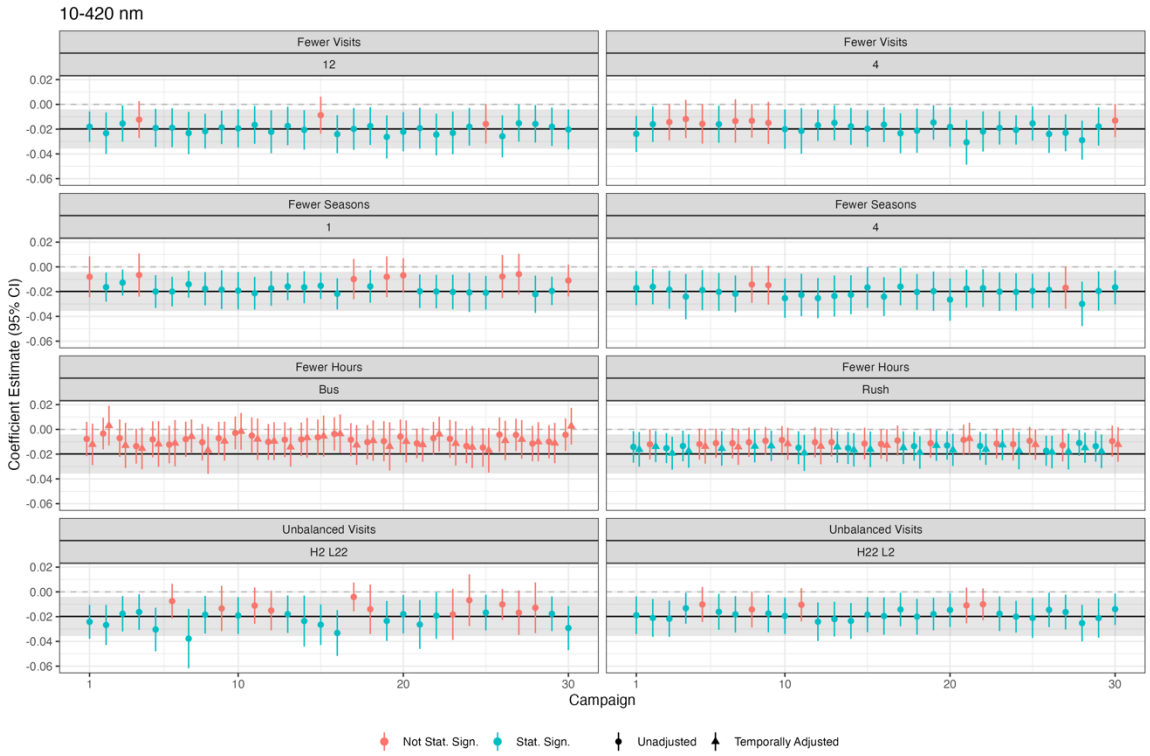


Figure S12. Health effect estimates produced from different sampling designs for the adjusted association between PNC (1,900 pt/cm3) and cognitive function (CASI-IRT) for primary (10-420 nm PNC) and sensitivity analyses. Percent difference is relative to the health effect estimate from the all-data exposure model. Boxes show the median and IQR, whiskers show the 10th and 90th percentiles.



(see below for caption)

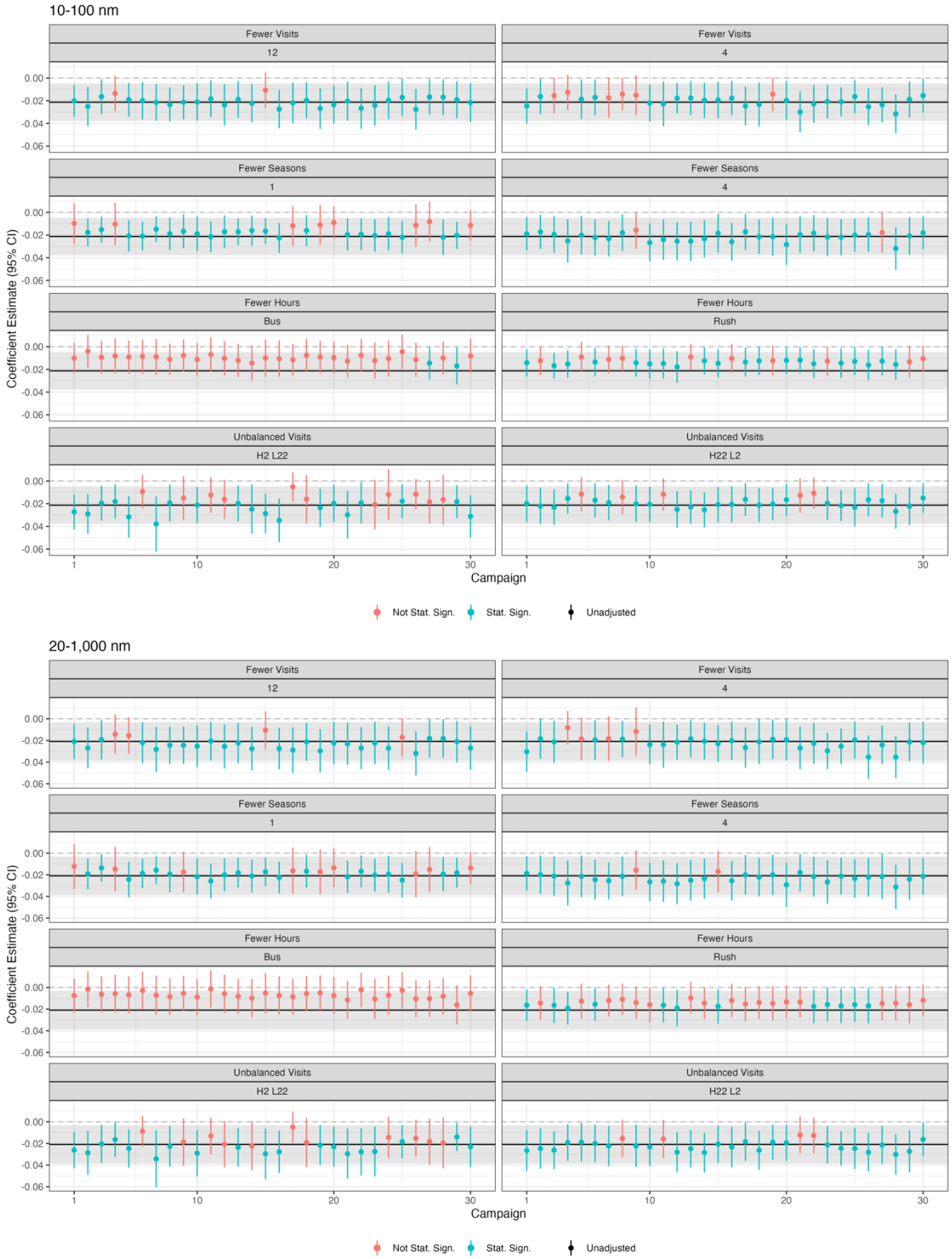


Figure S13. Health effect estimates for the association between UFP (1,900 pt/cm3) and CASI-IRT for sensitivity analyses after adjusting for age, calendar year, sex, education. The horizontal black line and shaded gray area are the point and 95% confidence interval estimates from the all-data campaign, respectively. Each point-range is the estimated health effect and 95% CI for a given sampling campaign.

3 References

US EPA. (2023). *Air Quality System Data Mart*. http://www.epa.gov/ttn/airs/aqsdatamart