Engaging Biostatistics in Maternal and Birth Outcome Research

Yan Yuan, PhD
School of Public Health
University of Alberta

April 4, 2018

Biostatistical Research and Collaboration

- Developing new biostatistical methodology
  - Risk prediction performance measures
  - Prediction algorithms
  - Trajectory modelling
- Applying biostatistical methods in health research, particularly cancer research using administrative data
- Providing biostatistical support to health researchers

Received: 3 September 2016 | Revised: 24 September 2017 | Accepted: 14 December 2017

DOI: 10.1002/sim.7606

RESEARCH ARTICLE

WILEY Statistics in Medicine

A threshold-free summary index of prediction accuracy for censored time to event data

Yan Yuan[1] | Qian M. Zhou[2,3] | Bingying Li[3] | Hengrui Cai[1] | Eric J. Chow[4] | Gregory T. Armstrong[5]

[1] School of Public Health, University of Alberta, Edmonton, AB T6G1C9, Canada

[2] Department of Mathematics and Statistics, Mississippi State University, Starkville, Mississippi 39762, USA

[3] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C. V5A1S6, Canada

[4] Fred Hutchinson Cancer Research Center, Seattle Children's Hospital, University of Washington, Seattle, Washington, USA

[5] Department of Epidemiology and Cancer Control, St. Jude Children's Research

Prediction performance of a risk scoring system needs to be carefully assessed before its adoption in clinical practice. Clinical preventive care often uses risk scores to *screen* asymptomatic population. The primary clinical interest is to predict the risk of having an event by a prespecified *future* time t_0. Accuracy measures such as positive predictive values have been recommended for evaluating the predictive performance. However, for commonly used continuous or ordinal risk score systems, these measures require a subjective cutoff threshold value that dichotomizes the risk scores. The need for a cutoff value created barriers for practitioners and researchers. In this paper, we propose a threshold-free summary index of positive predictive values that accommodates time-dependent event status and competing risks. We develop a nonparametric estimator and

JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

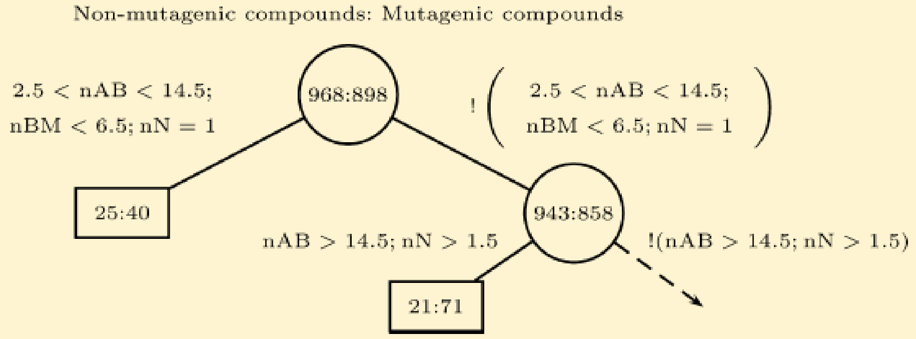
Article

pubs.acs.org/jcim

Harvesting Classification Trees for Drug Discovery

Yan Yuan,*[†] Hugh A. Chipman,[‡] and William J. Welch[§]

[†]Department of Public Health Sciences, University of Alberta, Edmonton, Alberta T6G 1C9, Canada
[‡]Department of Mathematics and Statistics, Acadia University, Wolfville, Nova Scotia B4P 2R6, Canada
[§]Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada



ABSTRACT: Millions of compounds are available as potential drug candidates. High throughput screening (HTS) is widely used in drug discovery to assay compounds for a particular biological activity. A common approach is to build a classification model using a smaller sample of assay data to predict the activity of unscreened compounds and hence select further compounds for assay. This improves the efficiency of the search by increasing the proportion of hits found among the assayed compounds. In many assays, the biological activity is dichotomized into a binary indicator variable; the explanatory variables are chemical descriptors capturing compound structure. A tree model is interpretable, which is key, since it is of interest to identify diverse chemical classes among the active compounds to serve as leads for drug optimization. Interpretability of a tree is often reduced, however, by the sheer size of the tree model and the number of variables and rules of the terminal nodes. We develop a "tree harvesting" algorithm to filter out redundant "junk" rules from the tree while retaining its predictive accuracy. This simplification can facilitate the process of uncovering key relations between molecular structure and activity and may clarify rules defining multiple activity mechanisms. Using data from the National Cancer Institute, we illustrate that many of the rules used to build a classification tree may be redundant. Unlike tree pruning, tree harvesting allows variables with junk rules to be removed near the top of the tree. The reduction in complexity of

1. Modelling Gestational Weight Trajectory

Maternal Weight Gain

- Adverse maternal outcomes
  - Obesity
  - C-section
  - Gestational hypertension
- Adverse birth outcomes
  - Small for gestation age
  - Large for gestation age

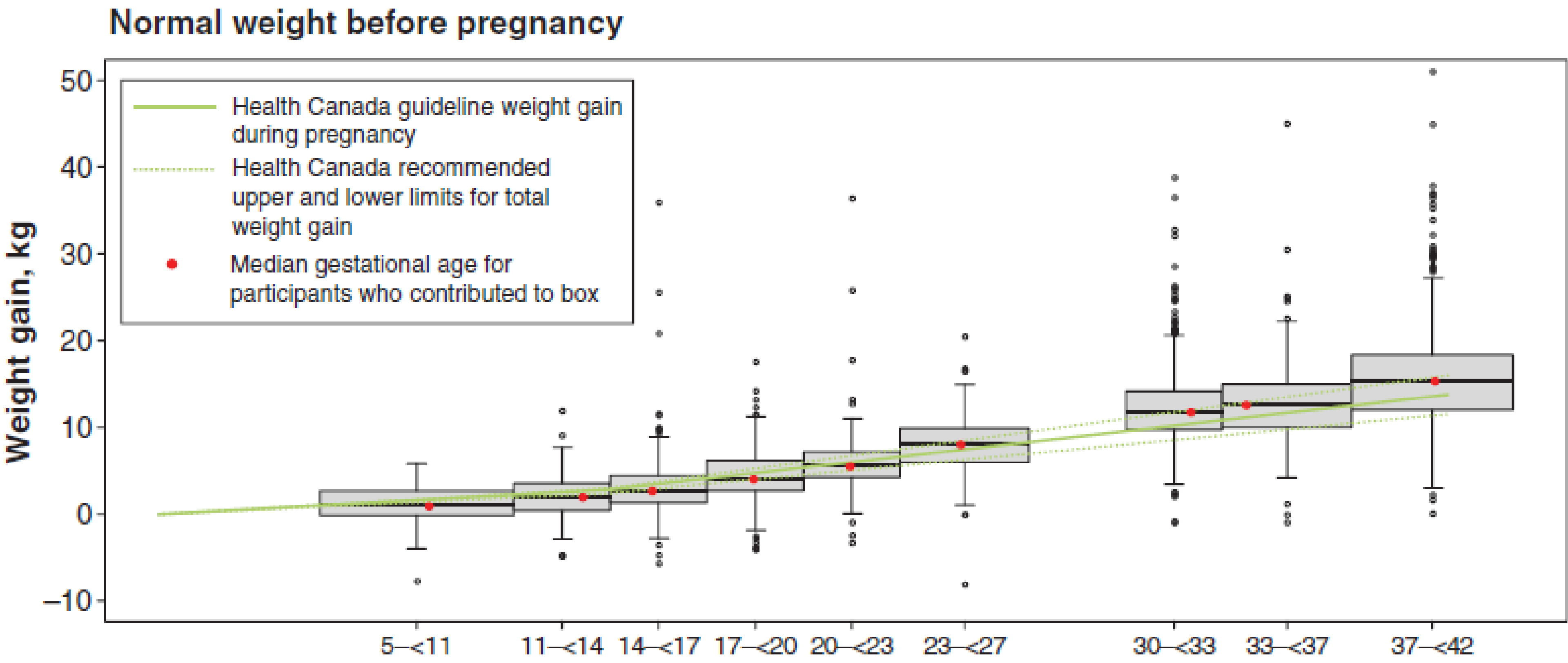
APrON study

Study objective

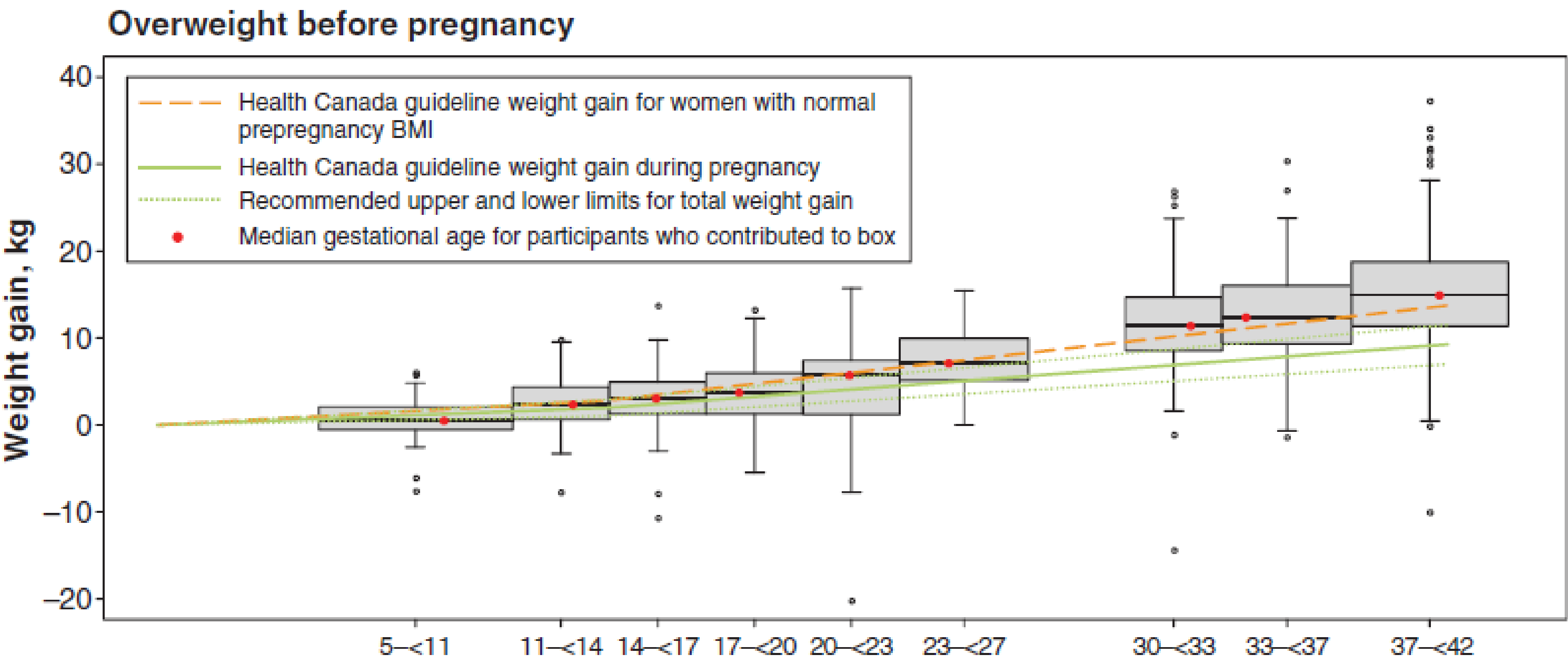
- Comprehensive assessment of maternal and offspring well-being, identification of risk factors prior to and during pregnancy and post-partum for adverse outcomes.

Study cohort

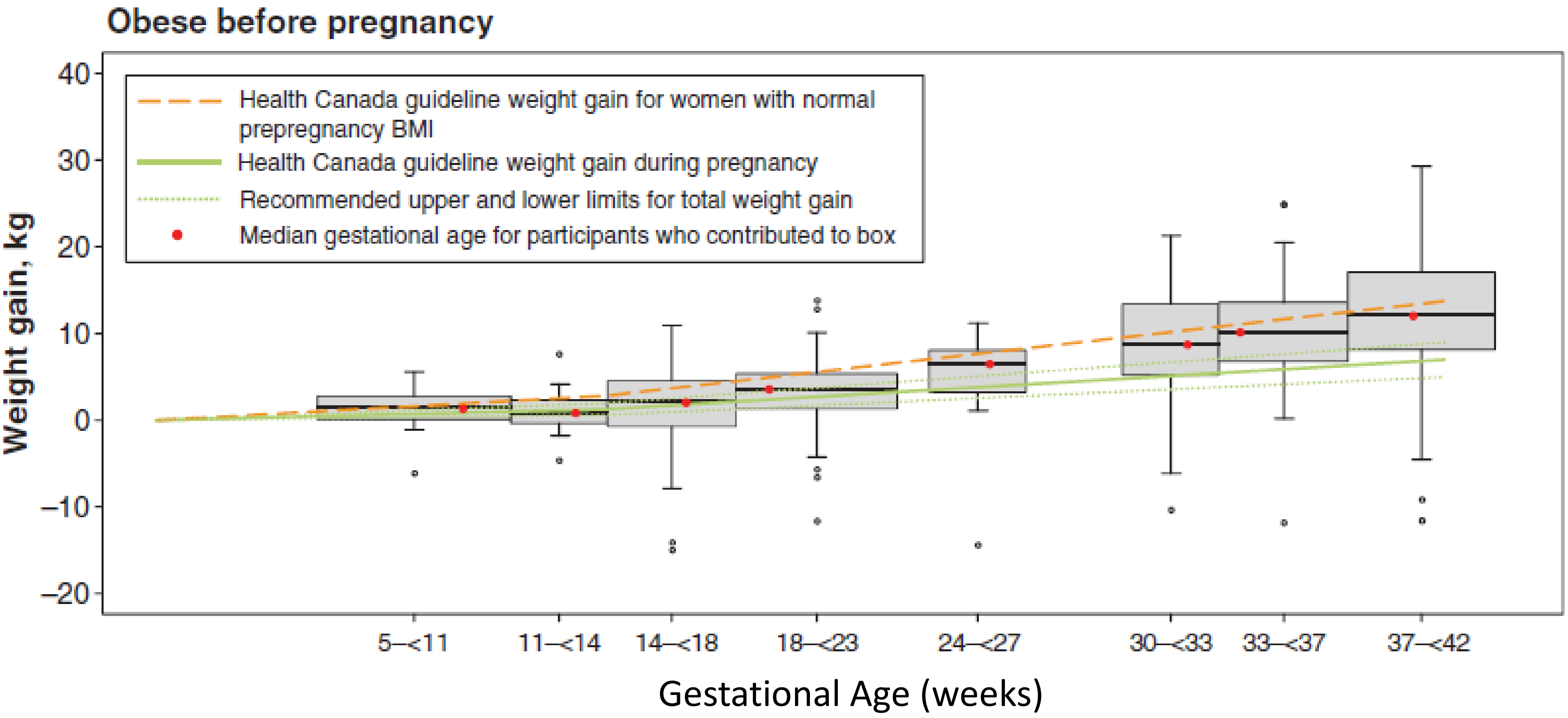
- A prospective cohort of 2189 adolescents and women and their infants during pregnancy and post-partum in Edmonton and Calgary.



Gestational Age (weeks)



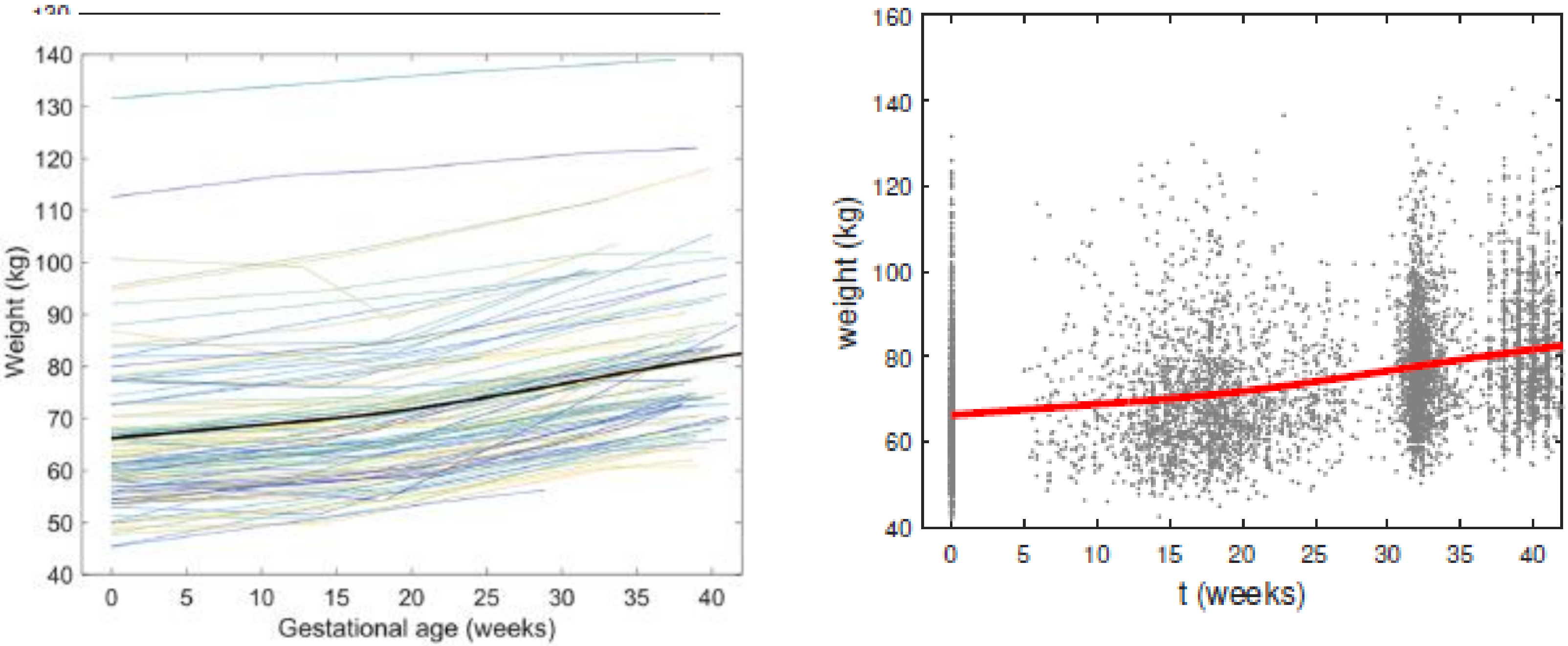
Gestational Age (weeks)



Obese before pregnancy

Jarman M. et al. (2016) *Canadian Medical Association Journal Open* 4(2):E338-345.
PMID: 27525254

Individual Trajectory



(a) (b)

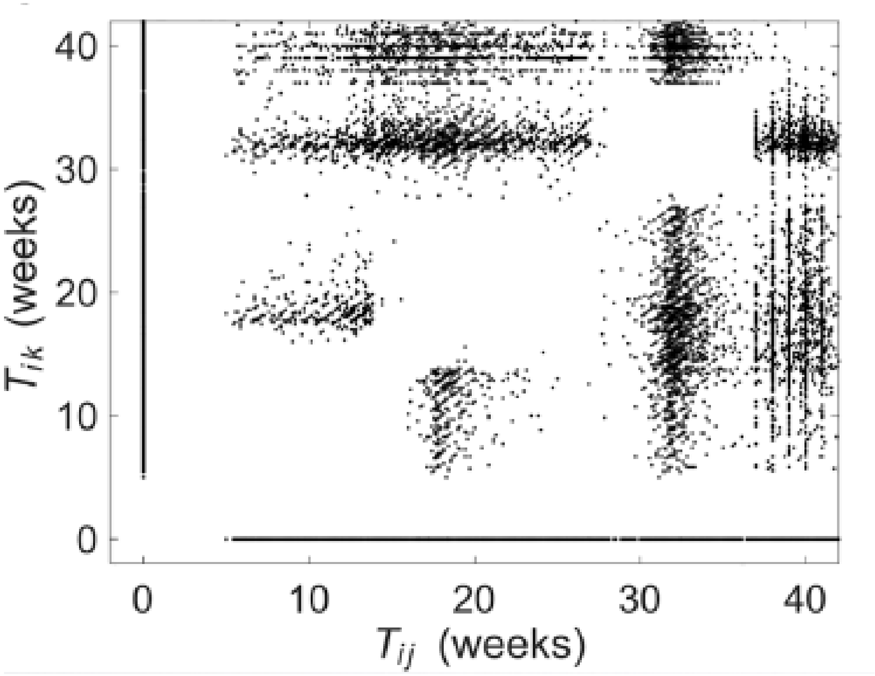
Figure: (a) Observed individual weight trajectories of randomly selected 100 subjects, overlaid with the smooth estimate of the mean function and (b) All the weight records overlaid with the smooth estimate of the mean function

Che M. *et al. (2017) PLoS One.* 12(10): e0186761

Traditional approaches and why they don’t work well

- Non-linear mixed model
- Longitudinal model

Challenges: Sparse irregularly-spaced data



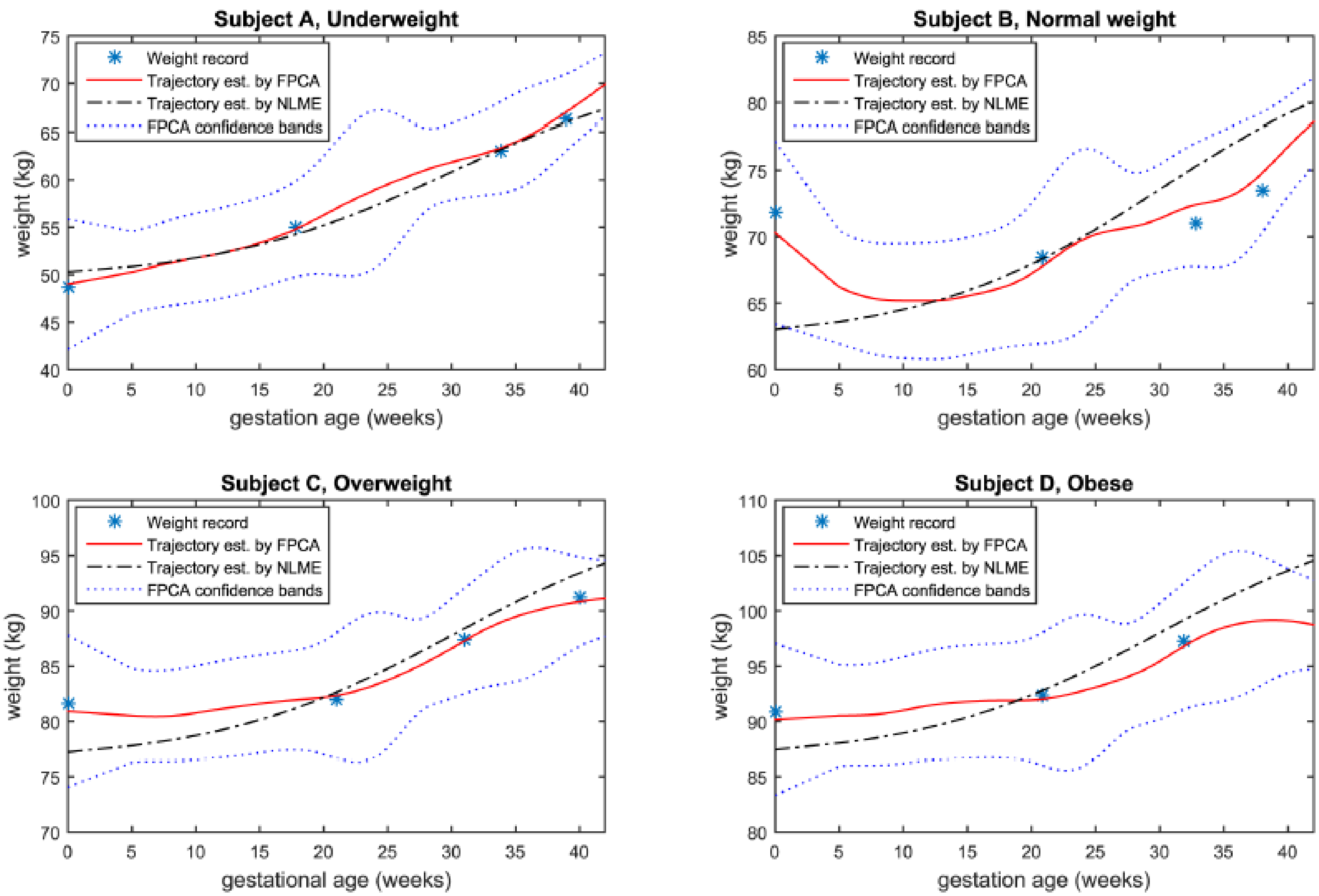


Figure: Predicted trajectories and confidence bands of the weight measurements of 4 random subjects

Functional Principal Component Analysis approach

$$Y_{ij} = X_i(T_{ij}) = \mu(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(T_{ij}) + \epsilon_{ij}.$$

$$\hat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_k(t).$$

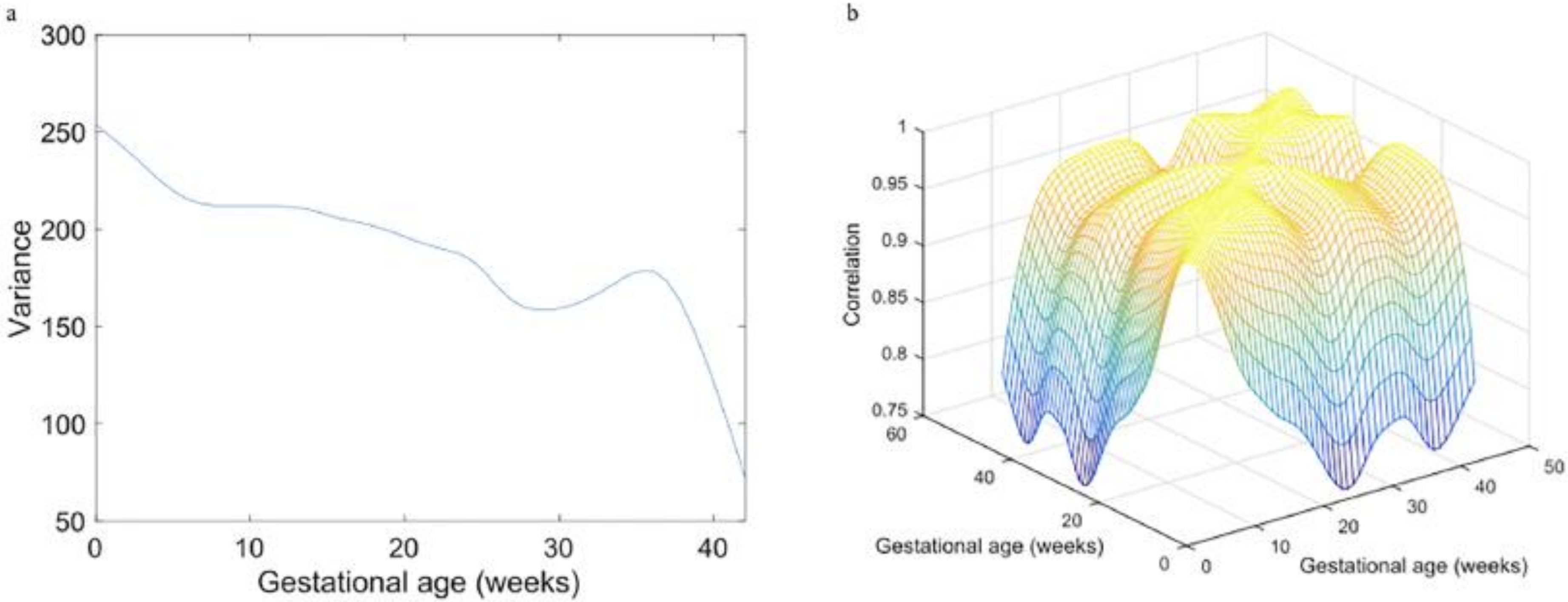


Fig 2. (a) Smooth estimate of the variance function of the weight data; (b) Smooth estimate of the correlation surface.

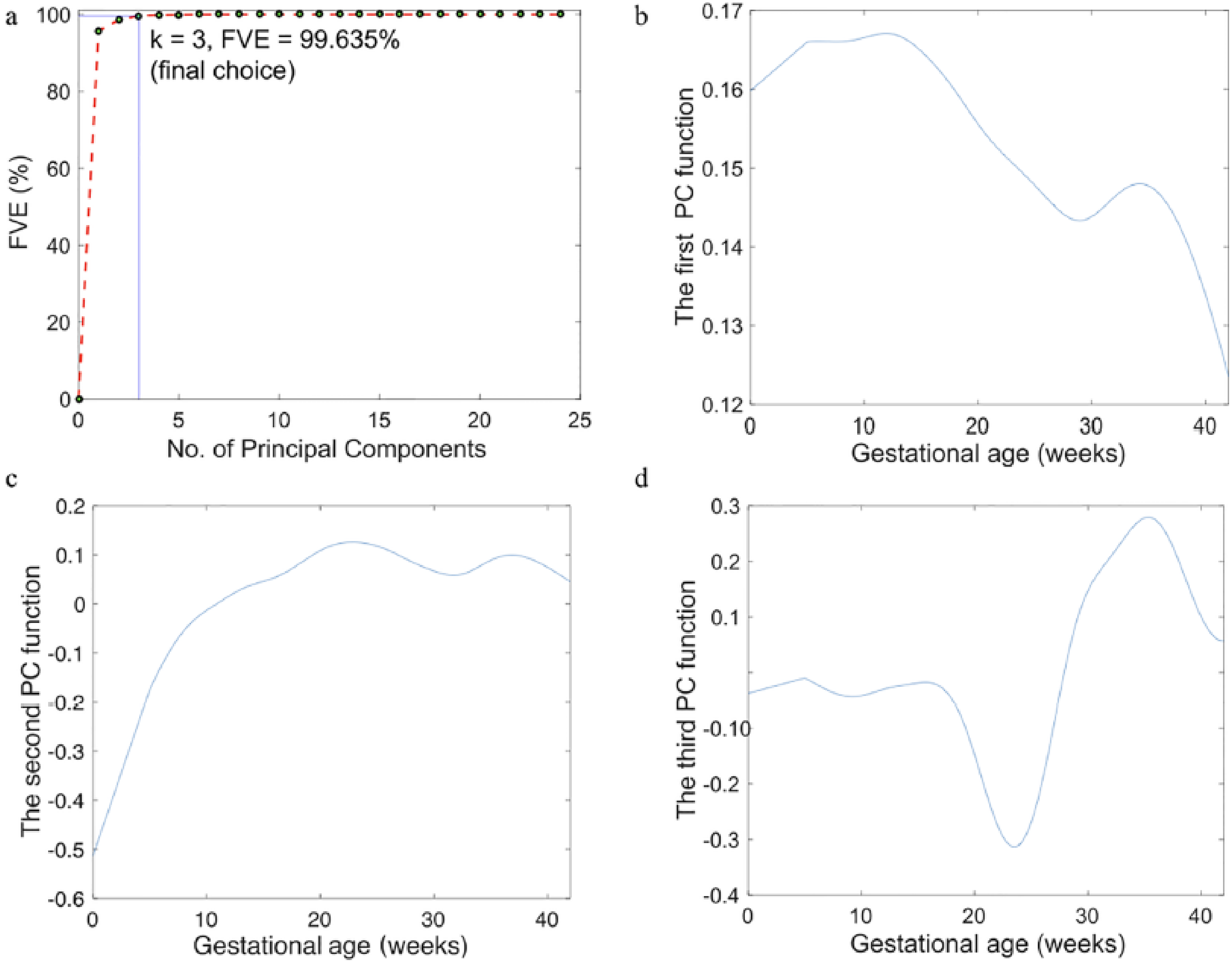
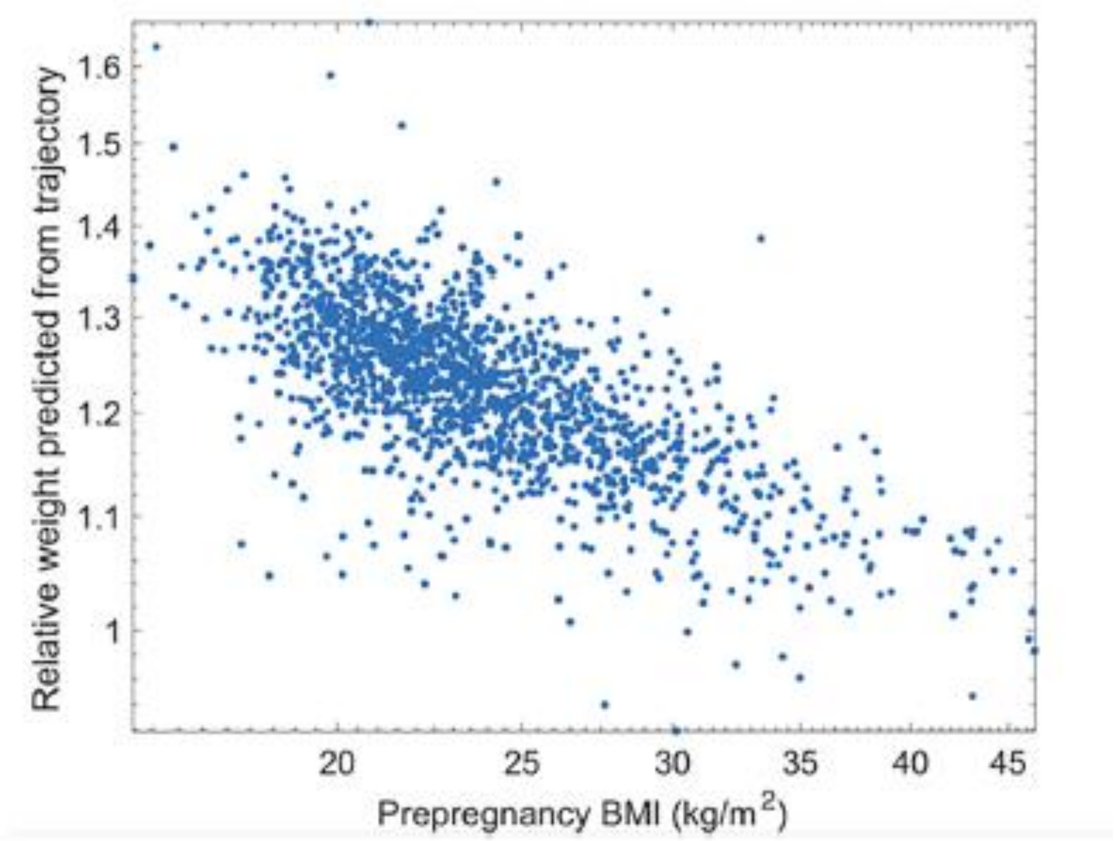


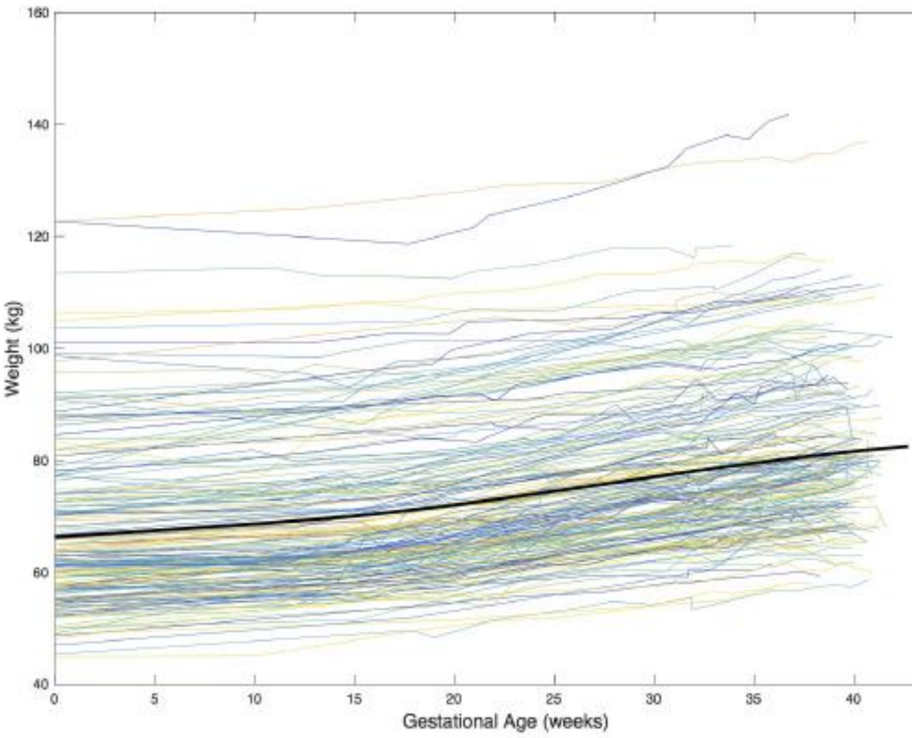
Fig 3. (a) Scree plot of the weight data and (b–d) The first, second and third principle component (PC) functions for the weight data which account for 95.7%, 2.8%, and 1.1% of the total variation, respectively.

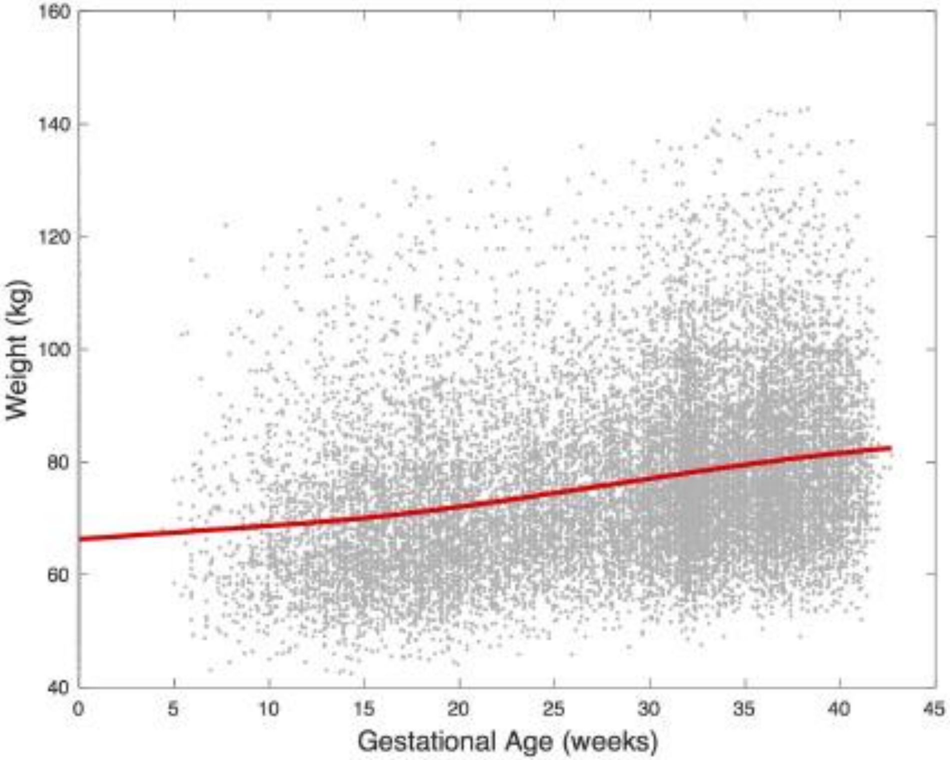
Modelling the total weight gain



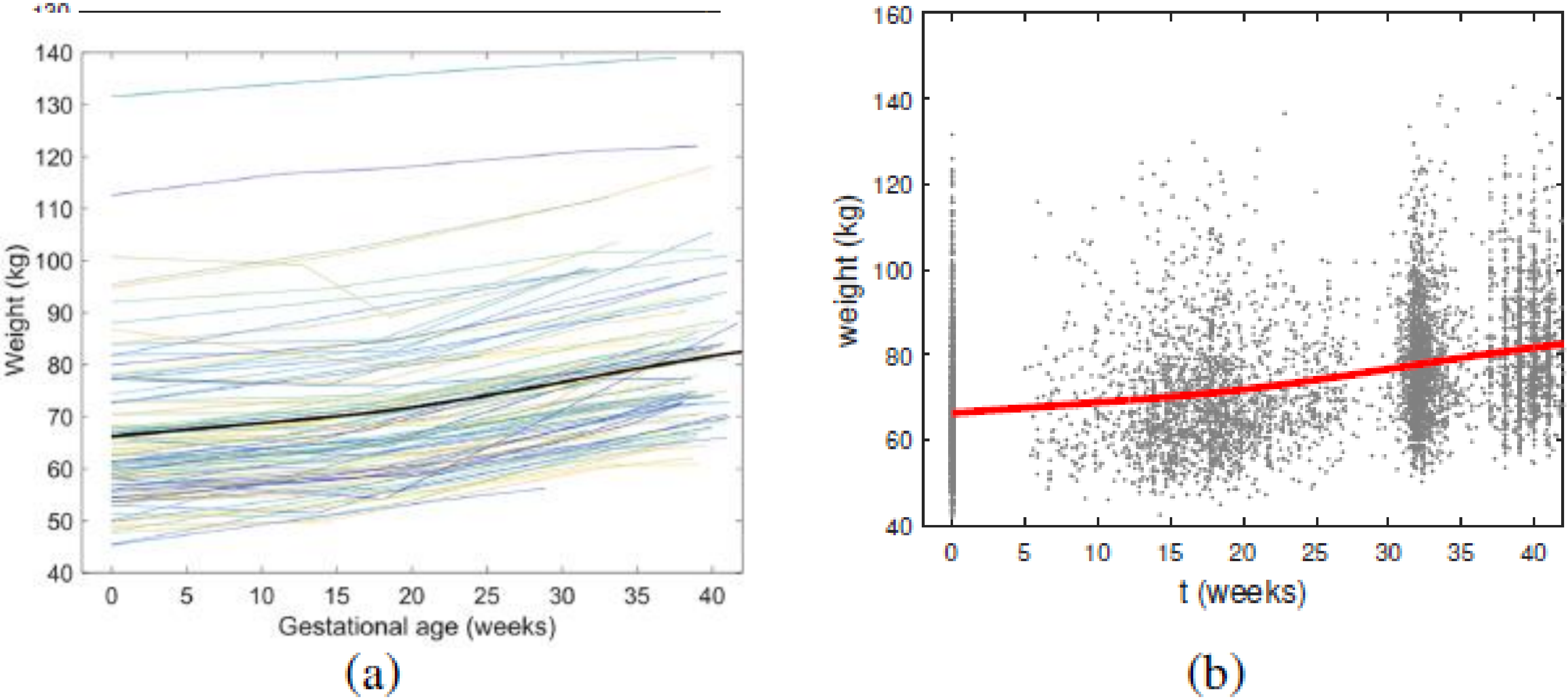
E[Log(weight at delivery/weight at pre-pregnancy)] = $\beta_0 + \beta_1$BMI
BMI alone accounted for 50% variance in Δlog(weight).

APrON + clinical weight data





APrON weight data



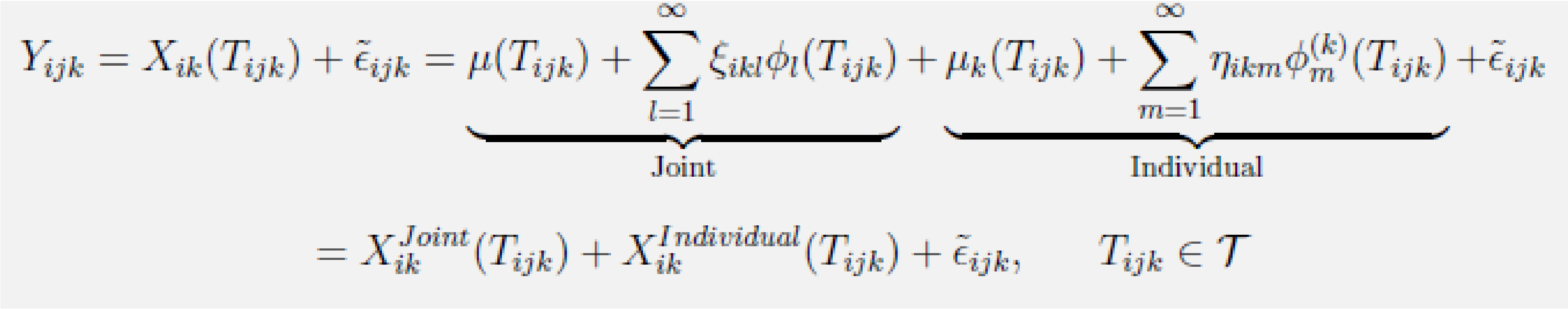
(a)

(b)

Che M. *et al. (2017) PLoS One.* 12(10): e0186761

Modifying FPCA

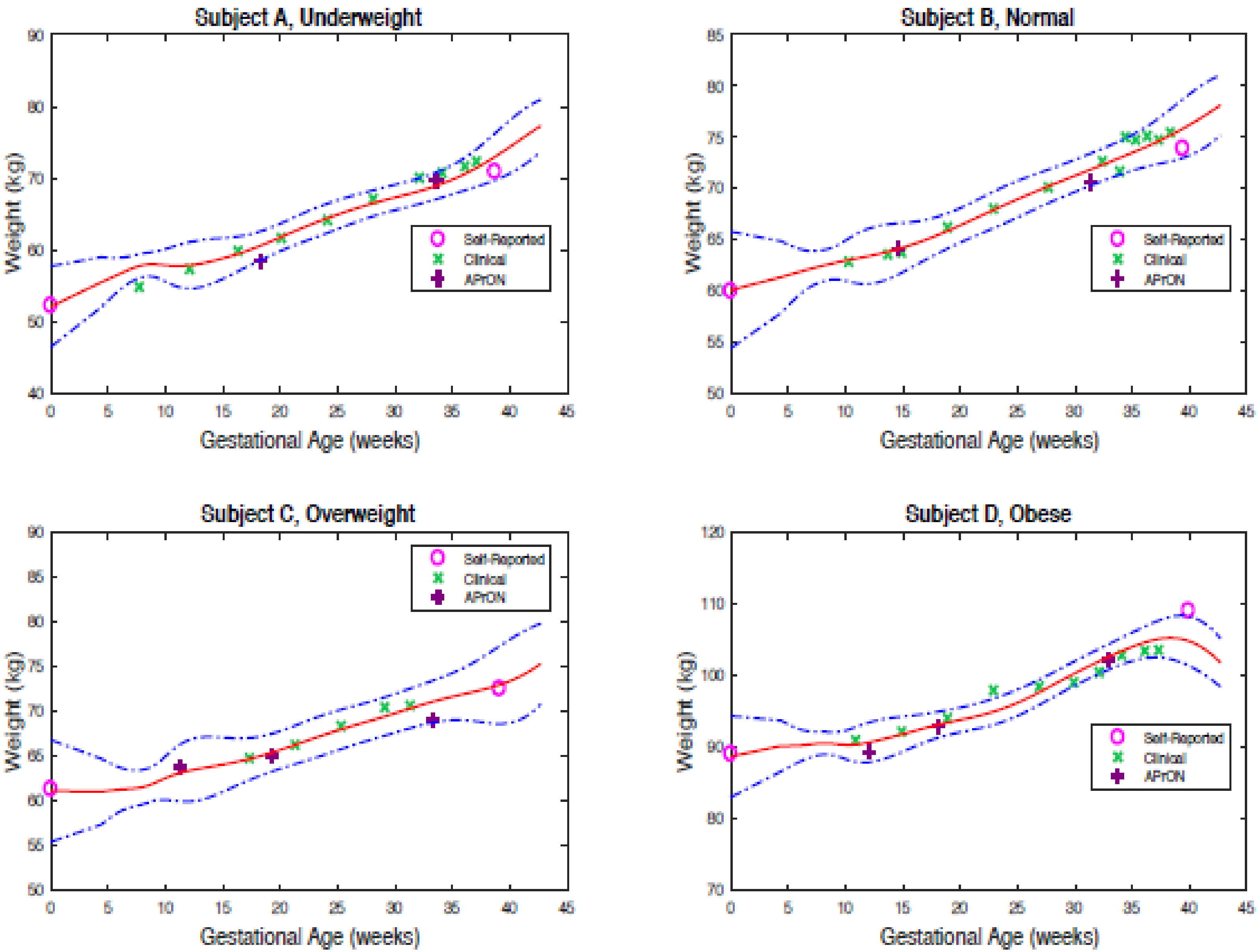
- BMI-category specific patterns



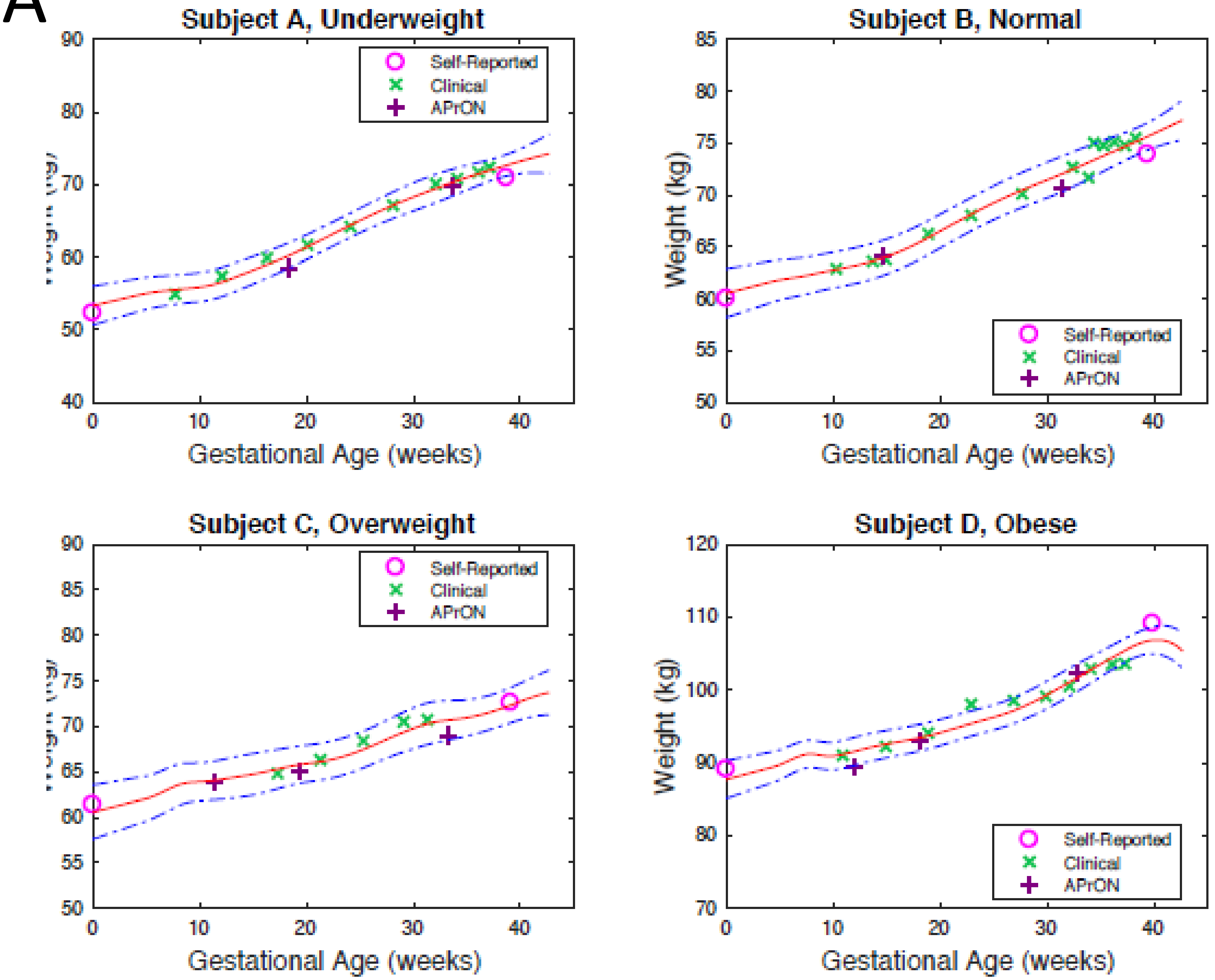
$$Y_{ijk} = X_{ik}(T_{ijk}) + \tilde{\epsilon}_{ijk} = \underbrace{\mu(T_{ijk}) + \sum_{l=1}^{\infty} \xi_{ikl}\phi_l(T_{ijk})}_{\text{Joint}} + \underbrace{\mu_k(T_{ijk}) + \sum_{m=1}^{\infty} \eta_{ikm}\phi_m^{(k)}(T_{ijk})}_{\text{Individual}} + \tilde{\epsilon}_{ijk}$$

$$= X_{ik}^{Joint}(T_{ijk}) + X_{ik}^{Individual}(T_{ijk}) + \tilde{\epsilon}_{ijk}, \qquad T_{ijk} \in \mathcal{T}$$

FPCA


Subject A, Underweight
Weight (kg)
Gestational Age (weeks)
40
50
60
70
80
90
0
5
10
15
20
25
30
35
40
45
Self-Reported
Clinical
APrON
Subject B, Normal
Weight (kg)
Gestational Age (weeks)
50
55
60
65
70
75
80
85
Self-Reported
Clinical
APrON
Subject C, Overweight
Weight (kg)
Gestational Age (weeks)
50
55
60
65
70
75
80
85
90
Self-Reported
Clinical
APrON
Subject D, Obese
Weight (kg)
Gestational Age (weeks)
70
80
90
100
110
120
Self-Reported
Clinical
APrON

FPCA 2.0


Subject A, Underweight
Self-Reported
Clinical
APrON
90
80
70
60
50
40
0
10
20
30
40
Gestational Age (weeks)
Subject B, Normal
Weight (kg)
85
80
75
70
65
60
55
50
Subject C, Overweight
90
85
80
75
70
65
60
55
50
Subject D, Obese
120
110
100
90
80
70

Comparing the Performance

Method	Mean square error	Standard deviation
FPCA	1.55	1.24
FPCA 2.0	0.93	0.96

2. The Link between Relative Risk and “Lift”: An Example of Industrial Chemical Emission and Adverse Birth Outcomes

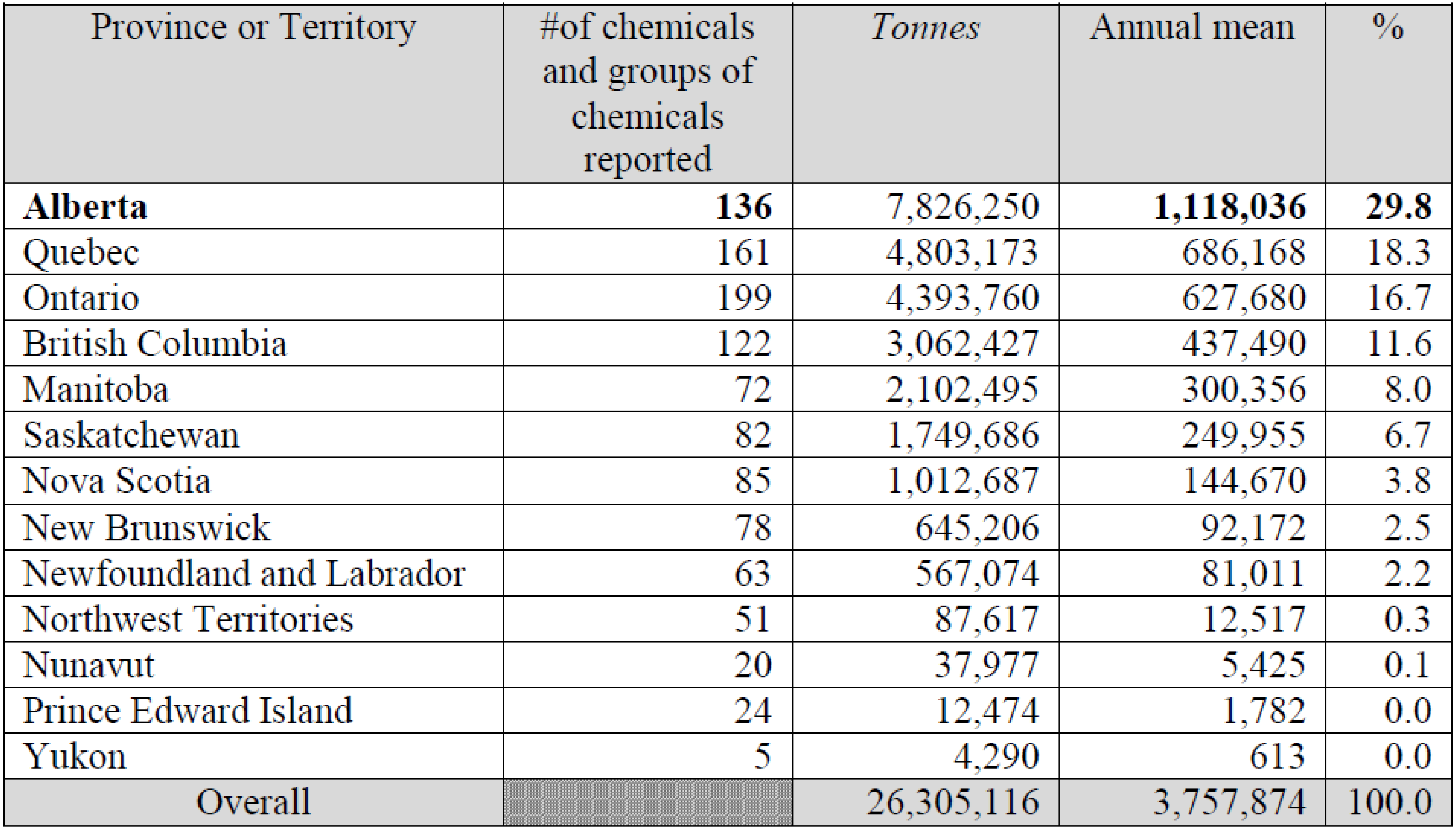
Motivation

- Data mining tools become increasingly popular in medical and health research in the era of big data.
- Association measures used in data mining field are different from those used in traditional medical and epidemiological field.

“TRANSLATION” NEEDED

Data Mining: Industrial Chemical Emission and Adverse Birth Outcomes

- To identify combinations of emitted industry chemicals that associated with adverse birth outcome, e.g. pre-term birth, small for gestation age and low birth weight.
- Data
  - Alberta industry plant locations
  - Mixture of chemicals emitted (type and quantity)
  - Wind direction and velocity
  - Birth outcome



Province or Territory	#of chemicals and groups of chemicals reported	*Tonnes*	Annual mean	%
Alberta	**136**	7,826,250	**1,118,036**	**29.8**
Quebec	161	4,803,173	686,168	18.3
Ontario	199	4,393,760	627,680	16.7
British Columbia	122	3,062,427	437,490	11.6
Manitoba	72	2,102,495	300,356	8.0
Saskatchewan	82	1,749,686	249,955	6.7
Nova Scotia	85	1,012,687	144,670	3.8
New Brunswick	78	645,206	92,172	2.5
Newfoundland and Labrador	63	567,074	81,011	2.2
Northwest Territories	51	87,617	12,517	0.3
Nunavut	20	37,977	5,425	0.1
Prince Edward Island	24	12,474	1,782	0.0
Yukon	5	4,290	613	0.0
Overall		26,305,116	3,757,874	100.0

*Source: Extracted from NPRI databases (2006-2012). Based on initial extraction data (before a complete evaluation of guidelines for all Provinces).

Industrial Sector	Tonnes	%	cum.%
Conventional Oil and Gas Extraction	3,177,490	40.6	40.6
Non-Conventional Oil Extraction (including Oilsands and Heavy Oil)	1,778,269	22.7	63.3
Electricity	1,623,774	20.7	84.1
Wood Products	310,845	4.0	88.0
Chemicals	241,637	3.1	91.1
Pulp and Paper	149,343	1.9	93.0
Petroleum and Coal Prod. Refining and Manufacturing	149,012	1.9	94.9
Oil & Gas Pipelines and Storage	101,502	1.3	96.2
Cement, Lime and Other Non-Metallic Minerals	84,288	1.1	97.3
All other activities*	210,090	2.7	100.0
Total	7,826,250	100.0	

Category	Chemical-class	Chemical name	CAS_Number
1	Sulphur dioxide	Sulphur dioxide	7446- 9-5
2	Nitrogen oxides	Nitrogen oxides (expressed as NO2)	111 4-93-1
3	Carbon monoxide	Carbon monoxide	63 - 8-
4	Particulate Matter	PM2.5 - Particulate Matter <= 2.5 Microns	NA - M1
		PM1 - Particulate Matter <= 1 Microns	NA - M 9
		PM - Total Particulate Matter	NA - M 8
5	Volatile Organic Compounds (non-PAHs)	1,1,2,2-Tetrachloroethane	79-34-5
		1,1,2-Trichloroethane	79- -5
		1,2,4-Trimethylbenzene	95-63-6
		1,2-Dichloroethane	1 7- 6-2
		1,3-Butadiene	1 6-99-
		1,4-Dioxane	123-91-1
		2-Butoxyethanol	111-76-2
		Acetaldehyde	75- 7-
		Acetonitrile	75- 5-8
		Acrolein	1 7- 2-8
		Aniline (and its salts)	62-53-3
		Benzene	71-43-2
		Biphenyl	92-52-4
		Carbon disulphide	75-15-

Association Measure

- Data mining

$$\text{Lift} \quad \overset{\text{def}}{=} \frac{P(OE)}{P(O)P(E)}$$

- Epidemiology

$$\text{RR} \overset{\text{def}}{=} \frac{P(O|E)}{P(O|\overline{E})}, \text{OR} \overset{\text{def}}{=} \frac{P(O|E)/P\left(\bar{O}\middle|E\right)}{P\left(O\middle|\overline{E}\right)/P\left(\bar{O}\middle|\overline{E}\right)}$$

O denotes event and $\bar{O}$ denotes event-free
E denotes exposure and $\bar{E}$ denotes no exposure

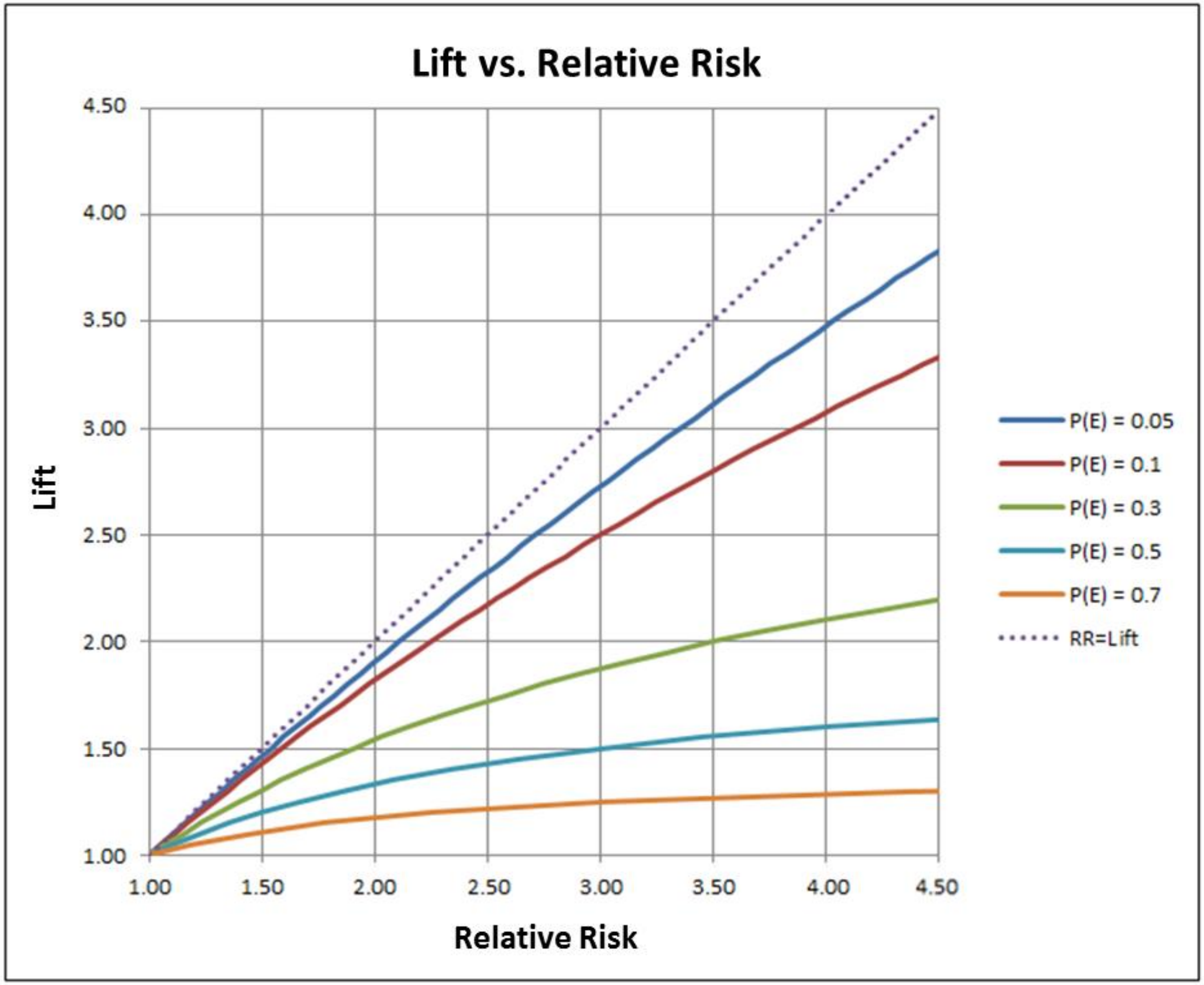
Relationship of the Measures

$$\text{Lift}_{(O|E)} \overset{\text{def}}{=} \frac{P(OE)}{P(O)P(E)} = \frac{P(O|E)}{P(O)}$$

It can be shown

$$\text{RR} = \frac{\big(1-P(E)\big)\text{Lift}}{1-P(E)\text{Lift}},$$

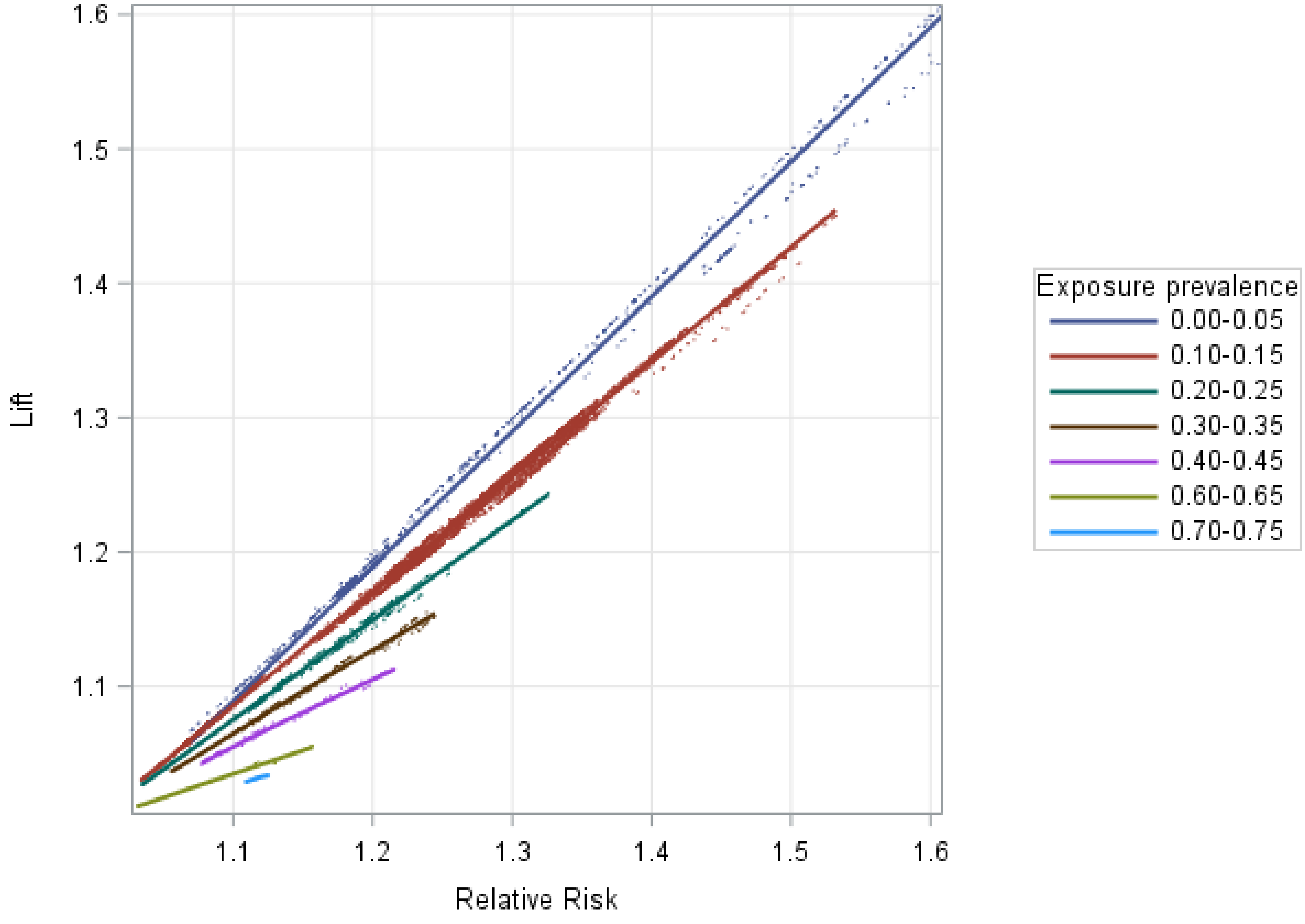
$$\text{OR} = \frac{\text{Lift}_{(O|E)}\left(1 - P(E)\text{Lift}_{(\bar{O}|E)}\right)}{\text{Lift}_{(\bar{O}|E)}\left(1 - P(E)\text{Lift}_{(O|E)}\right)}$$


Lift vs. Relative Risk
Lift
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
1.00
1.50
2.00
2.50
3.00
3.50
4.00
4.50
Relative Risk
P(E) = 0.05
P(E) = 0.1
P(E) = 0.3
P(E) = 0.5
P(E) = 0.7
RR=Lift

Small for gestational age

- The prevalence of SGA: 8.92% (CI:8.59, 9.25). Urban 9.20% *vs* rural 6.78%.
- 13156 one to four chemical combinations were found to be associated with SGA.

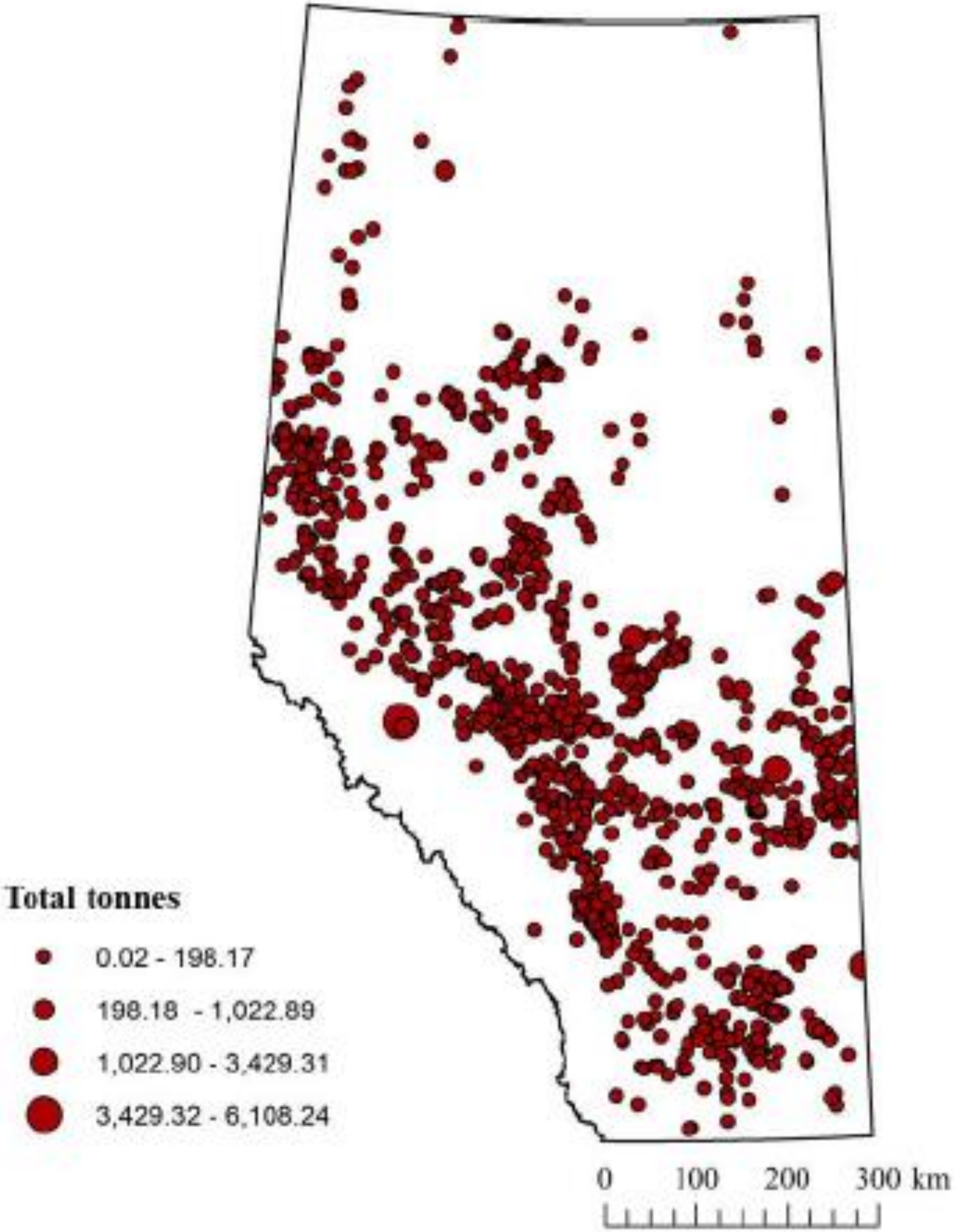
Associations between industrial chemical exposure and adverse birth outcomes from DoMINO study



High Prevalent Exposure

- Exposure to Total Particulate Matter
  - 325,249 births exposed (P(E) = 97%) with 29,616 SGA and 295,633 non-SGA.
  - Lift = 1.01; RR = 1.30; OR = 1.33

PM-mixtures. Alberta 2006-2012



Low Prevalent Exposure

- Exposure to the combination of [Lead and its compounds, Hydrochloric acid, Hydrogen sulphide, Sulphuric acid, Acrolein and n-Hexane]
  - 21,580 birth exposed (P(E) = 6.4%) with 2,787 SGA and 18,793 non-SGA.
  - Lift = 1.4; RR = 1.5; OR = 1.5

Next steps

- Inference
  - adjusting for multiple comparison via permutation and false discovery rate
  - adjusting for known factors, such as lowest SES, smoking during pregnancy, gestational hypertension, past-SGA, and pre-pregnancy mothers’ weight <45 kg.

Acknowledgement

Methodology collaborators

Dr. Khanh Vu, Post-doctoral fellow

Dr. Colin Bellinger, Post-doctoral fellow

Other Collaborators

Dr. Linglong Kong

Dr. Rhonda Bell

Dr. Alvaro Osornio Vargas

Dr. Osmar Zaiane

Graham Erickson, MSc.

Students

Menglu Che, MSc., PhD candidate

Lisa Shulman, MSc. candidate

Rebecca Clark, MSc. candidate

Lois Hole Hospital for Women

women & children's health research institute

CIHR IRSC

Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Thank You