ARTICLE

Optimal-design domain-adaptation for exposure prediction in two-stage epidemiological studies

Ron Sarafian1✉, Itai Kloog2 and Jonathan D. Rosenblatt1

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

BACKGROUND: In the first stage of a two-stage study, the researcher uses a statistical model to impute the unobserved exposures. In the second stage, imputed exposures serve as covariates in epidemiological models. Imputation error in the first stage operate as measurement errors in the second stage, and thus bias exposure effect estimates.

OBJECTIVE: This study aims to improve the estimation of exposure effects by sharing information between the first and second stages.

METHODS: At the heart of our estimator is the observation that not all second-stage observations are equally important to impute. We thus borrow ideas from the optimal-experimental-design theory, to identify individuals of higher importance. We then improve the imputation of these individuals using ideas from the machine-learning literature of domain adaptation.

RESULTS: Our simulations confirm that the exposure effect estimates are more accurate than the current best practice. An empirical demonstration yields smaller estimates of PM effect on hyperglycemia risk, with tighter confidence bands.

SIGNIFICANCE: Sharing information between environmental scientist and epidemiologist improves health effect estimates. Our estimator is a principled approach for harnessing this information exchange, and may be applied to any two stage study.

Keywords: Environmental epidemiology; Two-stage studies; Optimal-design; Domain-adaptation

INTRODUCTION

Environmental epidemiology (EEPI) is a branch of epidemiology that studies the health effects of environmental exposures, such as air pollution and temperature. Studies are typically observational, and relate exposure measurements to health outcomes at the individual-level. Common EEPI studies are usually conducted at the individual-level, and exposure assessment is geolocated to a particular point, such as an individual’s residence [1, 2].

Defining and measuring exposure accurately is far from trivial. Inaccurate exposure assessments will introduce measurement error, and will bias exposure effect (EE) estimates [3]. At the individual level, direct measurement of ambient exposure is usually impossible, and thus, an indirect measurement is performed. This indirect measurement can be thought of as an imputation, or prediction of the unknown exposure. A common approach to indirect measurement is predicting exposure from satellite imagery. For instance, indirect measurements of fine particulate matter (PM$_{2.5}$) are obtained by fitting a model that takes satellite products like aerosol optical depth (AOD) as predictors. The model is first calibrated using fixed ground monitoring stations, then used to predict PM$_{2.5}$ where ground stations are unavailable but AOD is [4, 5].

In a two-stage study the EE is estimated using indirect measurement, where the locations of individuals differ from the locations of monitors. Two-stage studies are tremendously popular in the EEPI community [6]. The current paradigm in EEPI research is that exposures are predicted once, and then serve multiple studies. Namely, in the first stage, the statistician provides a single exposure surface (e.g., a grid); then different epidemiologists use it for different studies.

The above imputation introduces measurement error. The measurement error literature offers many remedies [7]. For instance, [3, 8–10] compare various error correction methods. For satellite predictions, [11] use machine learning algorithms to improve predictions and thus reduce measurement error. The above either treat predicted exposures as fixed, or improve them uniformly for all studies. This, in contrast to the current proposal, where we tailor predictions to each study.

Ignoring the subsequent epidemiological study is the current practice in exposure modeling. Yet, Diao, and 16 other respectable epidemiologists, statisticians, and exposure scientists suggest that it should not be the case [12]. In the context of PM$_{2.5}$, they recommend that researchers from both communities work together when applying exposure predictions into health impact assessments. A similar argument was voiced by Szpiro et al. [13], which recommend that the development of models for exposure prediction and EE estimation should be considered simultaneously.

In this work, we try to answer these calls and suggest a framework for improving exposure predictions for EEPI studies. For improvement, we focus on the accuracy of EE estimates. This is
quite different from today’s state-of-the-art exposure models, which focus on first-stage (exposure) predictions.

The following example illustrates our argument: Assume the epidemiologist is fitting a linear model. We know that in simple linear regression, \( \text{Var}[\beta] = \sigma^2 / \sum(x_i - \bar{x})^2 \), where \( \beta \) is the estimated slope, \( x_i \) the observed predictors, and \( \sigma^2 \) the variance of departures from the linear trend. We thus see that for accurate effect estimates, extreme values of \( x_i \) are more influential than mean values. This suggests that the environmental epidemiologist should pay more attention to accurate predictions of \( x \) at its extremes than its typical values.

Unlike a designed experiment, the epidemiologist is not free to choose the exposure levels of the participants. However, in two-stage studies, the epidemiologist is free to improve the accuracy for “important” individuals. Identifying individuals of importance can be done with the theory of optimal-design. Having identified these individuals, we need to improve their exposure predictions. This can be done with the theory of domain-adaptation, a.k.a. transduction or dataset-shift. Our approach thus consists of two components: (i) At the second stage: use optimal-design to identify the observations of importance. (ii) At the first stage, use domain-adaptation to improve exposure predictions where required for the second stage. The optimal-design might depend on the exposures themselves, so an iterative structure is suggested.

Here, we cast our ideas as an estimation algorithm. We present simulation results that confirm its efficiency comparing to current (non-iterative) best practices, and apply it to empirically estimate PM2.5 effect on hyperglycemia risk.

METHODS
Problem Setup
Let \( y \in \mathcal{Y} \) be a health outcome. Usually \( \mathcal{Y} \subseteq \mathbb{R} \), or \( \mathcal{Y} = \{0,1\} \). We allow the exposure to be multivariate (e.g., PM2.5 and PM10; temperature, humidity, etc.), and denote it with \( x \in \mathcal{X} \subseteq \mathbb{R}^p \). Lastly, \( z \in Z \subseteq \mathbb{R}^q \) are some other covariates in the epidemiological model. Typically, we assume that \( y \)'s distribution, \( F \), depends on \( x \) and \( z \), and some parameter \( \theta \):

\[
y(x, z) \sim F(\theta; x, z). \tag{1}
\]

We will denote all the modeling assumptions in Eq. (1) with \( \mathcal{M} \). For instance, \( M \) may state \( y(x, z) \sim \mathcal{N}(\beta^T x + \phi(z), \sigma^2) \). Here \( F \) is Gaussian, and \( \theta = (\beta, \phi, \sigma^2) \). The EE in this case is \( \beta \). The epidemiologist has access to a sample of \( i = 1, \ldots, n \) individuals. For each \( i \), \( y_i \) and \( z_i \) are known, but the true exposure, \( x_i \), is unknown. We denote with \( \tilde{x}_i \) an indirect measurement of the exposure of individual \( i \), sometimes known as an imputation, surrogate, or proxy. We call \( D \) the second-stage dataset, that consists of \( n \) triplets \((y_i, x_i, z_i)\). Because \( \tilde{x}_i \) is used instead of \( x_i \), the epidemiologist may estimate \( \beta \), but not \( \beta \). A good proxy of \( x \) is one where \( \|\tilde{x}_i - x_i\| \approx 0 \).

Moving to the exposure scientist. We use \( J \) to denote measurements in the first stage. We think of \( x \) as the exposure at location \( s \), which is measured directly without noise; for instance, because \( s \) is the location of a ground monitoring station. We think of \( r \) as a vector of geographical covariates that includes satellite products (e.g., AOD), and other spatial predictors such as land-use, spline-basis functions, etc. Because \( x \) and \( r \) are spatial we may spatially unit \( x_i(s) \), \( r_i(s) \) when we want to emphasize the spatial structure of the problem. The data \( D' \), consists of \( n \) pairs \((x_i(s), r_i(s))\). The indirect measurement, \( \tilde{x}(s) \), is done by applying some statistical learning algorithm, on the training dataset \( D' \), and applying the resulting predictor \( h_{oracle} \) at \( s \). We denote this \( \tilde{x}(s) := h_{oracle}(s) \).

To fix ideas, \( h_{oracle} \) may be a Gaussian random field [5], a Gradient boosting machine [14], a deep network [15], or an ensemble of different machine-learning algorithms [16].

In this text we assume the first-stage learning paradigm is empirical risk minimization (ERM). Let \( h \) be a predictor from some hypothesis class, and \( l(.) \) the loss for predicting \( h(r) \) instead of \( x(s) \): \( l(h(r), x(s)) \rightarrow \mathbb{R}^+ \). The risk function can be written as

\[
\mathcal{R}(h) := \mathbb{E} [l(h(r(s)), x(s))]. \tag{2}
\]

A more explicit formulation that clarifies the sources of variability in Eq. (2) is given in the section “Adaptation to the optimal design”.

In a usual two-stage study, the exposure scientist’s problem is finding the predictor that minimize empirical risk on the training set:

\[
h_{oracle}^{opt} := \arg \min_h \frac{1}{n} \sum_{i=1}^n l(h(r(s))), x(s)). \tag{3}
\]

We denote the risk-minimizing predictor with \( h_{oracle}^{opt} \), where \( n \) is the number of iterations between the exposure scientist and the epidemiologist.

Alternatively, the exposure scientist may adopt the epidemiologist’s view, and seek a predictor that has a small second-stage estimation error:

\[
h_{oracle}^{sec} := \arg \min_h \mathbb{E} [l(h(r(s))), \beta_x]^2]. \tag{4}
\]

We call \( h_{oracle}^{sec} \) an oracle, because it requires knowledge of the unknown \( \beta_x \). Ideally \( h_{oracle}^{opt} \), and \( h_{oracle}^{sec} \) are close, and return similar predictions. In reality, \( h_{oracle}^{opt} \), and \( h_{oracle}^{sec} \) might be quite different, which will inflate \( \|\beta_x - \beta_{oracle}\| \) at the second stage.

We will now focus on generalized linear models (GLMs), for two reasons. First, they are very popular second-stage models in EEP studies. Second, optimal-design for GLMs have been extensively studied [17]. Extensions to the GLM are discussed in Section “Departing from GLMs’.”

A GLM has the following components: (1) An assumed distribution of health outcome, \( F \), that is restricted to the exponential family. We denote its mean by \( \mu \). (2) A linear predictor of exposure and other covariates: \( \eta = \beta \Phi(x, z) \), where \( \beta \in \mathcal{B} \) are unknown parameters. (3) A link function \( g(\mu) = \eta \) relating the linear predictor to the expected health outcome.

The vector \( \Phi(x, z) \) is the transformed features. Usually, \( \Phi(x, z) = (1, x, x^T) \), i.e., a \( 1 + p + q \) vector, but \( \Phi(x, z) \) may also include interactions and other transformations of covariates. In a GLM, the variance of \( y \) has the form

\[
\text{Var}[\eta|x, z] = \phi(V(\mu)),
\]

where \( \phi \) is a constant and \( V(.) \) is some function of \( \mu \), both determined by \( F \).

Optimal design

For ease of exposition, we assume that the second-stage model includes only the \( (p \times q) \)-dimensional exposure as explanatory variables with an intercept, so the linear predictor is given by \( \eta = \beta \Phi(x, z) = \beta_0 + x^T \beta_x \). In the section “The effect of second-stage covariates” we discuss the scenario where more covariates are included. An optimal design identifies the values of \( x \) and their sampling probabilities, so that uncertainty in estimates is minimal. These are called the support points of the design, and the corresponding probabilities are called the design weights. For \( j = 1, \ldots, J \) denote \( x_j \in X \) an exposure support point, and \( w_j \) a corresponding weight. The design weights meet: \( \sum_j w_j = 1 \).

A good proxy of \( x \) is one where \( \|\tilde{x}_j - x_j\| \approx 0 \). One of the nice properties of the GLM with design \( \xi \) is the general and compact form of the information matrix, \( I(\xi, \beta) \), it implies:

\[
l(\xi, \beta) = \sum_{j=1}^J w_j u(x_j) \Phi(x_j) \Phi(x_j)^T.
\]

where\( u(x) \) is the model weight of point \( x \):

\[
u(x) := \frac{1}{\phi(V(\mu))} \left( \frac{\partial \mu}{\partial \eta} \right)^2.
\]

and \( l(.) \) is the variance function defined in Eq. 5 [17].

An optimal design, \( \xi \), minimizes some function \( V(.) \) of the information, with respect to support point and weights, within the permissible region:

\[
\tilde{\xi} := \arg\max_{\xi} V(l(\xi, \beta)).
\]

Setting \( V(.) := \text{det}(.) \) is known as a D-optimal design, which is arguably the most popular; \( V(.) := \text{Tr}(.) \) is an A-optimal design; \( V(.) := \lambda_{max}(.) \), where \( \lambda_{max}(.) \) returns the minimum eigenvalue is an E-optimal design. A review is given in Fedorov and Leonov (2013, Ch. 2) [18].

In linear models, \( \tilde{\xi} \) is independent of any unknowns. In non-linear models, \( \tilde{\xi} \) depend on both the unknown \( \beta \) and the assumed distribution. Clearly, an inconvenient circularity arises if the optimal design depends on the quantity to be recovered by the design. The literature offers various remedies to this matter [17]. For instance, in sequential designs the idea is
to start from an initial static design to estimate the parameters of interest; then sequentially parameters and design are updated.

**Adaptation to the optimal design**

*Domain-Adaptation* deals with the learning of a predictor in one population, and its application in another. These are known as source and target distributions, respectively. In our case, we will learn to predict exposure in the first stage, and apply predictions in the second stage. This means, for instance, that the first (second) stage marginal distribution of $x$ represents exposures at monitoring stations (residences of subjects). We denote the joint source distribution of geographical covariates and exposure with $P_1(r,x)$. We use $P_1(r,x)$ for their joint target distribution. Our goal is to train a predictor with samples from $P_1(r,x)$, to predict the unknown exposure of samples drawn from $P_1(r,x)$, $P_1(r,x)$ and $P_1(r,x)$ might be quite different. For instance, when individuals are exposed to different levels than those measured in monitoring stations.

We can now write Eq. (2) more precisely, i.e., the risk function that the exposure scientist minimizes if unaware of the epidemiologist's needs:

$$R_S(h) := \int l(h(r),x)dP_1(r,x).$$

In $R_S(h)$, integration is taken w.r.t. the source distribution. If, on the other hand, the exposure scientist is aware of the target domain, then instead of Eq. (9), the risk is the average loss in the target population:

$$R_T(h) := \int l(h(r),x)dP_1(r,x).$$

$R_T(h)$ can be rewritten using integration w.r.t. the source distribution:

$$R_T(h) = \int \omega(x)l(h(r),x)dP_1(r,x),$$

where $\omega(x)$ are called the importance weights [19], and satisfy:

$$\frac{P_1(r,x)}{P_2(r,x)} = \frac{\omega_1(x)}{\omega_2(x)} = \omega(x) = \frac{P_1(x)}{P_2(x)}$$

When $P_1(r,x) = P_2(r,x)$, i.e., $\omega(x) = 1$, then no adaptation is required. This is actually implicit in today’s current best practices. In contrast, when $P_1(r,x) \neq P_2(r,x)$, the exposure scientist should know $\omega(x)$ in order to minimize the empirical counterpart of $R_T(h)$.

The issue is simplified if we assume a priori-shift (a.k.a. label shift) [20]. Under the prior-shift assumption, conditional distributions are assumed equal:

$$P_1(r,x) = P_2(r,x),$$

whereas the prior distributions of the exposures differ: $P_0(x) \neq P_1(x)$. More on this assumption in the section "Prior shift assumption". Substituting the prior shift assumption in Eq. (12) implies:

$$\omega(x) = \frac{P_1(x)}{P_2(x)}$$

$P_0(x)$ is the marginal distribution of exposures in the first-stage. We estimate it from $D$ and denote the estimator $\hat{P}_0(x)$. $P_1(x)$ is the marginal distribution of exposures in the target population, i.e., in the second stage. In Sarafian et al. [21] $P_1(\cdot)$ was known because we set it to be the locations of residence of second-stage subjects. In this contribution, we set the target to be the exposure support points identified by the second-stage's optimal design.

To set $P_1(x)$, we observe that the optimal design, $\xi$, is a (discrete) probability distribution over $X$. We could define $P_1(x) = \xi(x)$ and $\omega(x) = \xi(x) \times P_0(x)$. However, $\xi(x)$ and $P_1(x)$ do not agree on the support, thus, we suggest smoothing $\xi(x)$, for instance, with kernel density estimators.

By convolving $\xi$ with some kernel function $K$, we get a mixture distribution:

$$P_1(x; \xi) \propto \xi(x) \times K(x,x) = \sum_{j=1}^{k} w_j K(x,x_j),$$

where $K(x,x_j)$ is some kernel function that weights according to the distance of $x$ from the support point $x_j$ and $\propto$ means equality up to some normalizing constant.

The empirical counterpart of the weighted risk in Eq. (11) can now be derived, using $\omega(x) := P_1(x; \xi)/P_0(x)$.

$$h^{\text{DiW}}_{\text{opt}} = \arg \min_{h} \frac{1}{n} \sum_{i=1}^{n} \omega(x(i))(h(r(i)),x(i)).$$

**Algorithm**

Equipped with optimal-design and domain-adaptation theory, we now suggest an estimation algorithm that allows to estimate the EE accurately.

The crux is to iterate between the exposure scientist (Esther) and epidemiologist (Ephraim): (1) Ephraim provides exposure predictions. (2) Ephraim uses them to estimate the EE. (3) Ephraim uses optimal-design to mark data points of importance. (4) Esther uses domain-adaptation to improve predictions at those points.

The details of our algorithm are the following. Denote $h_{D,\omega}$ the ODIW predictor from Eq. (16), learned with importance weights $\omega$ and dataset $D^*$. Denote with $x \leftarrow h_{D,\omega}$ the vector of exposures predicted using $h_{D,\omega}$. Denote with $\omega^0$ an initialization of weights. Denote $\beta_l \rightarrow M_{l+}$, effect estimates in epidemiological model $M$, estimated using exposures $x$.

Denote with $\tilde{\xi} = \arg \max_{\xi}(\det(\xi(\xi; \beta_I)))$ the D-optimal design of $\xi$, defined in Eq. (8), with information matrix evaluated at $\beta_0$, as defined in Eq. (6). Finally, let $P_1(x; \tilde{\xi})$ be the estimated density implied by $\tilde{\xi}$ as defined in Eq. (15), and $P_1(x)$ a density estimate for $x$ in $D^*$.

Algorithm 1 has many design choices. The obvious ones are the epidemiological model, $M$, and the predictor’s hypothesis class. These have received enough attention in the literature so we will not elaborate. Design choices that are more specific to our setup include: initialization choices; optimization tuning; stopping rules; and the optimality criterion. These are discussed in the section “User selected tuning parameters”.

Algorithm 1. ODIW Estimator. function $\text{EE Estimator}(D^*, D, M, K, w^0)$. $x^l \leftarrow h_{D,\omega}$ initialize exposures. for $i \in \{1, \ldots, L\}$ do $\beta_i \rightarrow M_i$, Estimate EE with current exposures. $\tilde{\xi}^i = \arg \max_{\xi}(\det(\xi(\xi; \beta_i)))$ Find D-optimal design. $\omega^i \leftarrow P_1(x; \tilde{\xi}^i)/P_0(x^i)$, $\forall i = 1, \ldots, n^*$ Re-Weight $x_i$, $\omega_i^{i+1} \leftarrow h_{D,\omega}$ Update exposures using current weights. end for. return $\hat{\beta}$; end function

**RESULTS**

**Simulation analysis**

The following simulation examines the accuracy of Algorithm 1 in estimating the EE. For first-stage data, $D^*$, we simulate satellite data and spatial predictors, from a zero-mean multivariate normal distribution: $r_j \sim N(0,\Sigma)$, where $\Sigma = U'U$ and $U$'s entries are independent uniformly distributed: $U_{ij} \sim \text{Unif}[0,1]$. The true exposure, $x_i$ is some linear function of $r$ with additive Gaussian noise: $x_i = y_i r_i + e_i$, A binary health outcome, $y_i$ is simulated from a Bernoulli distribution with a logit link:

$$P(y_i = 1|x_i; \beta) = \frac{\exp(\beta_0 + \beta_x x_i)}{1 + \exp(\beta_0 + \beta_x x_i)}.$$ (17)

Predicted exposures are restricted to be linear in $r$. We also compared other classes, including support vector regression with non-linear kernels; results were qualitatively the same, and so not reported herein.

Figure 1 compares the accuracy of Algorithm 1’s ODIW-iterative estimator (ODIW) and the standard two-stage estimator (Naïve) in estimating $\beta_x$. Main finding is that a small number of iterations almost always improves accuracy compared to a non-iterative approach. This can be seen from the distribution of $\beta_0 - \hat{\beta}_x$ in the upper left display. The number of iterations seems to decrease $\hat{\beta}_0 - \hat{\beta}_x$ (lower right), but after enough iterations, overfitting may kick in, and errors will grow.

The upper right panel illustrates the true exposure $x$ versus predicted $\hat{x}$ after $L = 10$ iterations. It is not surprising that the Naïve
predictions are fairly accurate for all $x$ (green). ODIWI, on the other hand, does not try to give accurate predictions for all $x$, but rather, only at the support of the optimal design ($\xi$, in red). Interestingly, ODIWI gives worse $\beta_x$ on average, but improves $\beta_0xC0$$\beta_x$.

Figure 2 compares the bias in $\beta_0xC0$$\beta_x$ between the ODIWI and Naïve estimates, when $\beta_x$ change from 0 to 2. It can be seen that when the true EE is stronger, i.e., the relation between $x$ and $y$ is less linear, there is more to benefit from optimal-design, and so ODIWI’s estimates have lower bias than those of the Naïve.

To protect from first-stage overfitting, Algorithm 1 may includes some design choices. We consider here the following: (i) choosing a momentum ratio $\alpha$, which determines how much previous iterations affect estimates updates; (ii) choosing the bandwidth of the kernel operator, $\sigma_K$, which determines the importance weights’ decay around the support points; (iii) choosing the number of iterations.

Figure 3 presents the evolution of ODIWI’s estimates (red lines), comparing different momentum and bandwidths (Gaussian kernel’s standard deviation) setups. In most cases, ODIWI’s estimates approach $\beta_x$ (dashed line), and then stagnate closer to it, in relation to the Naïve estimate (green line). It can be seen that the bias in ODIWI’s estimates do not prone to increase sharply after dozens of iterations.

Figure 4 presents the distribution of the error $\beta_0xC0$$\beta_x$, when estimated by a single-step ($L = 1$) ODIWI (red line), and the Naïve (green line) estimators. We compare different data generative processes, by changing $\beta_x$ (columns), and the variance of the first-stage model noise, $\sigma_T^2$ (rows). The analysis suggests that one iteration improves the accuracy of the Naïve estimator: the centers of ODIWI estimators’ error distributions are closer to 0 in all generative setups.

Pollution and glucose example
Recent studies found positive association between PM$_{2.5}$ and fasting blood glucose [22, 23]. We now employ Algorithm 1, to re-estimate the EE of PM$_{2.5}$ on the potential risk for hyperglycemia. We are interested in comparing Algorithm 1’s estimates to the current (non-iterative) standards in the field.

Exposure data includes daily PM$_{2.5}$ measurements over the years 2003–2012 from 46 monitoring stations in Israel, alongside satellite measurements of AOD, normalized-difference-vegetation-index (NDVI), and other spatial and temporal features, which are also available at 1 km$^2$ resolution in the residences of the individuals in the study.

The epidemiological data, also used by Yitshak Sade et al. [24], includes over 0.5 million blood glucose tests performed by approximately 43,000 individuals during the years 2003–2012 in Southern Israel, along with other subject’s characteristics. The
The study was approved by the IRB committee of the Soroka University Medical Center.

We now describe the analysis setup. We follow [4] and predict PM$_{2.5}$ at the subjects’ residences with a linear-mixed-model using satellite measurements and other geospatial features.

The assumed second-stage epidemiological model has the following form:

$$\log \mathbb{P}(y = 1|x, \xi) = \beta_0 + \sum_{k=1}^{6} \beta_{z_k} z_k + \beta_x \bar{x},$$

where $y \in \{0, 1\}$ is a binary response indicating whether the subject’s blood glucose level is above 126 mg/dl (a clinical cutoff used in diagnosing diabetes); $\beta_0$ is an intercept; $z_1, \ldots, z_6$ are covariates capturing seasonal variables and subject’s health and socio-economic characteristics (such as age, smoking status, diabetes status, weight, BMI, and socio-economic group), with corresponding coefficients $\beta_{z_1}, \ldots, \beta_{z_6}$; $\bar{x}$ is the subject’s average predicted exposure to PM$_{2.5}$ over last 21 days before test, and $\beta_x$ is the corresponding EE.

Figure 5 presents the progress of Algorithm 1. The algorithm was initialized on the Naïve estimates (iteration 0), and was operated with $L = 5$ iterations. We report bootstrap confidence intervals, since the usual large-sample—parametric inference for GLMs does not account for uncertainty introduced by the iterative imputations. It can be seen that EE estimates stabilize after two iterations. Our estimates indicate that the effect of PM on subject’s blood glucose level is slightly smaller than the effect estimated in the Naïve approach, with slightly tighter intervals.

For sensitivity analysis, we repeated this analysis among different subgroups, stratifying along diabetes status or subjects’ age. We also checked the sensitivity of the results to specific confounding variables such as socio-economic or smoking status. Our findings suggest that a stabilization of the EE estimate after

---

**Fig. 3** $\hat{\beta}_x$ estimates over the number of iterations. Dashed line indicates the true exposure effect ($\beta_x = 2$). Values are the averages over 200 simulation repetitions of ODIWI (red line) and Naïve (green line) estimators; error bars are 95% range. Results are presented for different values of momentum rate $\alpha$ (columns), and the standard deviation $\sigma_\epsilon$ (rows) of the Gaussian kernel: $K(x, x_j) = (2\pi\sigma_k)^{-1}\exp\left\{-(x - x_j)^2/2\sigma_k^2\right\}$.
few iterations, on lower values with tighter intervals, is not sensitive to a specific stratification.

DISCUSSION
Motivated by the calls of leading epidemiologists and exposure scientists such as Diao et al. and Szpiro et al. [12, 13], we tried to improve EE estimates by tailoring predictions to the epidemiological task downstream. We use optimal-design theory to identify observations of importance, and domain-adaptation theory to improve predictions for those observations. Our simulations confirm the validity of this argument: our EE estimates are indeed more accurate than the non-iterative current best practice. An empirical demonstration yields smaller EE estimates, with tighter confidence bands. This is merely a preliminary demonstration, which is promising: the estimator does make a difference, but it does not completely invalidate previous methods and results.

Fig. 4 Densities of $\beta_x - \beta_e$ over 200 simulation repetitions of ODIWI (red lines) and Naïve (green lines) estimators. Results are presented for different values of exposure effect, $\beta_x$ (columns), and the variance of the first-stage model noise, $\sigma^2_e$ (rows).

Fig. 5 ODIWI estimates and 95% bootstrap confidence intervals.
Our exposure predictions are less accurate, on average, compared to non-iterative predictions. The paradox that worse predictions may improve EE estimates was reported, for instance, in ref. [1]. Casting the problem using optimal-design theory, perfectly explains this paradox.

A conceptual difficulty: exposures vary with each new study
The reader may wonder whether it is reasonable to let the exposure, a fixed quantity in reality, to vary from (second-stage) study to study. We argue in favor: allowing the exposure model (and hence the predictions) to vary between studies is not unique to our two-stage setup. For instance, in supervised-learning, the user is free to make different design decisions that will return different predictions. Namely: choosing a loss function, averaging loss versus median loss, etc.

Another example is due to “resolution”: even when global-scale models exist, researchers may prefer country-level predictions. These, implicitly, prioritize certain areas over others.

We thus argue that using different estimators for different tasks is not unprecedented. We merely state explicitly which is the estimation task at hand.

A practical difficulty: iterating between statistician and epidemiologist
Non-iterative two-stage studies are the current standard in EEI. Some of the technical and other barriers that made them so popular no longer exist. Storage, communication and computing technologies are less stringent. Unprecedented information sharing is now possible with cloud technology (e.g., “Copernicus DIAS” service https://www.copernicus.eu/access-data/dias). These improvements make our iterative approach feasible.

It is also possible to alleviate the computational burden by predicting only exposures that are required for each study, instead of predicting an entire spatiotemporal domain. Reducing first-stage’s training set may also be justified.

In addition, our estimator does not require the epidemiologist to share data, which may be sensitive, but only the importance preferences. This ensures that the two-stage decentralized nature is maintained, and eases privacy concerns.

Unprecedented exposures
It may be possible that an optimal design will return exposures that have never been seen in the data. We thus recommend adding a further restriction: that the optimal design be restricted to the convex-hull of data-points.

The number of support points
Choosing the number of support points, J, is a non-trivial matter. A careful discussion is given by Pukelsheim (2006, Section 8.3) [25].

We rely on Fedorov and Leonov (2013) [18] who state a simple and usually satisfied condition under which the maximum number reduces to \( |\Phi| (|\Phi| + 1) / 2 \).

The effect of second-stage covariates
First-stage data does not include personal covariates that are available at the second stage. This means that the optimal design may not depend on such covariates. We thus suggest either ignoring covariates when finding an optimal design, or optimizing predictions for a particular value of the covariates such as their median.

Generalization from pollution to other exposures
In this presentation we focused on satellite measurements of air-pollution and temperature, but the ideas are not limited to those examples. Versions of Algorithm 1 may be applied in any other two-stage study. Two-stage studies are used in environmental epidemiology for other exposures such as NO\(_x\), O\(_3\), air pollen, light at night, etc. Two-stage studies are also used in other epidemiological sub-fields such as nutritional epidemiology, occupational epidemiology, and more [6, 26]. All these fields may gain accuracy by using optimal-design and domain-adaptation theory to guide predictions where they truly matter.

Departing from GLMs
When the second stage is a linear model, our approach can be simplified. Unlike GLMs, in linear models information is maximized when sampling at the boundary of the design space, and independently from the unknown effects. This suggests that for second-stage linear models, it may be enough to provide good predictions for extreme exposure values, without iterating. Other than GLM, optimal-design theory exists for many other nonlinear second-stage models, including survival and longitudinal models, [17, 18].

When to use the estimator?
Two major components are required for our estimator: (1) Second-stage optimal-design theory. (2) First-stage domain-adaptation. The stronger the second-stage non-linearities, then more there is to gain from our iterations.

Caution should be taken if the second-stage model is misspecified, as estimated EE affects predicted exposures and vice-versa. Bias in the second stage may introduce bias in the first stage, a phenomenon known in the EEI literature as feedback [27].

User selected tuning parameters
In this section we discuss some of the design choices that can be made in Algorithm 1.

Initialization.
A natural importance weights initiation is uniform weights: \( \omega^0(x^i) := 1/n \). We did however find that averaging \( \hat{\beta} \) over multiple random initializations of \( \omega^0 \) is beneficial. This is because the two-stage estimator has many degrees of freedom, and is prone to overfitting. To see why this is the case, consider a null effect, \( \beta = 0 \). Because our estimator sequentially updates the design, an error in the initial estimate, \( \beta_1 \), will affect downstream iterations (feedback effect). We found that aggregating multiple initializations alleviates this feedback.

There are many ways to aggregate the multiple \( \hat{\beta} \). We considered two: average estimates after a single iteration (then continue serially), or average after the last iteration. Our simulation results suggest there is no significant difference in the estimation accuracy between the two.

Over-fitting.
Too many iterations can lead to overfitting of the training data. Unlike usual supervised learning problems, where one can hold a validation set to alarm when performance stops improving, in estimation problems, a holdout dataset does not protect from bias.

Yet, there are some choices in the optimization process we can make that reduce overfitting. One such is defining a low learning rate between iterations. A momentum method for \( \beta \) is suggested. Choosing higher \( \alpha \) dampens oscillations, and is more safe, as can be inferred from Fig. 3.

The kernel operator, \( K \), in Eq. (15) is another design choice that governs the tendency to overfit. Our simulation results suggest that the choice of the kernel function is less significant (we compared uniform, Gaussian, and triangle). The kernel’s bandwidth, on the other hand, is more influential. A wider bandwidth means slower convergence, and is more safe in general, as Fig. 3 suggests.

When the choice of number of iterations may be critical, in the spirit of one-step estimators [28], we advocate a single iteration from each initialization \( L = 1 \). Setting \( L = 1 \) may be suboptimal if \( M \) is highly non-linear. The fact that accuracy is improved after a single iteration, is confirmed in our simulation analysis, as can be seen in Fig. 4.
Optimal-design criterion. In Algorithm 1 we used local D-optimality because of its popularity and computational convenience. Our simulation results suggest the estimates are insensitive to the optimality criterion such as E-optimality, A-optimality, etc., but with multivariate exposures this may change.

Prior shift assumption
The prior-shift assumption, \( P_1(r|x) = P_2(r|x) \), means that the distribution of geographical covariates given an exposure, does not vary between first and second stage. Put differently, the difference between the joints is only caused by a change in the exposure marginal distribution. This assumption will often not hold, but Algorithm 1 may still remain useful.

We used the prior-shift assumption to derive the weights in Eq. (14); putting more importance on samples with similar exposures as those of the important samples in the second stage. This was merely a construction device. Our simulation shows that estimates may improve even when this assumption is invalidated. Moreover, a domain-adaptation of the first to the second stage can be achieved without this assumption using other methods [29]. For instance by estimating \( P(r|x) \) in both stages directly, using prior knowledge on the important individuals (e.g., about their residence).

Future research
This work is merely a proof of concept in an attempt to answer the call of Diao et al. and Szpiro et al. [12, 13]. A lot of work is still required on the statistical properties of the proposed estimator, recommended usage, sensitivity analysis of its assumptions, further comparisons with the non-iterative approach, and also with the full two-stage likelihood approach.

DATA AVAILABILITY
The data are not available for replication because of privacy issues.

REFERENCES
1. Montero J-M, Fernández-Avilés G, Mateu J. Spatial and spatio-temporal geostatistical modeling and kriging. Chichester: John Wiley & Sons; 2015.
2. Hodges JS. Richly parameterized linear models: additive, time series, and spatial models using random effects. Boca Raton, FL: CRC Press; 2013.
3. Szpiro AA, Sheppard L, Lumley T. Efficient measurement error correction with spatially misaligned data. Biostatistics. 2011;12:610–23.
4. Shitein A, Kanieli A, Katra I, Raz R, Levy I, Lyapustina A, et al. Estimating daily and intra-daily pm10 and pm2.5 in israel using a spatio-temporal hybrid modeling approach. Atmos Environ. 2018;191:142–52.
5. Sarafian R, Kloog I, Just AC, Rosenblatt JD. Gaussian markov random fields versus linear mixed models for satellite-based pm2.5 assessment: evidence from the Northeastern USA. Atmos Environ. 2019;205:30–35.
6. Szpiro AA, Paciorek CJ. Measurement error in twostage analyses, with application to air pollution epidemiology. Environmetrics. 2013;24:501–17.
7. Carroll RJ, Ruppert D, Stefanski LA. Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. Boca Raton, FL: CRC Press; 2006.
8. Buntin M, Smola A, Huang J, Schmittfull M, Borgwardt K, Schölkopf B. Covariate shift by kernel mean matching. Dataset Shift Mach Learn. 2009;3:5.
9. Spiegelman D. Approaches to uncertainty in exposure assessment in environmental epidemiology. Annu Rev Public Health. 2010;31:149–63.
10. Lopiano KK, Young LJ, Gotway CA. A comparison of errors in variables methods for use in regression models with spatially misaligned data. Stat Methods Med Res. 2011;20:29–47.
11. Just AC, Carli MMD, Shitein A, Dorman M, Lyapustina A, Kloog I. Correcting measurement error in satellite aerosol optical depth with machine learning for modeling pm2.5 in the northeastern usa. Remote Sens. 2018;10:803.
12. Diao M, Holloway T, Chian S, O'Neill SM, Al-Hamdan MZ, Van Donkelaar A, et al. Methods, availability, and applications of pm2.5 exposure estimates derived from ground measurements, satellite, and atmospheric models. J Air Waste Manag Assoc. 2019;69:1391–414.
13. Szpiro AA, Paciorek CJ, Sheppard L. Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology (Camb, MA). 2011;22:880.
14. Just AC, Arfe R, Jish D, Dorman M, Shitein A, Lyapustina A, et al. Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (pm2.5) using satellite data over large regions. Atmos Environ. 2020;239:117649.
15. Park Y, Kwon B, Heo J, Hu X, Liu Y, Moon T. Estimating pm2.5 concentration of the conterminous united states via interpretable convolutional neural networks. Environ Pollut. 2020;256:113395.