**Comparision of usage of individual and semi-summarized date in estimation of a nonlinear exposure-outcome relationship using instrumental variables**

Amy M. Mason, PhD1,2, Stephen Burgess, PhD3, 4

**Affiliations**

1 British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

2 National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge, UK

3 Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

4 MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

Corresponding author: Amy M. Mason, Strangeways Research Laboratory, Department of Public Health and Primary Care, Cardiovascular Epidemiology Unit, University of Cambridge, Wort’s Causeway, Cambridge CB18RN, UK.  
 Tel: [44-7825299644](mailto:44-7825299644); E-mail: [am2609@medschl.cam.ac.uk](mailto:am2609@medschl.cam.ac.uk)

**Words of abstract**: / **main text**:

**Numbers of tables**: / **figures**:

**Supplementary tables**: / **figure**:

**Numbers of references**:

**Funding source**: AMM is supported by EC‐Innovative Medicines Initiative (BigData@Heart). SB is supported by Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (204623/Z/16/Z).

**Author contributions:** Study conception and design: AMM & SB; data creation and analysis: AMM; drafting the manuscript and figures: AMM; reviewing the manuscript: SB & AMM

**Acknowledgments:**

**Data availability:**

**Conflict of interests:** All authors have no conflict of interest to disclose.

**Abstract**

Mendelian randomization uses instrumental variable analysis with genetic variants to determine the causal effect of an exposure on an outcome. Semiparametric methods have been designed to allow the estimation of nonlinear relationships between the exposure and the outcome, but these require individual level data. However, genetic data is often only published in summarized forms reducing the usability of these methods.   
Where data can be obtained in a semi-summarized form, with localized genetics effects on the exposure and outcome for each stratum of the population, then an adapted fractional polynomial method can be applied without need of individual level data.   
We demonstrate in a simulation study that use of semi-summarized data estimates the true exposure-outcome relationship with similar accuracy to the original fractional polynomial method.

**Keywords:** causal effects, fractional polynomials, genetic variants, mendelian randomization,

**Introduction**

The assumption of a linear relationship between exposure and outcome is typical in Mendelian randomization. Yet the true relationship between exposure and outcome of interest may be nonlinear, such as in the case of body mass index (BMI) and all-cause mortality (Sun, et al., 2019). Approaches using semiparametric methods for testing and estimating non-linear effects have been tested for individual level data (Staley & Burgess, 2017). One of these methods was the fractional polynomial methods, which performed a meta regression of the localised average causal effect (LACE) in strata of the population determined by their exposure distribution. The major disadvantage of this method was that it needs access to the full individual dataset of genetic values, outcome and exposure for each member of the population. This prevents its wider usage as a method as the sharing of individual data is fraught with complexities of data security and privacy concerns.

This paper tests the proposal that the fractional polynomial model could instead be applied to “semi-summarised” data; using the individual level data to calculate the genetic associations for the exposure and outcome data in the strata of the population and then using only that semi-summarised data, forming the LACE estimates and fitting the fractional polynomial. This would have the advantage of MR researchers being able to request and share the semi-summarized form of the data without sharing any one individual’s data and yet still being able to test and fit an accurate non-linear model.

**Methods**

Given an exposure X, an outcome Y and single genetic instrument G, we wish to investigate the shape of the exposure-outcome relationship. We will assume that this relationship is homogeneous in the population.

**Semi-summarized data and LACE estimates**

We will break the population up into strata based on the exposure distribution. We cannot stratify on the exposure X directly, as it is on the causal pathway from the genetic instrument G to the outcome Y by assumption (Sheehan & Didelez, 2020) and could induce an association between G and Y violating the exclusion restriction criteria.   
Instead we need the residual or “IV-free” value of the exposure – the expected value of exposure for that individual if they had no exposure-increasing alleles. This is the difference between the measured value and the fitted value obtained from regressing on the genetic instrument. These values allow us to stratify the population as if they all had the same genetic background for the exposure, and cannot induce a relationship between G and Y. The people in the same strata would, if they had had the same genotype, have similar values of the exposure.

Within each stratum, we require the mean of the exposure , and the beta coefficient and standard deviation for the regression of X on G and Y on G (called and ). This data is all that needs to be shared to create the LACE estimates – localised average causal effects, calculated as a standard linear mendelian randomisation Wald estimate for each strata: the genetic association with the outcome divided by the genetic association with the exposure.

Table 1 Example of semi-summarised data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Strata** |  |  | s.d() |  | s.d() |
| **1** | **2.45** | **0.253** | **0.006** | **0.154** | **0.049** |
| **2** | **2.75** | **0.227** | **0.004** | **0.249** | **0.050** |
| **3** | **2.97** | **0.226** | **0.003** | **0.205** | **0.048** |
| **4** | **3.14** | **0.234** | **0.002** | **0.197** | **0.048** |
| **…** |  |  |  |  |  |

**Fractional polynomials**

**SIMULATION STUDY**

To assess the performance of these methods in realistic scenarios

for Mendelian randomization, we performed a simulation

study.We simulated data for 10,000 individuals for an IV

*𝐺*, a continuous exposure *𝑋* that takes only positive values, a

continuous outcome *𝑌* , and a confounder *𝑈* (assumed to be

unmeasured). The data-generating model for individual *𝑖* is

*𝑥𝑖* = 2+0*.*25*𝑔𝑖* + *𝑢𝑖* + *𝜖𝑋𝑖*

*𝑦𝑖* = *ℎ*(*𝑥𝑖*) + 0*.*8*𝑢𝑖* + *𝜖𝑌 𝑖,*

where *𝑔𝑖* ∼ Bin(2, 0.3), *𝜖𝑋𝑖* ∼ Exp(1), *𝑢𝑖* ∼ Unif(0,1), *𝜖𝑌 𝑖* ∼

*𝑁*(0*,* 1), and *ℎ*(*𝑥𝑖*) is the function relating the exposure to

the outcome (the exposure-outcome relationship). Exposure

values were taken to be positive and away from zero so

that the outcome takes sensible values for log and negative

power functions. The IV explains 2.6% of the variance in the

exposure.

**3.1 Choice of exposure-outcome model**

For the fractional polynomial method, all possible fractional

polynomials of degrees 1 and 2 were considered as the functional

form of the exposure-outcome relationship. Combinations of effect sizes for the *𝛽* parameters were chosen

ranging from 0 to 2. For fractional polynomials of degree

2, we also considered effects in opposing directions for *𝛽*1

and *𝛽*2; these simulations yielded similar results to those discussed

here (results not shown). Fixed-effects metaregression

was used in the simulations, however, random-effects metaregression

yielded similar results (results not shown).

For the piecewise linear method and comparisons between

methods, linear, quadratic, square-root, and logarithmic

functions were considered as the functional form of the

exposure-outcome relationship, as well as a threshold

model:

*ℎ*(*𝑥𝑖*) =

{

0 if *𝑥𝑖* ≤ 3*.*65

*𝛽𝑥𝑖* if *𝑥𝑖 >* 3*.*65*.*

**3.2 Evaluating the performance of the**

**methods**

To evaluate the fractional polynomial method, we first fitted

the correct fractional polynomial model (i.e., with the correct

degree and powers) and assessed the bias and coverage

of the effect parameter estimates. Subsequently, we fitted all

fractional polynomials of the same degree and selected the

best-fitting polynomial based on the likelihood. We assessed

the proportion of simulations where the best-fitting fractional

polynomial was the correct fractional polynomial. If the correct

fractional polynomial was not the best-fitting fractional

polynomial, we tested whether itwas in the group of fractional

polynomials that fit the data almost as well as the best-fitting

polynomial; defined as those fractional polynomials where

twice the difference in the log-likelihood (compared with the

best-fitting polynomial)was less than the 90th percentile point

of a *𝜒*2

*𝑚*

distribution, where *𝑚* = 1 for comparing fractional

polynomials of degree 1 and *𝑚* = 2 for comparing polynomials

of degree 2.

For comparing the fit of the fractional polynomial and

piecewise linear models, we used the following heuristic

function:

Σ*𝐾*

*𝑘*=1

||

*̂̄*

*𝑦𝑘* − *𝑦̄𝑘*

||

*,* (3)

where summation is across the *𝐾* quantile groups, and *𝑦̄𝑘* is

the expected value of the outcome evaluated at themean value

of the exposure in each quantile group.

**3.3 Varying the number of strata**

In the initial simulations, the population was split based on

the IV-free exposure into decile groups. Further simulations

**3.4 Additional simulations to assess impact of**

**violations of assumptions**

We performed additional simulations in which the underlying

assumptions that the effect of the IV on the exposure and

the effect of the exposure on the outcome are fixed and independent

were relaxed. In these simulations, we assessed both

modeling assumptions by allowing the effect of the IV on the

exposure to vary (by drawing the effect parameter from a normal

distribution *𝑁*(0*.*25*,* 0*.*12) for each individual in the population),

and allowing the exposure-outcome relationship to

vary (by drawing the causal parameter from a normal distribution

*𝑁*(*𝛽,* 0*.*22) for each individual in the population). We

assessed the impact of allowing each of these parameters to

vary separately and both to vary together. In addition, we also

allowed variation in both parameters to be correlated by drawing

the parameters from a bivariate normal distribution with

correlation 0.2. For fractional polynomials of degree 2, only

the causal parameter for the second polynomial was allowed

to vary across individuals.

We also performed further simulations using a lowfrequency

genetic variant having a large effect on the exposure

(minor allele frequency = 0.03, linear effect on the exposure

= 0.75), and using the original genetic variant but having a

superadditive (first allele increases exposure by 0.1 units, second

by 0.3 units) and a subadditive (first allele increases exposure

by 0.3 units, second by 0.1 units) effect on the exposure

in the data-generating model.

**Methods**

Heuristic statistic

we used the following heuristic

function:

Σ*𝐾*

*𝑘*=1

||

*̂̄*

*𝑦𝑘* − *𝑦̄𝑘*

||

*,* (3)

where summation is across the *𝐾* quantile groups, and *𝑦̄𝑘* is

the expected value of the outcome evaluated at the mean value

of the exposure in each quantile group.

The heuristic statistic (mean (SD) across simulations) is the sum of the absolute values of the predicted value of the outcome minus the correct value of the outcome at the mean value of the exposure in deciles of the IV-free distribution

**Results**

Table 2 Simulation results for degree 1 polynomials

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Fitting correct FP | | | | Fitting all FPs | | | |
|  |  | Individual | | Semi-summarized | | Individual | | Semi-summarized | | |
|  |  | Mean (SD)  [Mean S.E] | Coverage | Mean (SD)  [Mean S.E] | Coverage | Powers | | Powers | | |
|  |  |  |  |  |  | Correct | Set | Correct | Set | |
| 0 | 0 | 0.00 (0.22) [0.21] | 0.95 | 0.00 (0.22) [0.21] | 0.95 | - | - | - | - | |
| 0 | 1 | 0.99 (0.21) [0.21] | 0.94 | 0.99 (0.21) [0.21] | 0.94 | 0.18 | 0.91 | 0.18 | 0.91 | |
| 0 | 2 | 1.97 (0.22) [0.21] | 0.95 | 1.97 (0.22) [0.21] | 0.95 | 0.36 | 0.90 | 0.36 | 0.90 | |
| 0.5 | 0 | -0.01 (0.23) [0.23] | 0.96 | -0.01 (0.23) [0.23] | 0.96 | - | - | - | - | |
| 0.5 | 1 | 0.99 (0.24) [0.24] | 0.94 | 0.99 (0.24) [0.23] | 0.94 | 0.17 | 0.91 | 0.17 | 0.91 | |
| 0.5 | 2 | 1.99 (0.23) [0.24] | 0.95 | 1.99 (0.23) [0.24] | 0.95 | 0.41 | 0.92 | 0.41 | 0.92 | |
| 1 | 0 | -0.01 (0.06) [0.06] | 0.95 | -0.01 (0.06) [0.06] | 0.95 | - | - | - | - | |
| 1 | 1 | 1.00 (0.07) [0.06] | 0.95 | 1.00 (0.07) [0.06] | 0.95 | 0.78 | 0.95 | 0.78 | 0.95 | |
| 1 | 2 | 2.00 (0.07) [0.07] | 0.94 | 2.00 (0.07) [0.07] | 0.94 | 0.93 | 0.97 | 0.93 | 0.97 | |
| 2 | 0 | 0.00 (0.01) [0.01] | 0.95 | 0.00 (0.01) [0.01] | 0.95 | - | - | - | - | |
| 2 | 1 | 1.02 (0.01) [0.01] |  | 1.02 (0.01) [0.01] | 0.74 | 1.00 | 1.00 | 1.00 | 1.00 | |
| 2 | 2 | 2.03 (0.02) [0.01] | 0.44 | 2.03 (0.01) [0.01] | 0.44 | 1.00 | 1.00 | 1.00 | 1.00 | |

*Notes*. Results for all the fractional polynomials of degree 1 are presented in Tables S1; this table is a summary of

results for the most commonly encountered powers. *𝑝* are the powers and *𝛽* are the effect parameters. Coverage refers to the proportion of replications where the true value of *𝛽* was contained within the corresponding 95% confidence interval. Correct refers to the proportion of replications where the best-fitting fractional polynomial was also the correct fractional polynomial. Set refers to the proportional where the twice the difference of log-likelihood of the model fit with correct fractional polynomial and the best-fitting fractional was less than the 90th percentile of the relevant *𝜒*2(1) distribution. SD, standard deviation; SE standard error; FP, fractional polynomial.

Supplementary

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Fitting correct FP | | | | Fitting all FPs | | | |
|  |  | Individual | | Semi-summarized | | Individual | | Semi-summarized | | |
|  |  | Mean (SD)  [Mean S.E] | Coverage | Mean (SD)  [Mean S.E] | Coverage | Powers | | Powers | | |
|  |  |  |  |  |  | Correct | Set | Correct | Set | |
| 0 | 0 | 0.00 (0.22) [0.21] | 0.95 | 0.00 (0.22) [0.21] | 0.95 | - | - | - | - | |
| 0 | 1 | 0.99 (0.21) [0.21] | 0.94 | 0.99 (0.21) [0.21] | 0.94 | 0.18 | 0.91 | 0.18 | 0.91 | |
| 0 | 2 | 1.97 (0.22) [0.21] | 0.95 | 1.97 (0.22) [0.21] | 0.95 | 0.36 | 0.90 | 0.36 | 0.90 | |
| 0.5 | 0 | -0.01 (0.23) [0.23] | 0.96 | -0.01 (0.23) [0.23] | 0.96 | - | - | - | - | |
| 0.5 | 1 | 0.99 (0.24) [0.24] | 0.94 | 0.99 (0.24) [0.23] | 0.94 | 0.17 | 0.91 | 0.17 | 0.91 | |
| 0.5 | 2 | 1.99 (0.23) [0.24] | 0.95 | 1.99 (0.23) [0.24] | 0.95 | 0.41 | 0.92 | 0.41 | 0.92 | |
| 1 | 0 | -0.01 (0.06) [0.06] | 0.95 | -0.01 (0.06) [0.06] | 0.95 | - | - | - | - | |
| 1 | 1 | 1.00 (0.07) [0.06] | 0.95 | 1.00 (0.07) [0.06] | 0.95 | 0.78 | 0.95 | 0.78 | 0.95 | |
| 1 | 2 | 2.00 (0.07) [0.07] | 0.94 | 2.00 (0.07) [0.07] | 0.94 | 0.93 | 0.97 | 0.93 | 0.97 | |
| 2 | 0 | 0.00 (0.01) [0.01] | 0.95 | 0.00 (0.01) [0.01] | 0.95 | - | - | - | - | |
| 2 | 1 | 1.02 (0.01) [0.01] | 0.73 | 1.02 (0.01) [0.01] | 0.74 | 1.00 | 1.00 | 1.00 | 1.00 | |
| 2 | 2 | 2.03 (0.02) [0.01] | 0.44 | 2.03 (0.01) [0.01] | 0.44 | 1.00 | 1.00 | 1.00 | 1.00 | |
| 3 | 0 | 0.00 (0.00)[0.00] | 0.95 | 0.00 (0.00)[0.00] | 0.95 |  |  |  |  | |
| 3 | 0 |  | 0.95 |  | 0.00 |  |  |  |  | |
| 3 | 1 | 1.03 (0.01) [0.01] | 0.00 | 1.03 (0.00) [0.01] | 0.00 |  |  |  |  | |
| 3 | 1 |  | 0.00 |  | 0.00 |  |  |  |  | |
| 3 | 2 | 2.07 (0.01) [0.01] | 0.00 | 2.07 (0.01) [0.01] | 0.00 |  |  |  |  | |
| 3 | 2 |  | 0.00 |  | 0.97 |  |  |  |  | |
| -0.5 | 0 | -0.03 (0.73) [0.74] | 0.97 | -0.03 (0.73) [0.74] | 0.97 |  |  |  |  | |
| -0.5 | 0 |  | 0.97 |  | 0.93 |  |  |  |  | |
| -0.5 | 1 | 1.04 (0.77) [0.75] | 0.93 | 1.04 (0.76) [0.74] | 0.93 |  |  |  |  | |
| -0.5 | 1 |  | 0.93 |  | 0.93 |  |  |  |  | |
| -0.5 | 2 | 1.99 (0.79) [0.75] | 0.93 | 1.99 (0.79) [0.75] | 0.93 |  |  |  |  | |
| -0.5 | 2 |  | 0.93 |  | 0.95 |  |  |  |  | |
| -1 | 0 | -0.04 (0.67) [0.65] | 0.95 | -0.04 (0.67) [0.65] | 0.95 |  |  |  |  | |
| -1 | 0 |  | 0.95 |  | 0.95 |  |  |  |  | |
| -1 | 1 | 0.99 (0.64) [0.64] | 0.95 | 0.99 (0.64) [0.64] | 0.95 |  |  |  |  | |
| -1 | 1 |  | 0.95 |  | 0.93 |  |  |  |  | |
| -1 | 2 | 1.96 (0.67) [0.64] | 0.93 | 1.96 (0.67) [0.64] | 0.93 |  |  |  |  | |
| -1 | 2 |  | 0.93 |  | 0.92 |  |  |  |  | |
| -2 | 0 | -0.10 (1.01) [0.93] | 0.92 | -0.10 (1.01) [0.93] | 0.92 |  |  |  |  | |
| -2 | 0 |  | 0.92 |  | 0.95 |  |  |  |  | |
| -2 | 1 | 1.01 (0.97) [0.93] | 0.95 | 1.01 (0.97) [0.93] | 0.95 |  |  |  |  | |
| -2 | 1 |  | 0.95 |  | 0.97 |  |  |  |  | |
| -2 | 2 | 1.89 (0.93) [0.93] | 0.97 | 1.89 (0.93) [0.93] | 0.97 |  |  |  |  | |
| -2 | 2 |  | 0.97 |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |
|  |  |  |  |  |  |  |  |  |  | |

is the power, and is the effect parameter. Coverage refers to the proportion of replications where the true was contained within the corresponding 95% CI – note that for FP with the coverage is zero due to the excessively small confidence intervals, the actual point estimates are close to the genuine value. Correct refers to the proportion of replications where the best-fitting fractional polynomial was also the correct fractional polynomial. Set refers to the proportional where the twice the difference of log-likelihood of the model fit with correct fractional polynomial and the best-fitting fractional was less than the 90th percentile of the relevant *𝜒*2(1) distribution. SD, standard deviation; SE standard error; FP, fractional polynomial.