Econometrics and Statistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Constructing a polygenic risk score for childhood obesity using functional data analysis

Sarah J.C. Craig[1,2,#], Ana M. Kenney[3,#], Junli Lin[3], Ian M. Paul[2,4], Leann L. Birch[5,†], Jennifer S. Savage[6,7], Michele E. Marini[7], Francesca Chiaromonte[2,3,8,*], Matthew L. Reimherr[2,3,*], Kateryna D. Makova[1,2,*]

[1] *Department of Biology, Penn State University, University Park*
[2] *Center for Medical Genomics, Penn State University, University Park, PA*
[3] *Department of Statistics, Penn State University, University Park, PA*
[4] *Department of Pediatrics, Penn State College of Medicine, Hershey, PA*
[5] *Department of Foods and Nutrition, University of Georgia, Athens, GA*
[6] *Department of Nutritional Sciences, Penn State University, University Park, PA*
[7] *Center for Childhood Obesity Research, Penn State University, University Park, PA*
[8] *EMbeDS, Sant'Anna School of Advanced Studies, Piazza Martiri della Libertà, Pisa, Italy*

ARTICLE INFO

Article history:
Received 21 October 2020
Revised 21 October 2021
Accepted 23 October 2021
Available online xxx

Keywords:
Feature screening and selection
Functional Data Analysis
Ultra-high dimensional statistics
Under-sampling
Polygenic Risk Score
Statistical genomics

ABSTRACT

Obesity is a highly heritable condition that affects increasing numbers of adults and, concerningly, of children. However, only a small fraction of its heritability has been attributed to specific genetic variants. These variants are traditionally ascertained from genome-wide association studies (GWAS), which utilize samples with tens or hundreds of thousands of individuals for whom a single summary measurement (e.g., BMI) is collected. An alternative approach is to focus on a smaller, more deeply characterized sample in conjunction with advanced statistical models that leverage longitudinal phenotypes. Novel functional data analysis (FDA) techniques are used to capitalize on longitudinal growth information from a cohort of children between birth and three years of age. In an ultra-high dimensional setting, hundreds of thousands of single nucleotide polymorphisms (SNPs) are screened, and selected SNPs are used to construct two polygenic risk scores (PRS) for childhood obesity using a weighting approach that incorporates the dynamic and joint nature of SNP effects. These scores are significantly higher in children with (vs. without) rapid infant weight gain—a predictor of obesity later in life. Using two independent cohorts, it is shown that the genetic variants identified in very young children are also informative in older children and in adults, consistent with early childhood obesity being predictive of obesity later in life. In contrast, PRSs based on SNPs identified by adult obesity GWAS are not predictive of weight gain in the cohort of young children. This provides an example of a successful application of FDA to GWAS. This application is complemented with simulations establishing that a deeply characterized sample can be just as, if not more, effective

* Co-corresponding authors: Kateryna Makova, Matthew Reimherr, Francesca Chiaromonte
E-mail addresses: fxc11@psu.edu (F. Chiaromonte), mreimherr@psu.edu (M.L. Reimherr), kdm16@psu.edu (K.D. Makova).
Authors contributing equally
† deceased

https://doi.org/10.1016/j.ecosta.2021.10.014

than a comparable study with a cross-sectional response. Overall, it is demonstrated that a deep, statistically sophisticated characterization of a longitudinal phenotype can provide increased statistical power to studies with relatively small sample sizes; and shows how FDA approaches can be used as an alternative to the traditional GWAS.

1. Introduction

Obesity is a rising epidemic that is increasingly affecting children. In 2018, 18% of children in the United States were obese and approximately 6% were severely obese (Hales et al., 2018)—a substantial increase from previous years (Ogden et al., 2016). Given the strong association between weight gain during childhood and obesity across the life course (Cunningham, Kramer and Narayan, 2014, Zhang et al., 2019), the search for early life risk factors has become a public health priority.

Obesity is a complex disease with an etiology influenced by environmental, behavioral, and genetic factors, which likely interact with each other (Ang, Wee, Poh and Ismail, 2012). For childhood obesity, dietary composition and sedentary lifestyle have often been cited as main contributors (Sahoo et al., 2015). Evidence also exists for a significant role of parents' socioeconomic status (Barriuso et al., 2015) and maternal prenatal health factors including gestational diabetes (Boney, Verma, Tucker and Vohr, 2005) and smoking (Kries and von Kries, 2002). In addition, obesity risk in children has been associated with appetite (Carnell and Wardle, 2007), which has been shown to be partially influenced by genetics (Wardle et al., 2008).

The heritability of obesity has been estimated to be between 50% and 90% (with the highest values reported for monozygotic twins and the lowest for non-twin siblings and parent-child pairs, reviewed in Maes *et. al*, 1997 (Maes, Neale and Eaves, 1997)). This is a much higher percentage than currently accounted for by known genetic variants (Pigeyre, Yazdi, Kaur and Meyre, 2016, Llewellyn, Trzaskowski, Plomin and Wardle, 2014). This discord is referred to as "missing heritability"—a broad discrepancy between the estimated heritability of the phenotype and the variability explained by genetic variants discovered to date. Indeed, the search for specific genetic variants that increase the risk of obesity, in adulthood as well as in childhood, is still ongoing. Using whole-genome sequencing, researchers have found variants in individual genes that contribute to severe, early-onset obesity (Saeed et al., 2018). Moreover, genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) that are significantly associated with obesity phenotypes such as increased body mass index (BMI), high waist-to-hip ratio, etc. (Locke et al., 2015, Meyre et al., 2009, Speliotes et al., 2010, Thorleifsson et al., 2009, the Early Growth Genetics (EGG) Consortium 2012, Warrington et al., 2015, Felix et al., 2016, Khera et al., 2019, Vogelezang et al., 2020). Albeit successful, these studies have not resolved the "missing heritability" discrepancy, and have other shortcomings; in particular, the individual contributions of the identified SNPs tend to be very small (Pigeyre, Yazdi, Kaur and Meyre, 2016), and the prevalent focus is still on adult cohorts—with only one childhood obesity study for every 10 adult obesity studies (Goodarzi, 2018).

One way to utilize the information gained from GWAS is to summarize the risk from multiple disease-causing alleles in polygenic risk scores (PRSs) that can be computed for each individual (Sugrue and Desikan, 2019). These scores are either simple counts (unweighted) or weighted sums of disease-causing alleles identified by GWAS. Notably, while several studies have constructed PRSs for childhood obesity (Khera et al., 2019, den Hoed et al., 2010, Elks et al., 2010, Belsky et al., 2012, Li et al., 2018), most have done so relying on SNPs identified by GWAS on adult BMI. Since SNPs affecting obesity risk in adults and children may differ (Andersson et al., 2010, Sovio et al., 2011, Graff et al., 2013), this may explain the limited (Pigeyre, Yazdi, Kaur and Meyre, 2016, Llewellyn and Fildes, 2017, Khera et al., 2019) and age-dependent (Frayling et al., 2007, Khera et al., 2019) explanatory power of such scores for children's weight gain status.

We contribute to bridging this gap by focusing specifically on SNPs affecting obesity risk in children. In contrast to the majority of obesity-related GWAS, we focus on a small but deeply characterized pediatric cohort (Paul et al., 2014, Savage et al., 2016, Paul et al., 2018) with growth measurements from birth up to three years of age. This provides the unique opportunity to incorporate the time-dependent nature of weight gain in the analysis. To fully exploit this aspect, we use highly effective Functional Data Analysis (FDA) (Billheimer, Ramsay and Silverman, 2007) techniques to construct children's growth curves and treat them as a longitudinal phenotype. FDA is a field of statistics that incorporates such structure and considers data as evaluations of curves over a discrete grid. Measurements are smoothed into curves, using the correlations of points to provide additional insights. FDA can fully leverage longitudinal information, extracting complex signals that can be lost in standard analyses of cross-sectional or summary measurements (e.g., BMI collected at a single time point). This increases power and specificity for assessing potentially complex and combinatorial genetic contributions. Moreover, FDA models genetic effects on the entire growth curve non-parametrically. This captures changes in effect size over time in a more flexible and effective manner than other statistical methods for longitudinal data. The growing amount of literature in biomedical and 'omics research utilizing FDA (Cremona et al., 2019, Vsevolozhskaya et al., 2016, Huang et al., 2017, Park et al.,

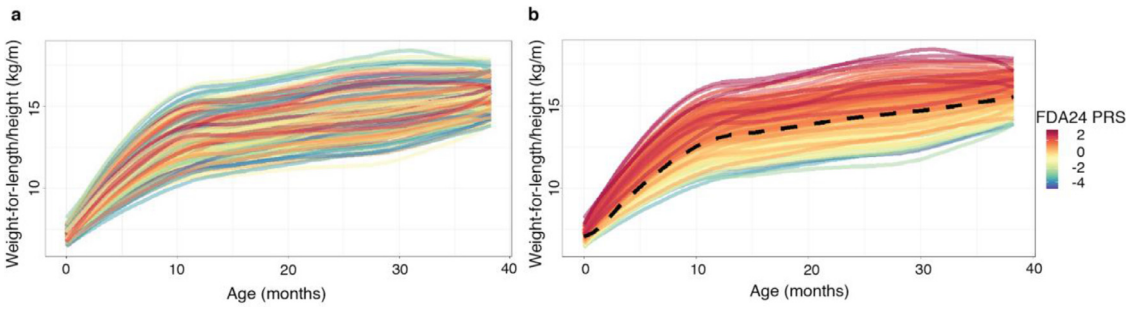

Fig. 1. Growth curves. Growth curves from birth to three years for 210 children enrolled in the INSIGHT study are shown (a) color-coded by participant's ID, and (b) color-coded based on a gradient corresponding to our FDA-based Polygenic Risk Score composed of 24 SNPs (i.e. FDA24 PRS). The dashed black line is the mean curve.

2018, Goldsmith and Schwartz, 2017, Wrobel, Zipunnikov, Schrack and Goldsmith, 2019, Gertheiss, Maity and Staicu, 2013) provides further evidence of the need to capture as much structure as possible from such challenging, complex datasets.

We propose an effective pipeline to select genetic variants significantly associated with children's growth curves and combine them into two novel PRSs that are predictive of growth patterns. The ultra-high dimensional (SNPs) and under-sampled (children cohort) setting we work with necessitates a multi-step procedure to carefully select relevant SNPs and build the scores, with each step utilizing the longitudinal structure. First we apply a computationally efficient feature screening method (Chu, Li and Reimherr, 2016) for function-on-scalar regression to eliminate a substantial portion of the hundreds of thousands of SNPs under consideration. Second, we use a regularization method (Parodi and Reimherr, 2018) for function-on-scalar regression to further select SNPs. Third, we incorporate the corresponding smooth functional estimates of selected SNPs into the PRS weight construction. While existing methods have utilized FDA when studying genetic variants (Vsevolozhskaya et al., 2016, Huang et al., 2017, Chu, Li and Reimherr, 2016), to our knowledge, we are the first to produce PRSs fully constructed with these techniques, and on such a high-dimensional, under-sampled, yet deeply characterized population. We present details of our proposal along with relevant background in Section 2 and address technical aspects of our implementation in Section 3. We provide further evidence of how functional outcomes can increase statistical power in identifying relevant SNPs in Section 4, through simulations mirroring the study characteristics. Section 5 contains our main findings along with several complementary results including the biological interpretation of selected SNPs, independent validation on separate datasets, and comparisons with other PRSs derived by "conventional" GWAS methods. We also investigate how environmental, behavioral, and clinical covariates compound with our novel score in effecting weight gain. We provide final remarks and discussion in Section 6.

1.1. INSIGHT Study and Problem Setup

We collected genetic information from 226 children recruited from the 279 families involved in the Intervention Nurses Start Infants Growing on Healthy Trajectories (INSIGHT) study (Paul et al., 2014). These children are full-term singletons born to primiparous mothers in Central Pennsylvania. INSIGHT is a randomized study which compared a responsive-parenting behavioral intervention aimed at the primary prevention of childhood obesity against a home safety control. The study collected clinical, anthropometric, demographic, and behavioral variables on the children between birth and the age of three years (Table 1). In our study we utilize children's weight and length or height measurements at eight time points; specifically, length was measured with a recumbent Shorr Productions length board at birth, 3-4 weeks, 16 weeks, 28 weeks, 40 weeks, and one year, and standing height was measured with a Seca 216 stadiometer at two and three years. We also utilize 11 covariates including maternal pre-pregnancy BMI, paternal BMI, maternal pregnancy health variables (gestational weight gain, gestational diabetes, and smoking status), family income (as a proxy for socioeconomic status), mode of delivery, child's sex, child's birth weight, INSIGHT intervention group (intervention or control), and mother-reported child's appetite at 44 weeks. The appetite score is an ordinal variable on a scale from 1-5 which summarizes the Child Eating Behavior Questionnaire (CEBQ) (Llewellyn et al., 2012). Domains on the CEBQ include food responsiveness, emotional over-eating, food enjoyment, desire to drink, satiety responsiveness, slowness in eating, emotional under-eating, and food fussiness.

Based on the ratio of weight for length or height (henceforth referred to *weight-for-length/height*) at the eight time points at our disposal, we constructed *growth curves* for each of the 226 children (Fig. 1a; see Section 3.1). Weight-for-length is the recommended measurement for identification of children at risk for obesity under the age of two years by the American Academy of Pediatrics (BMI is recommended afterwards) (Daniels and Hassink, 2015). Since six out of our eight time points fall into this category, we utilized weight-for-length/height ratio for all time points for consistency. In addition to growth curves, we computed the *conditional weight gain* for each child (change in weight between birth and 6 months, corrected for length, see Section 3.1). Conditional weight gain was shown to be an effective indicator of risk for developing obesity later in life in a previous study (Taveras et al., 2009). Also in our study, children who experienced *rapid infant weight gain*, i.e. those with a positive conditional weight gain, had a significantly greater weight at one ($p<2.2\times10^{-16}$), two ($p=9.1\times10^{-14}$), and

Table 1
Description of the study participants

		Rapid infant weight gain (N=104)	Non-rapid infant weight gain (N=122)
Children (N=226)			
	Sex: # males/ # females	53/51	67/55
	Birth weight (g): mean(S.D.)	3486 (449)	3420 (412)
	Delivery Mode: Vaginal (%)	36 (35)	33 (27)
	Weight-for-Length (kg/m) at 1 year: mean (S.D.)	13.67 (0.98)	12.23 (0.93)
	Weight-for-Height (kg/m) at 2 years: mean (S.D.)	15.13 (1.18)[1]	13.79 (1.21)
	Weight-for-Height (kg/m) at 3 years: mean (S.D.)	16.06 (1.29)[2]	14.70 (1.14)[3]
	Appetite score: mean (S.D.)	4.011 (0.92)[4]	3.38 (1.00)[5]
Mothers (N=226)			
	Pre-pregnancy BMI: mean (S.D.)	25.90 (5.75)	25.14 (4.97)
	Gestational weight gain status: did not gain enough/met guidelines/gained excess	16/29/59	25/43/54
	Gestational diabetes: had gestational diabetes (%)	6 (5.8)	7 (5.7)
	Smoked during pregnancy: smoked (%)	8 (7.7)	5 (4.1)
Fathers (N=209)			
	BMI: mean (S.D.)	29.53 (6.28)[6]	27.72 (4.79)[7]
Ethnicity of the children (as reported by their mothers)			
	Black	6 (5.8)	5 (4.1)
	White	94 (90.4)	109 (89.3)
	Native Hawaiian or Pacific Islander	1 (0.4)	0 (0)
	Asian	2 (0.9)	5 (4.1)
	Other	1 (0.4)	3 (1.3)
Annual Household Income			
	< $10,000	2	4
	$10,000 - $24,999	6	9
	$25,000 - $49,999	13	8
	$50,000 - $74,999	30	34
	$75,000 - $99,999	25	23
	$100,000 or more	24	38

1. Missing 3 measurements
2. Missing 9 measurements
3. Missing 7 measurements
4. Missing 16 measurements
5. Missing 20 measurements
6. Missing 9 measurements
7. Missing 8 measurements

three ($p=6.2\times10^{-13}$) years of age than children who did not experience rapid infant weight gain (one-tailed t-tests, Fig. S1). While the growth curves are the main response of interest in our analyses, we further support our results by demonstrating the association of our PRS with conditional weight gain (a scalar outcome) and rapid infant weight gain status (a binary outcome). Fig. 1b provides a snapshot of the relationship between one of our PRSs and the growth curves, demonstrating that curves of children with high PRS values are concentrated above the mean curve. More details are presented in Section 5.1.

Regarding the genetic information, which provides the predictors (or features) in our analyses, we isolated genomic DNA from blood samples from the 226 children and genotyped it on the Affymetrix Precision Medicine Research Array, which contains 920,744 SNPs across the genome. SNPs that had missing information, a minor allele frequency below 0.05, or were in the mitochondrial DNA were removed from the dataset—leaving a total of 329,159 SNPs for subsequent analyses (Fig. S2). With these SNPs we calculated individuals' relatedness to assess the presence of population substructure that may confound the analysis of genomic associations. After computing a relationship matrix we regressed conditional weight gain on the top five principal components of relatedness and found no significant correlation (R^2=-0.003, p=0.4992). This indicates that there is no need to incorporate a population stratification into downstream analyses. Lastly, to apply the FDA techniques, we restricted ourselves to a subset of 210 children and 79,498 SNPs for which we had complete information (no missing SNP values).

2. Background and Proposed Methodology

Let $\mathbb{H}$ be a real separable Hilbert space with norm $||\cdot||_{\mathbb{H}}$. We consider the function-on-scalar linear regression model:

$$Y_n = \sum_{i=1}^{I} X_{ni}\beta_i^* + \varepsilon_n, \ \ 1 \le n \le N, \tag{1}$$

where $Y_n \in \mathbb{H}$, $n = 1, ..., N$ are functional outcomes, $X_{ni} \in \mathbb{R}$, $i = 1, ..., I$ are scalar predictors, $\beta_i^* \in \mathbb{H}$ are nonparametric smooth coefficient functions, and ε_n are iid Gaussian random elements of $\mathbb{H}$ with mean function 0 and covariance operator C. In our problem, $Y \in \mathbb{H}^N$ is composed of the growth curves and $\boldsymbol{X} \in \mathbb{R}^{N\times I}$ is the complete set of SNPs. For a more in depth introduction to FDA and function-on-scalar regression, we refer the reader to Ramsay and Silverman (2007) (Billheimer, Ramsay and Silverman, 2007); Graves et al. (2009) (Ramsay, Hooker and Graves, 2009); Horvath and Kokoszka (2012) (Horváth and Kokoszka, 2012); Hsing and Eubank (2015) (Hsing and Eubank, 2015); and Kokoszka and Reimherr (2017) (Kokoszka and Reimherr, 2017).

We assume that only a small portion—without loss of generality and for ease of notation, say the first I_0, of the I features—are relevant, i.e., that only $\beta_1^*, ..., \beta_{I_0}^*$ are non-zero. Feature sparsity (that is, the existence of only a small portion of relevant SNPs) is a reasonable assumption for our problem; traditional GWAS considering millions of SNPs tend to discover only a handful of significant features (Zhang et al., 2019, den Hoed et al., 2010, Elks et al., 2010, Belsky et al., 2012) with few exceptions (Khera et al., 2019). From a practical standpoint, as will be discussed in Sections 3.4 and 6, a PRS built from a large number of SNPs can also be challenging to utilize and/or validate on new groups of individuals. Thus, we propose a pipeline to narrow down the set of SNPs that have a significant effect on the growth curves. In this section we present relevant background and details of our pipeline, which includes a feature screening method for function-on-scalar regression to quickly and efficiently filter out a substantial portion of the SNPs, a regularization method for more accurate SNP selection post screening, and a novel approach for PRS construction utilizing the nonparametric smooth coefficient functions corresponding to the final set of selected SNPs.

2.1. Feature screening for function-on-scalar regression

Feature screening procedures have gained wide popularity in recent years in part due to the increasingly common ultra-high dimensional problems produced in biomedical research. In the case of scalar outcomes, even popular and usually computationally efficient model selection methods such as the LASSO (Tibshirani, 2011) or SCAD (Fan and Li, 2001) are ineffective for ultra-high-dimensional problems (Fan and Lv, 2008, Fan, Samworth and Wu, 2009, Hall and Miller, 2009). This challenge is only heightened in the presence of complex, functional outcomes like our growth curves. The goal of a screening procedure is to quickly and effectively filter out as many irrelevant features as possible without removing those of importance. This provides a first, substantial reduction in problem size and often removes highly collinear features—paving the way for the effective use of other, more in-depth selection methods to further reduce the problem to a final predictor set.

While the literature on screening for scalar outcomes is extensive (Fan and Lv, 2008, Li, Zhong and Zhu, 2012, Shao and Zhang, 2014), approaches for longitudinal/functional outcomes are still under development. Existing proposals tend to rely on strict assumptions that are not realistic in our problem, e.g., independence among within-subject observations (Fan, Ma and Dai, 2014, Song, Yi and Zou, 2014). In addition, there are often baseline variables that should be controlled for and included within the procedure, for instance, sex may be incorporated to help screen SNPs whose effects are solely sex-based prior to the main selection method. This ability is not supported in existing screens for functional outcomes (Liu, Li and Wu, 2014). Thus, we utilize the screen proposed in (Chu, Li and Reimherr, 2016) that can incorporate baseline variables and utilize within-subject correlation while taking into account a time-varying error variance. This screen has both theoretical guarantees (i.e., the sure screening property (Fan and Lv, 2008)) and numerical evidence of increased screening accuracy by exploiting the correlation structure.

In short, we consider marginal function-on-scalar regression models for each of the I features:

$$Y_n = X_{ni}\beta_i + \sum_{j=1}^{q} Z_{nj}\gamma_j + \varepsilon_n, \ \ 1 \le n \le N. \tag{2}$$

The framework for the screening method is a specific instance of the scenario presented in (1) with $\mathbb{H}$ taken to be $L^2[0, 1]$. Here $Z_{nj} \in \mathbb{R}$, $j = 1, ..., q$ are the optional baseline variables to include in each marginal regression with corresponding smooth coefficient functions $\gamma_j \in \mathbb{H}$. The marginal regressions are fitted and used to provide a ranking of the features, after which the top d can be retained for further analysis. More specifically, a weighted least squares is used to estimate the coefficients for the I models in order to incorporate the covariance matrix of ε_n. The covariance itself is estimated using techniques in (Huang, Wu and Zhou, 2004). The corresponding weighted mean squared error is then used to generate feature rankings.

2.2. Functional linear adaptive mixed estimation (FLAME)

FLAME (Parodi and Reimherr, 2018) is a regularization approach that simultaneously selects important predictors and produces smooth estimates for function-on-scalar linear models by minimizing the target function $L(\beta)$:

$$L(\beta) = \frac{1}{2N}||Y - \boldsymbol{X\beta}||_{\mathbb{H}}{}^2 + \lambda \sum_{i}^{I} \tilde{w}_i ||\beta_i||_{\mathbb{K}}. \tag{3}$$

Here $Y \in \mathbb{H}^N$ and $X \in \mathbb{R}^{N\times I}$ are as previously defined. Let K be a compact linear operator from $\mathbb{H} \to \mathbb{H}$ assumed to be positive-definite and self-adjoint. Then, by the spectral theorem, K can be decomposed as:

$$K = \sum_{i=1}^{\infty} \theta_i v_i \otimes v_i, \tag{4}$$

where $\theta_1 \geq \theta_2 \geq ... \geq 0$ are the ordered eigenvalues with corresponding eigenfunctions $v_i \in \mathbb{H}$. $\mathbb{K}$ is defined as a subspace of $\mathbb{H}$ where:

$$\mathbb{K} := \{h \in \mathbb{H} : \sum_{i=1}^{\infty} \frac{< h, v_i >^2}{\theta_i} := ||h||_{\mathbb{K}}{}^2 < \infty\}. \tag{5}$$

The weights $\tilde{w}_i$ in (3) are akin to those in adaptive LASSO (Zou, 2006, Fan and Reimherr, 2017) and can be calculated in different data driven ways. Here we follow the default settings in FLAME, and first run an un-weighted/non-adaptive step taking $\tilde{w}_i = ||\hat{\beta}_i||_{\mathbb{K}}^{-1}$, where the $\hat{\beta}_i$, $i = 1, ..., I$ are initial coefficient estimates. FLAME utilizes the norm $||\cdot||_{\mathbb{K}}$ to induce both sparsity and smoothness, the degree of which is controlled using the tuning parameter λ. In contrast, previous approaches had to focus on one or the other (Parodi and Reimherr, 2018) and some can be computationally intensive (Chen, Goldsmith and Ogden, 2016), making high-dimensional problems difficult to compute. The implementation of FLAME relies on a functional coordinate descent algorithm, making it computationally efficient and applicable to ultra-high dimensional problems. We note that there are also several regularization procedures (Goldsmith and Schwartz, 2017, Gertheiss, Maity and Staicu, 2013, Mousavi and Sørensen, 2017) with functional features and/or outcomes, but this is not the scenario we consider for this work.

In addition, FLAME allows for a general K, creating a flexible framework that can incorporate various assumptions. Common choices include Sobolev, Gaussian, exponential, and periodic kernels which are discussed in more detail in (Parodi and Reimherr, 2018). We used the recommended Sobolev following the details in (Berlinet and Thomas-Agnan, 2011). Let $\mathbb{H} = L^2(D)$ where D is a compact subset of $\mathbb{R}^d$ and take $\mathbb{K} \subset L^2(D)$ to be the subset of differentiable functions with up to and including m^{th} order derivatives also contained in $L^2(D)$. Let α be a d-dimensional vector of nonnegative integers with $\sum_{i=1}^{d} \sigma_{\alpha_i}\alpha_i \leq m$ where σ_{α_i} are nonzero weights. We can then define a family of norms on $\mathbb{K}$:

$$||x||_{\mathbb{K}}^2 = \sum\nolimits_{|\alpha|\leq m} \frac{1}{\sigma^2{}_{\alpha}} \int_D |x^{(\alpha)}(\boldsymbol{s})|^2 ds. \tag{6}$$

With this norm $\mathbb{K}$ is an RKHS if and only if $m > d/2$. In our case, we again follow the default settings in FLAME and take $D = [0, 1]$, $d = 1$, $m = 1$, and $\sigma = 8$. Then the kernel K can be written explicitly as

$$K(s,t) = \begin{cases} \frac{\sigma}{\sinh(\sigma)} \cosh(\sigma(1-s)) \cosh(\sigma t) & t \leq s \\ \frac{\sigma}{\sinh(\sigma)} \cosh(\sigma(1-t)) \cosh(\sigma s) & t > s \end{cases}, \tag{7}$$

which is solved numerically to identify eigenfunctions and eigenvalues.

2.3. FDA-based Polygenic Risk Scores

In traditional studies, after a GWAS is completed and the final set of SNPs are selected, the PRS is computed through a weighted or unweighted approach. The latter is straightforward; the score of an individual is computed by taking the sum across relevant SNP counts. In the former, some measure of the effect size is used to create weights for each relevant SNP to incorporate varying levels of importance. These are usually the marginal effects on the (binary or scalar) phenotype considered, and are not derived from a joint model. We take a different approach and construct a novel PRS with weights based on effect sizes from a joint model that incorporates the dynamic nature of SNP effects on growth curves through the smooth coefficient curves produced by FLAME. The concept is intuitive: we select weights that maximize the squared covariance between the PRS and the growth curves. In more detail, consider the true set of I_0 features $x \in \mathbb{R}^{I_0 \times 1}$(the SNPs) and the population outcome (growth curve) expressed as $Y(t) = \sum_{i=1}^{I_0} \beta_i{}^*(t)x_i + \varepsilon(t)$. Note, here we narrow the setting of (1) taking $\mathbb{H} = L^2[0, 1]$ to better fit our problem scenario with functions over time. Let $w^* \in \mathbb{R}^{I_0 \times 1}$ be the weight vector, then

we have

$$w^* = arg\ max_w \int \left[Cov(\sum_{i=1}^{I_0} w_i x_i, Y(t)) \right]^2 dt\ s.t.\ w^T w = 1. \tag{8}$$

We can write

$$\int \left[Cov(\sum_{i=1}^{I_0} w_i x_i, Y(t)) \right]^2 dt = \int \left[Cov(\sum_{i=1}^{I_0} w_i x_i, Y(t)) \right]^2 dt = \int \left[Cov(\sum_{i=1}^{I_0} w_i x_i, \sum_{i=1}^{I_0} \beta_i^*(t) x_i + \varepsilon(t)) \right]^2 dt$$
$$= \int \left[Cov(w^T x, \beta(t)^T x) \right]^2 dt = \int [w^T \Sigma_x \beta(t) \beta^T(t) \Sigma_x w dt = w^T (\Sigma_x B \Sigma_x) w, \tag{9}$$

where $\Sigma_x = Cov(x)$ and $B_{ij} = \int \beta_i^*(t) \beta_j^*(t)\ dt$. We can therefore maximize the quadratic form in (7) which when combined with the constraint in (6), attains a maximum of λ_{max}; the maximum eigenvalue of $\Sigma_x B \Sigma_x$. Optimal weights are then easily computed taking $w^* = v_{max}$, the eigenvector corresponding to λ_{max}. For further discussion of the computation of weights, including results when maximizing the correlation rather than covariance, see Section S1.1 in the Supplementary Materials.

3. Implementation

In this section, we provide technical details involved in the implementation of our pipeline. Code for carrying out the key steps in our proposal (screening, the application of FLAME, PRS construction and evaluation) can be found at https://github.com/makovalab-psu/InsightPRSConstruction.

3.1. Functional and non-functional outcome generation

To construct growth curves, we utilized the anthropometric data collected by INSIGHT to calculate weight-for-length/height ratio for each child at each time point. These measurements, considered longitudinally, were then used to create a curve for each child using the Principal Analysis by Conditional Estimation (PACE) algorithm (Yao, Müller and Wang, 2005), which pools information across subjects for more accurate curve reconstruction. We used the implementation of PACE provided in the *fdapace* package in *R* with default settings. The resulting 210 curves (51 cubic splines functions with evenly spaced knots) are shown in Fig. 1a.

Conditional weight gain z-scores (scalar variable) were calculated as the standardized residuals from a regression of age- and sex-specific weight-for-age z-score at six months on the weight-for-age z-score at birth (determined using the World Health Organization sex-specific child growth standards (Savage et al., 2016)). Length-for-age z-score at six months, length at birth, and precise age at the 28-week visit were considered as cofactors in this regression, and thus only the change in weight between birth and six months was captured (Savage et al., 2016, Griffiths et al., 2009). These scores are approximately normally distributed and have, by construction, a mean of 0 and a standard deviation of 1. Positive conditional weight gain z-scores correspond to a greater than average weight gain and are used to define rapid infant weight gain status (binary), which is a risk factor for developing obesity later in life (Baird et al., 2005, Ong and Loos, 2006, Zhou et al., 2016).

3.2. Feature generation

Blood from a fingerstick was collected at the child's one year clinical research visit as part of INSIGHT. Genomic DNA was isolated (Qiagen DNeasy Blood and Tissue Kit) and genotyped on the Affymetrix Precision Medicine Research Array (PMRA). Initial quality filtering was performed using the following criteria: we removed SNPs with minor allele frequency <0.05 and/or present in less than 5% of individuals and SNPs located in mitochondrial DNA. All quality filtering steps were performed in PLINK v1.9 (Purcell et al., 2007, Chang et al., 2015) with 329,159 SNPs remaining after quality filtering.

We calculated the relatedness of the INSIGHT individuals using the –make-rel command in PLINK 1.9 (Purcell et al., 2007, Chang et al., 2015). Principal components (eigenvalues and eigenvectors) of the relationship/relatedness matrix were then computed using the `eigen` function in R.

3.2. SNP screening

For the feature screening procedure described in Section 2.1, we considered marginal models for each SNP and included sex as a baseline feature. After ordering SNPs according to the weighted least squares error of their corresponding marginal model, the top $d = 10,000$ were selected and used for further analysis. We chose 10,000 based on the largest set of SNPs FLAME could sustain before suffering losses in solution quality (e.g., shrinking all coefficient curves to zero). Our implementation of this approach in R is available online in the GitHub repository mentioned above.

3.3. FLAME details and selection stability

Prior to applying FLAME, we standardized the predictors (now consisting only of the top 10,000 SNPs retained by the screening in Section 3.2) and centered the growth curves. We used the default choice of kernel for FLAME (Sobolev), and tuned the penalty for sparsity and smoothness splitting our observations into training (75%) and test (25%) sets and selecting the λ with minimum error on the test set. This resulted in a set of 24 SNPs which were used to construct our FDA-based PRS, i.e. FDA24 PRS (Table 2). Due to the standardization step required for FLAME, coefficient curves were re-computed regressing growth curves on the 24 SNPs using raw SNP counts. These curves were used to compute the final PRS weights, now on the original scale. Coefficient curves and an extended discussion are illustrated in Fig. S3 and Section 1.2 of the Supplementary Materials.

We also assessed the statistical robustness of SNP selection through a 20-fold sub-sampling scheme akin to a 20-fold cross-validation. Specifically, we randomly split the subjects into 20 equal parts (folds) and applied FLAME to perform SNP selection 20 times, each time omitting a different fold. To compare selection results produced through similar levels of induced sparsity and smoothness, we always used the same penalty parameter λ previously selected with the training-test sets split (see above). To ascertain stability of SNP selection, we counted how many times (out of 20) each SNP was selected. Five "top" SNPs, which presented both high selection frequency and large weight magnitude (see Section 5.1), were used to create a second, more parsimonious and possibly more stable FDA-based PRS, i.e. FDA5 PRS. This approach was implemented by modifying existing open source implementations of FLAME in R. This sub-sampling scheme is also used to assess out-of-sample prediction performance. After SNP selection and weight computations on left-in training folds, we compute the PRS for each individual in the left-out test fold. Next we fit a function-on-scalar regression of the growth curves on the PRS values in the left-out test fold, and compute its R^2.. This is repeated on all folds, providing multiple "prediction" R^2 values that we average and compute standard errors for. For further discussion and comparable results obtained changing the number of folds in the sub-sampling scheme see Section S1.3 and Figure S4 in the Supplementary Materials.

3.4. Computing Polygenic Risk Scores

To compute our FDA-based PRSs, we follow the weighting scheme proposed in Section 2.3, taking $w^* = v_{max}$ where v_{max} is the eigenvector of $\Sigma_x B \Sigma_x$ corresponding to the maximum eigenvalue. In practice, Σ_x is replaced with its Maximum Likelihood estimate, and $B_{ij} = \int \beta_i^*(t)\beta_j^*(t)\, dt$ is formed replacing $\beta^*(t)$ with the smooth coefficient estimates of SNPs selected by FLAME and re-computed on the original scale (see Section 3.2). The pairwise integrals are evaluated numerically over a grid of time points.

In Section 5.3, we study the relationship between the children's growth curves and other proposed scores—Belsky PRS (Belsky et al., 2012), Elks PRS (Elks et al., 2010), den Hoed PRS (den Hoed et al., 2010), Li PRS (Li et al., 2018), and Khera PRS (Khera et al., 2019). In order to calculate these PRSs on the INSIGHT cohort we employed the Allelic Scoring function in PLINK v1.9 (Purcell et al., 2007, Chang et al., 2015). For Belsky PRS, Elks PRS, and Li PRS some proxy alleles had to be used in place of SNPs that were not assessed on the PMRA. Such proxies were determined using linkage disequilibrium with LDlink (Machiela and Chanock, 2015). Tables describing the composition of each PRS can be found in the Supplementary Materials (Tables S1-S4). For SNPs for which no appropriate proxy could be found, we utilized imputed data (these are also denoted in Tables S1-S4).

For imputation, we used 294,987 of the 329,159 quality-controlled SNPs (see Section 1.1; 34,172 SNPs were removed because they were not found in the 1000 Genomes Project reference). Children's genotypes were first phased leveraging pedigree information using SHAPEIT2 (Delaneau, Marchini and Zagury, 2011, O'Connell et al., 2014)(genotypes were also collected for mother and father in most cases, and for some younger siblings), The phased haplotypes were then used for imputation using the 1,000 Genomes Project phase 3 data (1000 Genomes Project Consortium 2015) as a reference panel with IMPUTE2 (Howie, Donnelly and Marchini, 2009). SNPs with imputation probability $<$90% were removed. Following imputation, we had information for 12,479,343 SNPs. However, when calculating the Khera PRS we were still only able to use 751,735 SNPs—which is only 37% of the very large total number of SNPs (as many as 2,100,302 (Khera et al., 2019)) included in this score.

4. Simulation Study to Evaluate Sample Size vs Depth

While our sample size is smaller than those of most recent GWAS studies, there is much to gain from the use of the longitudinal information in growth curves. We demonstrate this through a simulation study that builds upon our actual data, as to guarantee realistic settings. In Section 4.1 we attempt to quantify these gains in terms of selection accuracy under a setting where there is a set I_0 of "true" features (e.g., SNPs) in the underlying model. Here FLAME is applied for selection and compared against a corresponding cross-sectional case with an adaptive Lasso (Zou, 2006). This comparison is implemented under a scenario where both selection methods operate on a pre-screened subset containing the true features (thus, complete selection accuracy is possible). We separately compare screening results with a threshold proportional to the one used in our analysis. Here marginal correlations are used for screening in the cross-sectional case.

In Section 4.2, we shift from selection to the risk score, where, assuming a "true" PRS has been constructed that does in fact explain the phenotype, statistical significance is compared between longitudinal and cross-sectional scenarios. Ad-

Table 2
24 SNPs identified as significant predictors of children's weight gain patterns by functional data analysis. SNPs and corresponding weights in bold are the top five SNPs and are used in FDA5 PRS.

SNP	Associated gene[1]	Chr	Position (bp)	Effect Allele	Weight	GWAS association[2]	Folds[3]	Up-down- stream gene[4] & associated GWAS trait[2]
	DNAJC6	1	65826847	C	0.111		12	*LEPR* - early onset extreme obesity
rs72679478								
	ZNF648	1	181952711	C	0.221		17	*CACNA1E* - longitudinal BMI measurement
rs12039940								
	NR5A2	1	199663441	A	0.152	Waist-hip ratio	9	
rs10494802								
rs4915535	*NAV1*	1	201500537	T	-0.218	BMI, waist circumference, obesity related traits	12	
rs638348	*RHOU*	1	229100364	T	-0.358 **-0.691**	Type 2 diabetes	20	
	NOL10	2	10763339	-	0.230 **0.421**		19	*HPCAL1* - type 2 diabetes
rs113822101								
rs9837708	*FOXP1*	3	71487582	T	-0.181	LDL cholesterol, subcutaneous adipose tissue	14	*EIF4E3* - eosinophil counts
rs921551	*TNIP3*	4	122110975	C	-0.159		11	*NDNF* - body weight; *ANXA5* - waist-hip ratio, type 2 diabetes; *PRDM5* - body weight
	FABP6	5	159637815	G	0.115		8	*CCNJL* - BMI-adjusted waist circumference
rs17057519								
	MTCH1	6	36948528	G	0.158	HDL cholesterol	18	*FGD2* - HDL cholesterol, hypothyroidism
rs16889349								
	RNF144B	6	18404450	T	0.126	BMI	10	
rs17626544								
rs4716760	*PTPRN2*	7	157376644	T	0.213	Energy intake, BMI, type 2 diabetes	19	*UBE3C* - BMI, type 2 diabetes
rs1701822	*PPP1R3A*	7	112947322	G	0.135	BMI	11	
	POR	7	75533388	C	0.217	Eosinophil counts	20	*HIP1* - BMI;
rs62475261					**0.396**			*RHBDD2* - eosinophil counts

(*continued on next page*)

Table 2 (*continued*)

SNP	Associated gene[1]	Chr	Position (bp)	Effect Allele	Weight	GWAS association[2]	Folds[3]	Up-down- stream gene[4] & associated GWAS trait[2]
	SEC61G	7	54678150	T	-0.322		14	*VSTM2A* - BMI-adjusted waist-hip ratio
rs10227226								
	OLFM1	9	138038493	C	0.189	Obesity-related traits (estradiol measurement)	13	*COL5A1* - waist-hip ratio
rs58307428								
rs9409226	*BRINP1*	9	122850776	A	0.173 **0.308**		19	*FBXW2* - BMI-adjusted hip circumference; *PHF19* - birth weight; *CDK5RAP2* - asthma
rs2389157	*LINC00557*	13	95433968	A	0.183	Body composition measurement	11	*GPC6* - body composition measurement, visceral adipose tissue measurement
	RBFOX1	16	7459683	C	0.205	BMI, visceral adipose tissue adjusted for BMI, subcutaneous adipose tissue measurement, body weight, body composition measurement	10	
rs17648524								
	DNAH9	17	11751186	A	-0.155		20	*SHISA6* - insulin sensitivity measurement
rs72815409					**-0.305**			
rs4969367	*BAIAP2*	17	79032015	A	0.200	BMI, lean body mass	18	*RPTOR* - BMI, obesity
	TGIF1	18	3451388	AA	-0.162		13	*DLGAP1* - obesity-related traits (igfbp-3 measurement)
rs141177192								
rs1539759	*TIAM1*	21	32895851	C	0.204	Hypertension, renal sinus adipose tissue measurement	17	*SCAF4* - LDL cholesterol; *HUNK* - type 2 diabetes, periodontitis
rs133709	*ELFN2*	22	37827248	A	-0.296		7	*LGALS1* - body fat distribution

[1]As determined by Affymetrix
[2]Association determined by the NHGRI-EBI GWAS catalog of published genome-wide association studies (https://www.ebi.ac.uk/gwas/)
[3]The number of times the SNP is selected in 20-fold cross validation
[4]Determined via LD-Link (Machiela and Chanock, 2015)

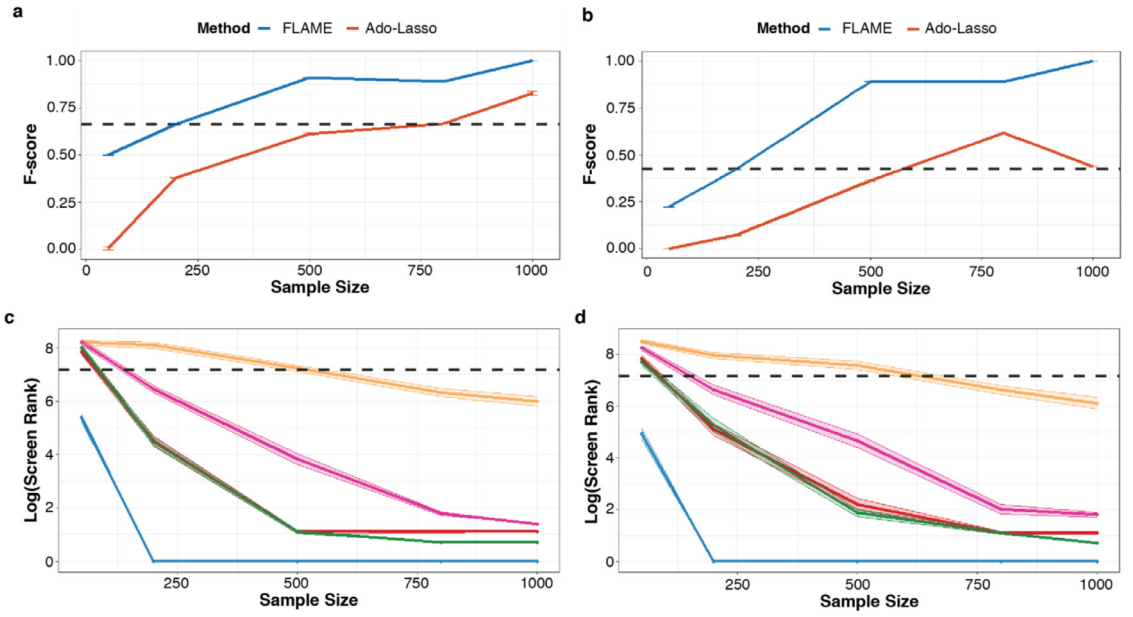

Fig. 2. Average F-scores and median screening ranks across sample sizes (simulated data). Average F-scores and standard error bounds (note the bounds are tight) as a function of sample size, from 100 replications of a simulation procedure with (a) 100 and (b) 1000 potential features, of which only 5 affect the outcome. Median screening rank of the 5 "true" features (out of 10000) and standard MAD error bounds (on the logarithmic scale) as a function of sample size are based on (c) function-on-scalar and (d) scalar regression models. In this simulation exercise, synthetic growth curves, as well as cross-sectional responses, are regressed on artificial features representing SNPs. Dashed black horizontal lines correspond to the average F-score obtained at a sample size of 200 (similar to that of the INSIGHT cohort) in (a)-(b), and denote the 1300th rank in (c)-(d) (1300/10000 is proportional to the threshold of 10000/79498 used in our main analysis).

ditional noise is added to the latter to account for the variability in less controlled studies. This section explores a setting in which a proposed PRS may be biologically valid, but its effect is better captured using longitudinal data and FDA methods. For instance, a score may not validate (or have weaker validation results) if the validation dataset is derived from a cross-sectional study versus a longitudinal one.

4.1. Feature Selection

As mentioned above, we built our numerical experiments from our actual data. We considered the 210 growth curves employed in our analysis, and re-sampled them to create simulated samples of curves with different sizes N. In each such sample, we altered the curves associating them to $I_0 = 5$ artificial features which act as the "true" set of important SNPs in this abstract setting. This was done taking $\tilde{Y}_n = Y_n + X_n{}^T \hat{\beta}_n$ where Y_n is the n-th original growth curve, X_n is a vector of size I corresponding to the n-th value of the artificial features, and the strength of the association, $\hat{\beta}_n$, is an $I \times m$ matrix consisting of the estimated coefficient curves of our FDA5 PRS (see Section 5.1) evaluated over the corresponding $m = 8$ time points for the n-th individual. Since there are $I_0 = 5$ "true" features, only 5 rows of each $\hat{\beta}_n$ are non-zero. Next, to simulate a comparable scenario with a scalar, "cross-sectional" response W, we randomly selected one time point in each curve $\tilde{Y}_n$ —this corresponds to a measurement at a given age A.

We varied the sample size $N = 50, \ldots, 1000$, generating 100 instances for each fixed N, and considered a smaller scenario with $I = 100$ and larger scenario with $I = 1000$. The features, X, are independently sampled from a normal distribution with mean 0 and identity covariance matrix. We applied FLAME under the settings described in Sections 2.2 and 3.3, regressing $\tilde{Y}$ on X in each instance. In the cross-sectional case, we first regressed W on A, computed the residuals W_{res}, and applied an adaptive Lasso to the regression of W_{res} on X using the `glmnet` (Friedman, Hastie and Tibshirani, 2010) package in R. For a fairer comparison, we mimicked the adaptive step in FLAME where in the first run, Lasso is tuned via cross-validation and the resulting estimates are used as weights in the adaptive step taking $1/|\hat{\beta}_{lasso}|$ (any X_j with corresponding $\hat{\beta}_{lasso_j} = 0$ are removed from consideration). This adaptive step is also tuned via cross-validation. To evaluate performance, in each instance we computed the F-score

$$F_{score} = \frac{TP}{TP + 0.5(FP + FN)}, \tag{10}$$

which is a combination of true and false positives. Results are summarized in Fig. 2, which shows averages and standard errors for the F-scores across sample sizes. We observe that a small (say $N = 200$, close to our 210 from INSIGHT) but deeply characterized sample, where a longitudinal phenotype is recorded and exploited with FDA methods, can produce F-

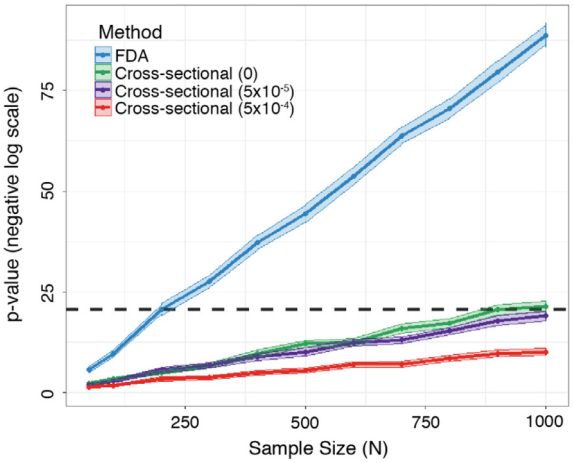

Fig. 3. Significance curves from simulations across sample sizes. Average significance (-log(p-value)) and standard error bands as a function of the sample size, from 100 replications of a simulation procedure. Synthetic growth curves, as well as cross-sectional responses with varying levels or noise, are regressed on an artificial feature representing a polygenic risk score. The dashed black horizontal line corresponds to a p-value of 7×10^{-9}, which is the one obtained using growth curves and a sample size of 200 (similar to that of our study).

scores similar to those produced with much larger samples where only a cross-sectional response is measured; specifically, samples around 4 times larger when $I = 100$, and more than 3 times larger when $I = 1000$. Thus, smaller studies can recover comparable selection accuracy exploiting richer phenotypes.

In the above comparison FLAME and the adaptive Lasso operate on a pre-screened subset of features containing the true set. To contextualize these results, we also applied the functional screening approach to this setting with additional features to total $I = 10,000$ predictors to be ranked. For the cross-sectional case, we ranked the features based on their correlation to the scalar response (after first regressing out age, as above). The median ranking of the 5 true features along with standard median absolute deviations (MAD) are shown in Fig. 2 (on a logarithmic scale due to a large variation range across sample sizes). All but one of the five true features would be included both in the longitudinal and the cross-sectional case if we set a cutoff at the top 1300 features (this 1300/10000 is proportional to the 10000/79498 in our main analysis, but we could reduce the cutoff here with similar results). However, the rankings are stronger in the longitudinal case, and the most difficult feature would be included at a lower sample size. For instance, the median ranking of this feature at a sample size of $N = 500$ is 1386/10000 in the longitudinal case compared to 1982/10000 in the cross-sectional case.

4.2. PRS Significance

Here we follow a simulation framework similar to that of Section 4.1, but associate the curves to an artificial feature P which plays the role of a "pseudo" PRS. This was done taking $\tilde{Y}_n = Y_n + P_n\widehat{\beta}_n$ where Y_n is again the n-th original growth curve, $P_n \sim N(0, 0.5^2)$ is the n-th value of the artificial feature, and the strength of the association, $\widehat{\beta}_n$, is a vector of size m consisting of the estimated coefficient curve of our FDA24 PRS (see Section 5.1) evaluated over the corresponding $m = 8$ time points of the n-th individual. As before, we also simulate a comparable scenario with a scalar, cross-sectional response W by randomly selecting one time point in each curve $\tilde{Y}_n$. Thus, in either case, we assume that the longitudinal and corresponding cross-sectional study are utilizing a truly relevant PRS.

To quantify how the effect of this PRS differs across sample sizes in the longitudinal case, we performed a functional regression for $\tilde{Y}$ on P and recorded the resulting p-values. In the cross-sectional case, we performed a standard regression of W on P and age A and recorded p-values for P. Thus, while this regression does not use longitudinal information, it does correct for age when evaluating the effect of P on the cross-sectional phenotype W.

To more realistically account for the variability in less controlled studies (e.g. based on Electronic Medical Records), we also generated cross-sectional responses with larger variation/noise, adding Gaussian errors with mean 0 and variance s^2 to the cross-sectional response. In a first scenario, the variance was "calibrated" on the INSIGHT data; we took $s^2 = 5 \times 10^{-5}$, the within-day variation estimated from a mixed effects model for the weight-for-length/height ratio (fixed effect for age, random effects for individual and observation number—for visits where multiple measurements were taken). In a second scenario, meant to mimic a study where measurements are less accurate than those collected in INSIGHT, we took $s^2 = 5 \times 10^{-4}$. In both scenarios we again performed a standard regression on W and A, and recorded the p-value for W. The whole procedure—creating simulated samples of different sizes, generating the artificial explanatory feature (pseudo-PRS) and the various functional and scalar responses, performing the various regressions and recording p-values—was repeated 100 times, allowing us to compute averages and standard errors for the negative log p-values plotted in Fig. 3.

We observe very similar gains in power to those seen in Section 4.1. Here the smaller sample (say $N = 200$, close to our 210 from INSIGHT) can produce results with a significance equivalent to those produced with a sample more than 4 times larger where only a cross-sectional response is measured—even when this response contains no additional noise. Adding noise akin to that in INSIGHT increases this factor to more than 5 times, and one would need sample sizes well beyond $N = 1000$ to obtain comparable significance at higher levels. These results help us gauge how smaller longitudinal studies may truly compare to larger cross-sectional ones, and also suggest that truly effective scores may appear less impactful when validated on datasets where only scalar phenotypes are available.

5. Linking FDA-based PRSs to Childhood Obesity

In this section we present the main findings obtained applying our proposed pipeline to the data from the INSIGHT children cohort. We assessed the in-sample association between growth curves and our FDA-based polygenic risk scores fitting function-on-scalar linear models (Kokoszka and Reimherr, 2017). The significance of the PRSs as predictors of the growth curves was determined based on three tests (Choi and Reimherr, 2018) employing different types of weighted quadratic forms. One employs a simple L2 norm of the parameter estimate (L2), another uses principal components to reduce dimension prior to a Wald-type test (PCA), and the last blends the two through the addition of a weighted scheme in the PCA (Choi). We reported the more conservative of the three values, which in this case was Choi. We complemented results on growth curves considering alternative phenotypes, including curves based on BMI rather than the weight-for-length/height ratio, the scalar conditional weight gain score, and the binary rapid infant weight gain (see Section 1.1). To demonstrate generalizability, we validated results on two independent cohorts. For both, PRSs were calculated using the score function in PLINK v1.9 (Purcell et al., 2007, Chang et al., 2015), and proxies for missing SNPs were determined using LDLink (Machiela and Chanock, 2015) (see Table S5). We also analyzed the relationship between other PRSs proposed in prior literature and the growth curves in our INSIGHT cohort. Finally, we studied the contributions of environmental, behavioral, and clinical covariates when considered jointly with our scores.

5.1. Main results

A screening procedure (Chu, Li and Reimherr, 2016) (see Section 2.1) was used prior to FLAME (Parodi and Reimherr, 2018) (see Section 2.2) to reduce the analysis from 79,498 to the top 10,000 potentially relevant SNPs (Fig. S2). FLAME then identified 24 SNPs as significant predictors of children's growth curves (Table 2). Using information from the 24 selected SNPs, we constructed our novel FDA24 PRS as a *weighted* sum of allele counts (see Section 2.3).

Our FDA24 PRS is indeed a strong predictor for growth curves, with a significant positive effect on weight-for-length/height ratios across time, especially between ∼10 and ∼30 months of age (Fig. 4a) (function-on-scalar regression, in-sample R^2=0.52, p=9.2×10^{-5} – note that, as with any study and GWAS studies in particular, in-sample R^2 values must be taken with caution as they are unrepresentative of true predictive power, see below and in Section 5.2 for "prediction" R^2 evaluated through a sub-sampling scheme and validation datasets). This can also be observed noting that growth curves of children with high PRS values are concentrated above the mean curve (Fig. 1b). Moreover, FDA24 PRS is significantly larger for children with rapid infant weight gain compared to those without (one-tailed t-test, p=3.3×10^{-8}; Fig. 4b), and is positively correlated with conditional weight gain (in-sample R^2=0.16, p<1×10^{-05}; Fig. 4c) as well as with weight-for-length/height ratio at one (in-sample R^2=0.50, p<1×10^{-5}), two (in-sample R^2=0.53, p<1×10^{-5}), and three (in-sample R^2=0.46, p<1×10^{-5}) years of age (Fig. S5).

As described in Section 3.3, in order to assess the robustness of our FDA-based SNP selection, we performed a sub-sampling stability analysis akin to a 20-fold cross-validation. Notably, for the 24 SNPs included in our FDA24 PRS, the weights computed to construct the PRS correlate with the number of times the SNPs are selected in this sub-sampling scheme (Fig. 5). The frequency of selection captures how stable the effect of a genetic variant is amid the complex and combinatorial signals in this type of data. SNPs which have both the highest selection frequency and the largest weights may be the most important to interpret and validate in future studies. Furthermore, this sub-sampling scheme used for stability can also be utilized to assess out-of-sample prediction performance, computing average "prediction" R^2 values over left-out folds. We follow the proposed procedure using 20, 10, and 5 folds. There are benefits and drawbacks to each, given that more folds provide more instances to average over, but reduce the size of each test sample (making the function-on-scalar regression more challenging). Averages and standard errors of "prediction" R^2 values with different numbers of folds are provided in Table 3. As mentioned in Section 3.3, we provide comparable results in the Supplementary Materials (see Figure S4). The R^2 values reported in Table 3 are an order of magnitude smaller than the in-sample R^2 summarized above. However, this reduction in explanatory power and the magnitudes of these "prediction" values are very much in line with those found in most GWAS studies concerning complex phenotypes such as weight gain (Felix et al., 2016, Goodarzi, 2018, Justice et al., 2019). Indeed, these are of the same magnitude as those found in our external validation results in Section 5.2.

In addition to calculating a PRS based on the full complement of 24 SNPs selected by FLAME (Parodi and Reimherr, 2018), we computed a PRS restricted to the top five SNPs in terms of selection frequency and weight magnitude, as highlighted by our stability analysis (Fig. 5; see also Table 2). These five SNPs are rs72815409, rs638348, rs9409226, rs113822101, and rs62475261, and we refer to the PRS calculated on them as *FDA5 PRS*. Like FDA24 PRS, FDA5 PRS also has a significant positive effect on weight-for-length/height ratios across time (function-on-scalar regression, in-sample R^2=0.21, p=4.3×10^{-5};

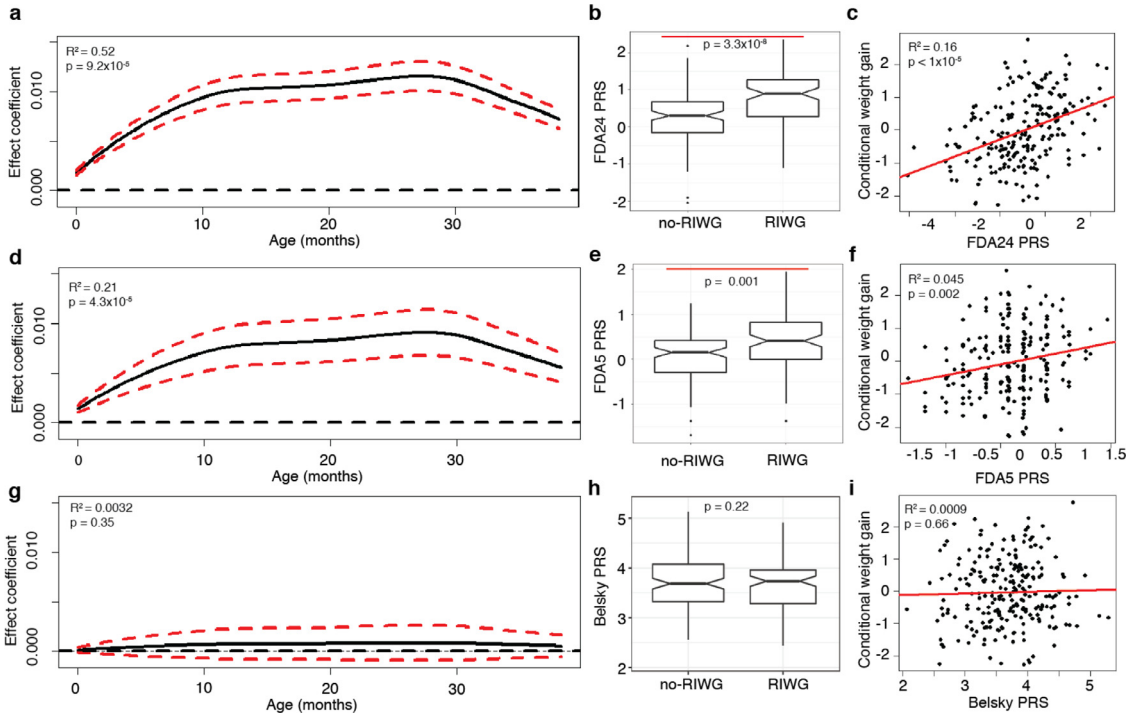

Fig. 4. Polygenic risk scores (PRSs) and children's growth patterns. Estimated effect coefficient for a PRS as a predictor of children's growth curves in a function-on-scalar regression for (a) FDA24 PRS, (d) FDA5 PRS, and (g) Belsky PRS. Boxplots comparing a PRS between children with vs. without rapid infant weight gain (RIWG vs. no-RIWG) for (b) FDA24 PRS, (e) FDA5 PRS, and (h) Belsky PRS. Scatterplot of conditional weight gain vs. a PRS using **(c)** FDA24 PRS, **(f)** FDA5 PRS, and **(i)** Belsky PRS.

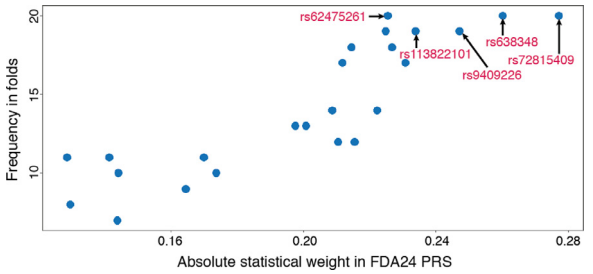

Fig. 5. Statistical validation of FDA-based SNP selection. The frequency with which the 24 SNPs included in the FDA24 PRS are re-selected in a 20-fold sub-sampling scheme is plotted against their absolute statistical weight in the FDA24 PRS—showing a strong positive association. The SNPs with both the largest weights and the highest re-selection frequency (top five SNPs marked by arrows) may be the most important to interpret and validate in future studies.

Table 3
Average and standard errors of "prediction" R^2 values with different choices of the number of folds in our sub-sampling scheme. Average "prediction" R^2 values and their standard errors (in parentheses) using the sub-sampling scheme proposed in Section 3.3. The number of folds utilized varies (20, 10, and 5 folds). The corresponding numbers of samples in training and test sets are given for reference.

No. Folds	n^{train}	n^{test}	R^2 (se)
20	200	10	0.049 (0.025)
10	189	21	0.031 (0.013)
5	168	42	0.024 (0.013)

Fig. 4d), a positive correlation with conditional weight gain (in-sample R^2=0.045, p=0.002; Fig. 4f), and values that are significantly higher for children with rapid infant weight gain compared to those without (one-tailed t-test, p=0.001; Fig. 4e). We note here that explanatory power evaluated after model selection (in our case, post selection of SNPs) and in-sample can be highly inflated, just as in any GWAS. However, as part of the biological validation analysis presented in the following section, we confirmed the predictive performance of our scores on two completely independent data sets.

Based on the NHGRI-EBI GWAS catalog (https://www.ebi.ac.uk/gwas/), one of the FDA5 PRS SNPs is located in a gene linked to a metabolic disorder: rs638348 is located within gene *RHOU* associated with type 2 diabetes (Table 2). Additionally, two other SNPs of the FDA5 PRS are downstream of genes also associated with diabetes: rs72815409 is downstream of *SHISA6* (associated with an insulin sensitivity measurement) and rs113822101 is downstream of *HPCAL1* (associated with type 2 diabetes). The fourth SNP in FDA5 PRS, rs9409226, is upstream of *FBXW2* (associated with BMI-adjusted hip circumference), *PHF19* (associated with birth weight), and *CDK5RAP2* (associated with asthma). The fifth SNP in FDA5 PRS (rs62475261) is located upstream of *HIP1* (associated with BMI) and downstream of *RHBDD2* (associated with eosinophil counts, an asthma-related trait). Thus, rs9409226 and rs62475261 are located in the vicinity of genes associated with obesity-related traits and asthma. There is unequivocal epidemiological evidence linking obesity with asthma (reviewed (Peters, Dixon and Forno, 2018)), but a shared genetic underpinning has been challenging to elucidate (Melén et al., 2010); our results suggest further investigation is warranted.

Among the 19 SNPs included in our FDA24 PRS but not in the FDA5 PRS (Table 2), twelve are located within genes linked to obesity-related traits such as BMI (rs4915535, rs471670, rs17648524, rs4969367, rs17626544, rs1701822), cholesterol levels (rs9837708 and rs16889349), body composition measurement (rs2389157), waist-to-hip ratio (rs10494802), hypertension (rs1539759), and estradiol measurement (rs58307428). Seven additional SNPs are located in the vicinity of other obesity-related genes: for example, rs72679478 is located just upstream of the leptin receptor gene (*LEPR*) which has been associated with early-onset adult obesity (Wheeler et al., 2013). In summary, while none of the SNPs we identified are located in the most typical and well-known obesity genes (e.g., *FTO* (*Frayling et al., 2007*, *Fall and Ingelsson, 2014*) and *MC4R* (*Locke et al., 2015*, *Fall and Ingelsson, 2014*)), all of them are located either within, or in the vicinity of, genes linked to obesity or metabolic disorders in previous studies.

5.2. Biological Validation of the FDA-based Polygenic Risk Score

BMI in our cohort

To provide an initial biological validation of the FDA24 PRS constructed using weight-for-length/height ratio growth curves, we considered growth curves for the children in our INSIGHT cohort constructed using a different (albeit highly related) measure of weight gain, i.e. BMI. As mentioned above, weight-for-length/height ratio is recommended for children under two years of age by the American Academy of Pediatrics (Daniels and Hassink, 2015), however, our cohort is also observed at ages two and three, when BMI is recommended as the most meaningful measurement (Daniels and Hassink, 2015). Notably, our weight-for-length/height FDA24 PRS is also a strong predictor for the BMI growth curves (R^2=0.40, p=2.9×10^{-5}, function-on-scalar regression)—suggesting a reasonable consistency between the information conveyed by the two measurements, at least up to the age of three years. FDA5 PRS is a strong predictor for BMI-based growth curves as well (R^2=0.18, 9.1×10^{-5}).

BMI in two independent cohorts

Among freely publicly available datasets, none provides genome-wide SNP data and longitudinal weight and length or height measurements for children under the age of three. Notwithstanding the unavailability of a good match to our study design, we were able to successfully validate FDA5 PRS on two independent dbGaP cohorts consisting of older children and adults. The first dataset consists of 283 children between the ages of 8 and 9 from the Philadelphia Neurodevelopment Cohort (dbGaP study phs000607.v3.p2 (Glessner et al., 2010, Calkins et al., 2014, Calkins et al., 2015)) who are identified as European Americans. The average FDA5 PRS, when individuals are grouped according to BMI deciles, exhibits an increasing trend as BMI increases (Fig. 6a). Moreover, the FDA5 PRS of children in the highest BMI decile is significantly higher than in the lowest one (p=0.041, one-tailed t-test), and there is a marginally significant, positive correlation between FDA5 PRS and BMI (R^2=0.011, p=0.081).

The second dataset consists of 2,486 adults (≥18 years of age) from the eMERGE study (dbGaP study phs000888.v1.p1) who identify as white. We see again a significant difference in average FDA5 PRS between the lowest and highest deciles of BMI (p=0.03, one-tailed t-test; Fig. 6b). The correlation between FDA5 PRS and BMI, though still marginally significant, is weaker in this adult cohort than in the Philadelphia children's cohort considered above (R^2=0.0012, p=0.087)—perhaps due to the larger difference in age with individuals in our study. Nevertheless, and remarkably, the FDA5 PRS based on our children's weight gain patterns is predictive of extreme obesity much later in life.

Considering the broader FDA24 PRS comprising all 24 SNPs instead of the FDR5 PRS, we did find a significant and in fact more pronounced difference between the first and 10th BMI decile in the eMERGE adults' cohort. However, we did not find a significant difference between those BMI deciles in the Philadelphia children's cohort (Table S6). The latter result may be due to the difficulty of validating a score based on a larger number of SNPs on small cohorts; for the 283 Philadelphia children, there are fewer allele counts across all 24 SNPs—in fact, some of the FDA24 PRS SNPs are completely missing. This issue does not arise for the 2,486 eMERGE adults. Notably, if we do not filter based on race and analyze the full cohort of 3,098 extremely obese and non-obese adults from eMERGE, decile differences and correlations are even more significant (Table S6). We conducted a number of other tests on these two validation cohorts (e.g., contrasting underweight and obese individuals), which further demonstrated the presence of a predictive signal in our scores (see Table S6). BMI groups were determined using sex-specific, BMI-for-age percentiles as described by the Centers for Disease Control and Prevention.

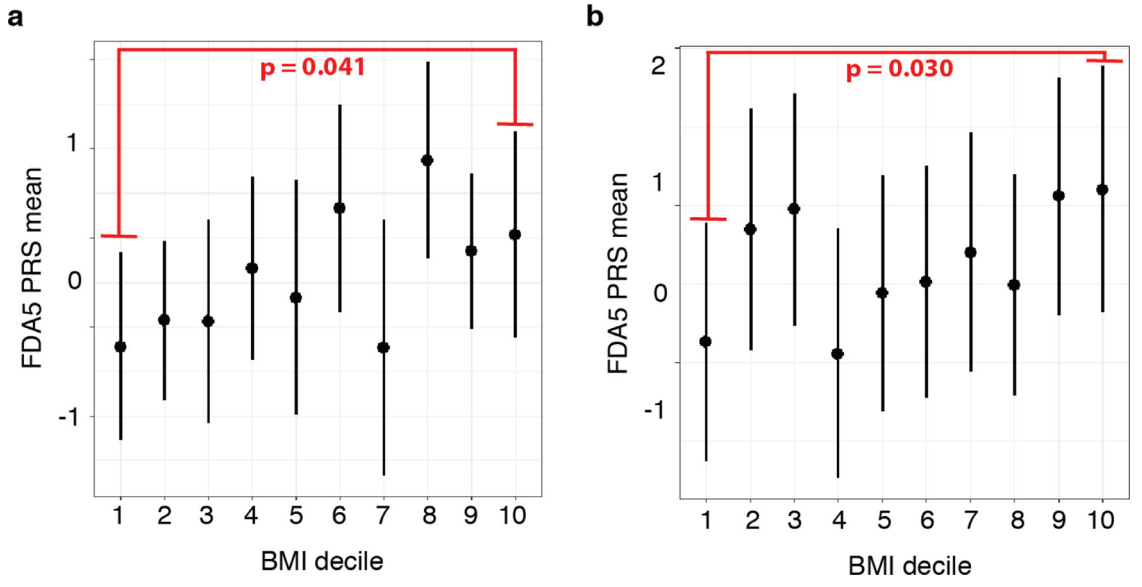

Fig. 6. FDA-based Polygenic Risk Score and obesity in adolescent and adults validation cohorts. (**a**) Distributions of FDA5 PRS in adolescents (age 8 and 9 years) from The Philadelphia Neurodevelopment Cohort by BMI decile (N=28 per decile). (**b**) Distributions of FDA5 PRS in adults (over 18 years of age) either classified as normal or extremely obese in the eMERGE study by BMI decile (N=284 per decile).

5.3. Other Polygenic Risk Scores

While our FDA5 PRS based on children's weight gain trajectories does validate in independent cohorts of older children and adults, PRSs based on adult GWASs do not validate in our cohort of children. First, we considered *Belsky PRS*—a weighted PRS based on 29 SNPs identified through adult obesity GWAS as described by Belsky and colleagues (Belsky et al., 2012). This PRS was shown to correlate with BMI outcomes from age three to 38, so we hypothesized it may also be a good predictor of weight outcomes in very early life. However, Belsky PRS is not a significant predictor of our children's growth curves from birth through age three (R^2=0.0032, p=0.35, function-on-scalar regression, Fig. 4g). Furthermore, Belsky PRS is not significantly larger for children with rapid infant weight gain compared to those without (one-tailed t-test, p=0.22; Fig. 4h) and does not display significant correlations with conditional weight gain (R^2=0.0009, p=0.66; Fig. 4i) and weight-for-length/height ratio at one (R^2=0.0064, p=0.25), two (R^2=0.0036, p=0.37), and three (R^2=0.0009, p=0.71) years of age (Fig. S6).

In addition to Belsky's PRS, we considered four other previously published PRSs specific to childhood obesity—Elks PRS (Elks et al., 2010), den Hoed PRS (den Hoed et al., 2010), Li PRS (Li et al., 2018), and the recent "life-long" Khera PRS (Khera et al., 2019). Similar to the Belsky PRS, the den Hoed, Li, and Khera PRSs were not significantly associated with our children's growth curves (Fig. S7d,g, j), were not significantly different between children with vs. without rapid infant weight gain (Fig. S7e,h,k), and did not have a significant correlation with conditional weight gain (Fig. S7f,i,l). The Elks PRS showed weak but significant association with our children's growth curves (R^2=0.021, p=0.019; Fig. S7a) and correlation with conditional weight gain scores (R^2=0.02, p=0.032; Fig. S7c), but no significant difference between children with vs. without conditional weight gain (two-tailed t-test, p=0.20; Fig. S7b).

5.4. Contributions of environmental, behavioral, and clinical covariates

In addition to genetics, children's weight gain patterns can be affected by a variety of environmental and behavioral factors. To evaluate their potential effects on our results, we considered a regression of conditional weight gain (Taveras et al., 2009) on FDA24 PRS plus 11 potential confounding covariates, namely: maternal pre-pregnancy BMI, paternal BMI, child's birthweight, maternal gestational weight gain, maternal gestational diabetes, maternal smoking during pregnancy, mode of delivery, the child's sex, mother-reported child's appetite score, INSIGHT intervention group, and family socioeconomic status (Table 1).

We applied best subset selection (Hastie, Tibshirani and Friedman, 2009) to the regression of conditional weight gain scores (Savage et al., 2016) on these potentially confounding covariates via the `leaps` package in R (Lumley and Lumley, 2013). The size of the best subset was tuned using the Bayesian Information Criterion (BIC). We included (separately) Belsky PRS, FDA24 PRS, or FDA5 PRS as a 12th predictor in the selection procedure. Only the FDA24 PRS ($p=8.5\times10^{-8}$) and appetite score ($p=2.4\times10^{-3}$) were identified as significant predictors. A regression comprising these two features had an R^2 of 0.22 (Table S7), only six percentage points higher than the one obtained with FDA24 PRS alone (R^2=0.16). Very similar results

were obtained using FDA5 PRS in place of FDA24 PRS (see Table S7). However, and not unexpectedly given its lack of association with children's growth patterns, when we reran the analysis using the Belsky PRS in place of the FDA24 PRS, we did not identify it as a significant predictor. Best subset selection for the regression of conditional weight gain on the Belsky PRS plus the 11 environmental and behavioral covariates retained only appetite score as a positive and significant predictor ($p=5.53 \times 10^{-5}$); all other predictors, including the Belsky PRS itself, were eliminated.

Obviously a high polygenic risk score does not deterministically imply that an individual will develop a particular disease, and our study does not establish causal links between any particular SNP and childhood obesity. However, one can assess whether an intervention on individuals with a high PRS can be successful in mitigating disease progression. Due to the nature of our cohort (collected for a randomized, early-life intervention clinical trial for the prevention of obesity (Paul et al., 2014)), this question can be answered in a retrospective manner. We found that, among children with an above-average FDA24 PRS, those who were part of the intervention group ($N = 115$) had a significantly lower conditional weight gain than those who were part of the control group ($N = 111$) (two-sided t-test, p-value = 0.036). This suggests that a screen based on the FDA24 PRS could potentially be used in future studies proposing intervention to identify individuals for whom it would be most beneficial.

6. Discussion

Genetics of childhood and adult obesity

In this study, we used FDA techniques to construct a novel polygenic risk score (FDA24 PRS) which includes 24 SNPs selected based on children's longitudinal weight gain patterns. Among our study participants, this score explains approximately 52% of the *in-sample* variability in growth curves from birth to the age of three years, and approximately 16% of the in-sample variability in conditional weight gain. We also assessed the stability of our SNP selection and constructed a second score (FDA5 PRS) using the 5 most stable SNPs among the 24. This restricted score explained approximately 21% and 4% of the in-sample variability in growth curves and conditional weight gain, respectively.

As with all genetic studies, our in-sample figures for explained variability are inflated and do not reflect predictive performance at large. When we assess predictive performance through our proposed sub-sampling scheme and external validation datasets, the predictive performance decreases substantially – by as much as an order of magnitude. However, these values are still comparable to other GWAS regarding complex phenotypes, given that signals in these problem settings are expected to be weak. We were in fact able to validate our FDA5 PRS and FDA24 PRS in two independent datasets comprising older children and adult individuals, but our results should still be considered preliminary. Replication in a large, prospective infant cohort would be of great benefit to show the generalizability of our risk scores as clinical markers for childhood obesity.

Interestingly, while our risk scores did validate on older individuals, none of the published PRSs based on adult BMI SNPs showed a significant association with growth curves and conditional weight gain measurements in our children cohort—with the exception of the Elks PRS (Elks et al., 2010), which did show a weak but detectable association signal. Notably, Elks PRS was based on SNPs identified in adult BMI GWAS but sub-selected specifically for their association with weight in children, which may explain its improved performance over other adult PRSs. Previous studies also supported only a weak relationship between PRSs based on adult BMI SNPs and childhood weight gain status (den Hoed et al., 2010, Elks et al., 2010, Belsky et al., 2012, Li et al., 2018)—and pointed out that the relationship became weaker the younger the age of the children (Llewellyn, Trzaskowski, Plomin and Wardle, 2014, den Hoed et al., 2010, Belsky et al., 2012). In fact, Belsky and colleagues (Belsky et al., 2012) found no relationship between their PRS and BMI at birth and a very weak relationship at three years of age. Similarly, Khera and colleagues (Khera et al., 2019) documented relatively weak (albeit significant) associations between their PRS and birth weight and stronger, more significant associations at age eight. These PRSs based on adult SNPs are consistent in that they show increasing effects on obesity-related phenotypes as individuals age, but almost no effect on the very young children comprising our cohort. In contrast, our FDA PRSs based on childhood SNPs show a significant association with obesity-related phenotypes later in life. Even though, based on the validation datasets at our disposal, the association appears to weaken with increasing age, SNPs affecting weight gain in early childhood (i.e. those included in our scores) do retain some predictive power. This is consistent with the notion that early life weight gain, and hence its genetic underpinning, predispose to obesity across the lifecourse (Cunningham, Kramer and Narayan, 2014).

The 24 SNPs identified by our study do not appear in prior PRSs for either childhood or adult obesity, except for the genome-wide Khera PRS, which contains 13 of our SNPs (as part of their set of 2,046,991 SNPs). However, all 24 are located in, or in the vicinity of, genes linked to obesity-related or metabolic disorder phenotypes in previous GWAS studies (Table 2). As with all GWAS, it is important to note that some of the identified SNPs may not be truly "causal", but may be in linkage disequilibrium with causal SNPs—and the genes in the immediate vicinity of such SNPs may not be those through which the phenotype is influenced (e.g., rs72679478 located upstream of the leptin receptor gene). Additionally we showed that there were performance differences between the FDA24 PRS (with 24 SNPs) and FDA5 PRS (with only the five most stable SNPs from FDA24 PRS). Comparing the two, the FDA24 PRS was more effective on the growth curves and conditional weight gain score within our study sample as well as the adult validation cohort. However, FDA5 PRS successfully validated on both child and adult cohorts in spite of the smaller sample size ($N = 283$ for children vs 2486 for adults).

An important advantage of using a score comprising a small number of SNPs, such as ours, is that it is much more practical to compute on individuals belonging to other studies (for comparison purposes) as well as in clinical settings (for screening and potential intervention purposes). If some, potentially several, of the SNPs included in a PRS are not available for an individual, one must choose to either omit them from the calculation or identify and use proxies in their place. The larger the number of SNPs included in a score, the more SNPs may need to be omitted or proxied, reducing the fidelity of the calculation (Chagnon et al., 2018) and the usefulness of the score for prediction. For instance, even having as many as 12.5 million imputed SNPs for our study cohort, we were only able to build the Khera PRS with 37% of the 2 million SNPs included in that score. In principle, this could be one explanation as to why this score did not validate in our dataset.

The power of FDA-based GWAS

Our results demonstrate a key advantage of GWAS employing longitudinal information and FDA techniques over traditional GWAS. We faced an ultra-high dimensional problem—with many more predictors (i.e. SNPs) than observations (i.e. individuals). By integrating FDA techniques into every step of the analysis, from the screening and selection of SNPs through the construction of the PRS, we were able to utilize a more dynamic and information-rich phenotype than the ones used in traditional cross-sectional analyses. In turn, this allowed us to unveil subtler, more complex effects with a limited sample size. We illustrated this with simulations built upon our actual data—to guarantee realistic settings. Using longitudinal data and FDA techniques, a sample size around 200 can be as effective as a cross-sectional study with more than four or five times as many individuals. This potentially expands the scope of GWAS to studies that do not comprise tens of thousands of individuals—but instead hundreds of deeply characterized participants (Reimherr and Nicolae, 2014). With FDA we exploit a series of observations collected longitudinally for each individual, with benefits that include a better understanding of within-subject variability, an estimation of time-varying effects in the form of smooth curves (which reduces noise in the phenotype measurements), and a substantial gain in power. This can be very useful for investigating populations that are difficult to sample—of course with the drawback that collecting longitudinal data requires low participant dropout and may induce other costs or complications. Even so, this approach can provide an effective alternative to traditional GWAS, in light of the costs and benefits of collecting a single measurement on hundreds of thousands of individuals versus longitudinal measurements on a fraction of the subjects.

Other contributing factors and perspectives

Behavioral and environmental factors are important variables to consider when investigating the etiology of complex diseases. In our study we considered 11 such factors that could influence child weight gain trajectories and found that, while the FDA24 PRS is by far the dominant predictor, an appetite score computed on our cohort (see Section 1.1) has a significant effect. It has been shown that a child's appetite behavior impacts early weight gain and may have a strong genetic basis (Llewellyn and Fildes, 2017, Wardle et al., 2008). In agreement with this, a recent study found a positive relationship between a childhood obesity PRS and appetite (Llewellyn et al., 2014). In our study, a child's appetite behavior was reported by their mother—which could have introduced some biases. Because appetite is emerging as an interesting predictor of child weight gain status, it should be explored in more detail in future studies.

In addition to the type of environmental and behavioral factors considered in our study, other factors may interact with genetics in shaping obesity risks. These include the microbiome, the metabolome, and the epigenome. We found previously that children's oral microbiota composition is associated with growth curves (Craig et al., 2018) and that metabolites such as butyrate are linked to child weight outcomes (Nandy et al., 2021). Moreover, we are collecting data on the epigenomes of the children in our study cohort. Our overarching goal is to develop a multi-omic model to comprehensively understand the development of childhood obesity and identify a combination of risk factors that can be used for accurate identification of children who would benefit most from early life intervention programs.

Our FDA-based polygenic risk score was computed considering the longitudinal change in weight-for-length/height ratio from birth through three years of age. An ongoing follow-up of our study participants, with weight and height collected at later time points, will allow us to further evaluate the predictive power of the FDA24 PRS as age progresses. Additionally, we note that our cohort (Table 1), as well as the cohorts of older children and adults used for validation, consisted predominantly of individuals of European ancestry. It will be important to conduct similar analyses on individuals of non-European ancestries, and identify differences and commonalities in the genetic factors contributing to obesity risks among different ethnicities.

The critical advantage of a PRS based on childhood vs. adult weight gain information is that the former is potentially more actionable. While INSIGHT (Paul et al., 2014) was not designed to test an obesity intervention on individuals with high genetic risk, we were able to observe a significant weight gain pattern difference between individuals with high genetic risk in the INSIGHT intervention vs. control groups. To fully understand the clinical implications of using a PRS as a screening tool for obesity intervention additional clinical trials are needed that would combine genetic screening with early life intervention.

Declarations of interest

none

Data Availability

Phenotypic and Genetic data are/will be available under dbGaP study number: phs001498.v2.p1.

Code for carrying out the statistical methods (screening, applying FLAME, PRS construction and evaluation) can be found at https://github.com/makovalab-psu/InsightPRSConstruction.

Ethics Statement

This project has been approved by Penn State IRB (PRAMS 34493).

Funding

This project was supported by grants R01DK88244 and R01DK099354 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Funding was also provided by Penn State Institute for Computational and Data Sciences, Penn State Eberly College of Sciences, and the Huck Institutes of Life Sciences at Penn State. Additionally, this project was funded in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement and CURE funds. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions. Additional funding was provided by NSF DMS 1712826. AK was supported by the NIH 5T32LM012415-03 predoctoral training grant.

Author contributions

SJCC, KDM, IMP, LLB, FC, and MR conceived the project and devised the project study design. AK, SJCC, JL, MR, FC, and KDM were involved in the data analysis. SJCC, AK, KDM, FC, and MR contributed to the writing of the manuscript with comments from co-authors. IMP, LLB, JS, and MM provided resources such as access to the study population and the associated data.

Competing Interests

The authors declare no competing interests.

Acknowledgements

We are grateful to the INSIGHT study participants and nurses for their participation in this project. We would also like to thank B.Higgins, C.Reimer, R. Bruhans, A.Shelly, P.Carper, J.Beiler, J. Stokes, N.Verdiglione, and L.Hess for their assistance. The Philadelphia Neurodevelopment Cohort: Support for the collection of the data for Philadelphia Neurodevelopment Cohort (PNC) was provided by grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson. Subjects were recruited and genotyped through the Center for Applied Genomics (CAG) at The Children's Hospital in Philadelphia (CHOP). Phenotypic data collection occurred at the CAG/CHOP and at the Brain Behavior Laboratory, University of Pennsylvania. eMERGE: The eMERGE Network was initiated and funded by NHGRI through the following grants: U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center). Samples and data in this obesity study were provided by the non-alcoholic steatohepatitis (NASH) project. Funding for the NASH project was provided by a grant from the Clinic Research Fund of Geisinger Clinic. Funding support for the genotyping of the NASH cohort was provided by a Geisinger Clinic operating funds and an award from the Clinic Research Fund. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000380.v1.p1.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2021.10.014.

References

1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74.
Andersson, E.A., et al., 2010. Do gene variants influencing adult adiposity affect birth weight? A population-based study of 24 loci in 4,744 Danish individuals. PLoS One 5, e14190.
Ang, Y.N., Wee, B.S., Poh, B.K., Ismail, M.N, 2012. Multifactorial Influences of Childhood Obesity. Curr. Obes. Rep. 2, 10–22.
Baird, J., et al., 2005. Being big or growing fast: systematic review of size and growth in infancy and later obesity. BMJ 331, 929.
Barriuso, L., et al., 2015. Socioeconomic position and childhood-adolescent weight status in rich countries: a systematic review, 1990-2013. BMC Pediatr 15, 129.
Belsky, D.W., et al., 2012. Polygenic risk, rapid childhood growth, and the development of obesity: evidence from a 4-decade longitudinal study. Arch. Pediatr. Adolesc. Med. 166, 515–521.
Berlinet, A., Thomas-Agnan, C., 2011. Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer Science & Business Media.
edited by Billheimer, D., Ramsay, J.O., Silverman, B.W., 2007. Functional Data Analysis. Biometrics 63, 300–301.
Boney, C.M., Verma, A., Tucker, R., Vohr, B.R., 2005. Metabolic syndrome in childhood: association with birth weight, maternal obesity, and gestational diabetes mellitus. Pediatrics 115, e290–e296.
Calkins, M.E., et al., 2014. The psychosis spectrum in a young U.S. community sample: findings from the Philadelphia Neurodevelopmental Cohort. World Psychiatry 13, 296–305.
Calkins, M.E., et al., 2015. The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. J. Child Psychol. Psychiatry 56, 1356–1369.
Carnell, S., Wardle, J., 2007. Measuring behavioural susceptibility to obesity: Validation of the child eating behaviour questionnaire. Appetite 48, 104–113.
Chagnon, M., O'Loughlin, J., Engert, J.C., Karp, I., Sylvestre, M.-P., 2018. Missing single nucleotide polymorphisms in Genetic Risk Scores: A simulation study. PLoS One 13, e0200630.
Chang, C.C., et al., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7.
Chen, Y., Goldsmith, J., Ogden, R.T., 2016. Variable selection in function-on-scalar regression. Stat 5, 88–101.
Choi, H., Reimherr, M., 2018. A geometric approach to confidence regions and bands for functional parameters. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80, 239–260.
Chu, W., Li, R., Reimherr, M, 2016. Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data. The Annals of Applied Statistics 10, 596–617.
Craig, S.J.C., et al., 2018. Child Weight Gain Trajectories Linked To Oral Microbiota Composition. Sci. Rep. 8, 14030.
Cremona, M.A., et al., 2019. Functional data analysis for computational biology. Bioinformatics 35, 3211–3213.
Cunningham, S.A., Kramer, M.R., Narayan, K.M.V, 2014. Incidence of childhood obesity in the United States. The New England journal of medicine 370, 1660–1661.
Daniels, S.R., Hassink, S.G., 2015. COMMITTEE ON NUTRITION. The Role of the Pediatrician in Primary Prevention of Obesity. Pediatrics 136, e275–e292.
Delaneau, O., Marchini, J., Zagury, J.-F., 2011. A linear complexity phasing method for thousands of genomes. Nat. Methods 9, 179–181.
den Hoed, M., et al., 2010. Genetic susceptibility to obesity and related traits in childhood and adolescence: influence of loci identified by genome-wide association studies. Diabetes 59, 2980–2988.
Elks, C.E., et al., 2010. Genetic markers of adult obesity risk are associated with greater early infancy weight gain and growth. PLoS Med 7, e1000284.
Fall, T., Ingelsson, E., 2014. Genome-wide association studies of obesity and metabolic syndrome. Mol. Cell. Endocrinol. 382, 740–757.
Fan, J., Li, R, 2001. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association 96, 1348–1360.
Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 849–911.
Fan, J., Ma, Y., Dai, W, 2014. Nonparametric Independence Screening in Sparse Ultra-High Dimensional Varying Coefficient Models. J. Am. Stat. Assoc. 109, 1270–1284.
Fan, J., Samworth, R., Wu, Y., 2009. Ultrahigh dimensional feature selection: beyond the linear model. J. Mach. Learn. Res. 10, 2013–2038.
Fan, Z., Reimherr, M., 2017. High-dimensional adaptive function-on-scalar regression. Econometrics and Statistics 1, 167–183.
Felix, J.F., et al., 2016. Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. Hum. Mol. Genet. 25, 389–403.
Frayling, T.M., et al., 2007. A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science 316, 889–894.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1.
Gertheiss, J., Maity, A., Staicu, A.-M., 2013. Variable Selection in Generalized Functional Linear Models. Stat 2, 86–103.
Glessner, J.T., et al., 2010. Strong synaptic transmission impact by copy number variations in schizophrenia. Proc. Natl. Acad. Sci. U. S. A. 107, 10584–10589.
Goldsmith, J., Schwartz, J.E, 2017. Variable selection in the functional linear concurrent model. Stat. Med. 36, 2237–2250.
Goodarzi, M.O., 2018. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. Lancet Diabetes Endocrinol 6, 223–236.
Graff, M., et al., 2013. Genome-wide analysis of BMI in adolescents and young adults reveals additional insight into the effects of genetic loci over the life course. Human Molecular Genetics 22, 3597–3607.
Griffiths, L.J., Smeeth, L., Hawkins, S.S., Cole, T.J., Dezateux, C., 2009. Effects of infant feeding practice on weight gain from birth to 3 years. Arch. Dis. Child. 94, 577–582.
Hales, C.M., Fryar, C.D., Carroll, M.D., Freedman, D.S., Ogden, C.L, 2018. Trends in Obesity and Severe Obesity Prevalence in US Youth and Adults by Sex and Age, 2007-2008 to 2015-2016. JAMA 319, 1723–1725.
Hall, P., Miller, H, 2009. Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems. J. Comput. Graph. Stat. 18, 533–550.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Series in Statistics.
Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications. Springer Science & Business Media.
Howie, B.N., Donnelly, P., Marchini, J, 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5, e1000529.
Hsing, T., Eubank, R., 2015. Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics doi:10.1002/9781118762547.
Huang, C., et al., 2017. FGWAS: Functional genome wide association analysis. Neuroimage 159, 107–121.
Huang, J.Z., Wu, C.O., Zhou, L, 2004. POLYNOMIAL SPLINE ESTIMATION AND INFERENCE FOR VARYING COEFFICIENT MODELS WITH LONGITUDINAL DATA. Stat. Sin. 14, 763–788.
Justice, A.E., et al., 2019. Genetic determinants of BMI from early childhood to adolescence: the Santiago Longitudinal Study. Pediatr. Obes. 14, e12479.
Khera, A.V., et al., 2019. Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. Cell 177, 587–596 e9.
Kokoszka, P., Reimherr, M., 2017. Introduction to Functional Data Analysis. CRC Press.
Kries, R.von, von Kries, R., 2002. Maternal Smoking during Pregnancy and Childhood Obesity. American Journal of Epidemiology 156, 954–961.
Li, A., et al., 2018. Parental and child genetic contributions to obesity traits in early life based on 83 loci validated in adults: the FAMILY study. Pediatr. Obes. 13, 133–140.
Li, R., Zhong, W., Zhu, L, 2012. Feature Screening via Distance Correlation Learning. J. Am. Stat. Assoc. 107, 1129–1139.

Liu, J., Li, R., Wu, R, 2014. Feature Selection for Varying Coefficient Models With Ultrahigh Dimensional Covariates. J. Am. Stat. Assoc. 109, 266–274.
Llewellyn, C.H., Fildes, A., 2017. Behavioural Susceptibility Theory: Professor Jane Wardle and the Role of Appetite in Genetic Risk of Obesity. Curr. Obes. Rep. 6, 38–45.
Llewellyn, C.H., Trzaskowski, M., Plomin, R., Wardle, J., 2014. From modeling to measurement: developmental trends in genetic influence on adiposity in childhood. Obesity 22, 1756–1761.
Llewellyn, C.H., Trzaskowski, M., van Jaarsveld, C.H.M., Plomin, R., Wardle, J., 2014. Satiety mechanisms in genetic risk of obesity. JAMA Pediatr 168, 338–344.
Llewellyn, C.H., van Jaarsveld, C.H.M., Plomin, R., Fisher, A., Wardle, J., 2012. Inherited behavioral susceptibility to adiposity in infancy: a multivariate genetic analysis of appetite and weight in the Gemini birth cohort. Am. J. Clin. Nutr. 95, 633–639.
Locke, A.E., et al., 2015. Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197–206.
Lumley, T., Lumley, M.T., 2013. Package 'leaps'. *Regression Subset Selection*. Thomas Lumley Based on Fortran Code by Alan Miller. *Available online* http://CRAN.R-project.org/package=leaps . *(accessed on 18 March 2018)*.
Machiela, M.J., Chanock, S.J., 2015. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31, 3555–3557.
Maes, H.H., Neale, M.C., Eaves, L.J, 1997. Genetic and environmental factors in relative body weight and human adiposity. Behav. Genet. 27, 325–351.
Melén, E., et al., 2010. Analyses of shared genetic factors between asthma and obesity in children. Journal of Allergy and Clinical Immunology 126, 631–637 e8.
Meyre, D., et al., 2009. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nat. Genet. 41, 157–159.
Mousavi, S.N., Sørensen, H., 2017. Multinomial functional regression with wavelets and LASSO penalization. Econometrics and Statistics 1, 150–166.
O'Connell, J., et al., 2014. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet 10, e1004234.
Ogden, C.L., et al., 2016. Trends in Obesity Prevalence Among Children and Adolescents in the United States, 1988-1994 Through 2013-2014. JAMA 315, 2292–2299.
Ong, K.K., Loos, R.J.F, 2006. Rapid infancy weight gain and subsequent obesity: systematic reviews and hopeful suggestions. Acta Paediatr 95, 904–908.
Park, S.Y., Xiao, L., Willbur, J.D., Staicu, A.-M., Jumbe, N.L, 2018. A joint design for functional data with application to scheduling ultrasound scans. Comput. Stat. Data Anal. 122, 101–114.
Parodi, A., Reimherr, M, 2018. Simultaneous variable selection and smoothing for high-dimensional function-on-scalar regression. Electron. J. Stat. 12, 4602–4639.
Paul, I.M., et al., 2014. The Intervention Nurses Start Infants Growing on Healthy Trajectories (INSIGHT) study. BMC Pediatr 14, 184.
Paul, I.M., et al., 2018. Effect of a Responsive Parenting Educational Intervention on Childhood Weight Outcomes at 3 Years of Age: The INSIGHT Randomized Clinical Trial. JAMA 320, 461–468.
Peters, U., Dixon, A.E., Forno, E., 2018. Obesity and asthma. J. Allergy Clin. Immunol. 141, 1169–1179.
Pigeyre, M., Yazdi, F.T., Kaur, Y., Meyre, D., 2016. Recent progress in genetics, epigenetics and metagenomics unveils the pathophysiology of human obesity. Clin. Sci. 130, 943–986.
Purcell, S., et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.
Ramsay, J., Hooker, G. & Graves, S. Functional Data Analysis with R and MATLAB. (2009) doi:10.1007/978-0-387-98185-7.
Reimherr, M., Nicolae, D, 2014. A functional data analysis approach for genetic association studies. The Annals of Applied Statistics 8, 406–429.
Saeed, S., et al., 2018. Loss-of-function mutations in ADCY3 cause monogenic severe obesity. Yearbook of Paediatric Endocrinology doi:10.1530/ey.15.11.5.
Sahoo, K., et al., 2015. Childhood obesity: causes and consequences. J Family Med Prim Care 4, 187–192.
Savage, J.S., Birch, L.L., Marini, M., Anzman-Frasca, S., Paul, I.M., 2016. Effect of the INSIGHT Responsive Parenting Intervention on Rapid Infant Weight Gain and Overweight Status at Age 1 Year: A Randomized Clinical Trial. JAMA Pediatr 170, 742–749.
Shao, X., Zhang, J, 2014. Martingale Difference Correlation and Its Use in High-Dimensional Variable Screening. J. Am. Stat. Assoc. 109, 1302–1318.
Song, R., Yi, F., Zou, H, 2014. On Varying-coefficient Independence Screening for High-dimensional Varying-coefficient Models. Stat. Sin. 24, 1735–1752.
Sovio, U., et al., 2011. Association between common variation at the FTO locus and changes in body mass index from infancy to late childhood: the complex nature of genetic association through growth and development. PLoS Genet 7, e1001307.
Speliotes, E.K., et al., 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. 42, 937–948.
Sugrue, L.P., Desikan, R.S., 2019. What Are Polygenic Scores and Why Are They Important? JAMA doi:10.1001/jama.2019.3893.
Taveras, E.M., et al., 2009. Weight Status in the First 6 Months of Life and Obesity at 3 Years of Age. PEDIATRICS 123, 1177–1183.
the Early Growth Genetics (EGG) Consortium, 2012. A genome-wide association meta-analysis identifies new childhood obesity loci. Nat. Genet. 44, 526–531.
Thorleifsson, G., et al., 2009. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat. Genet. 41, 18–24.
Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc. Series B Stat. Methodol. 73, 273–282.
Vogelezang, S., et al., 2020. Novel loci for childhood body mass index and shared heritability with adult cardiometabolic traits. PLoS Genet 16, e1008718.
Vsevolozhskaya, O.A., et al., 2016. Uncovering Local Trends in Genetic Effects of Multiple Phenotypes via Functional Linear Models. Genet. Epidemiol. 40, 210–221.
Wardle, J., et al., 2008. Obesity Associated Genetic Variation inFTOIs Associated with Diminished Satiety. The Journal of Clinical Endocrinology & Metabolism 93, 3640–3643.
Warrington, N.M., et al., 2015. A genome-wide association study of body mass index across early life and childhood. Int. J. Epidemiol. 44, 700–712.
Wheeler, E., et al., 2013. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat. Genet. 45, 513–517.
Wrobel, J., Zipunnikov, V., Schrack, J., Goldsmith, J., 2019. Registration for exponential family functional data. Biometrics 75, 48–57.
Yao, F., Müller, H.-G., Wang, J.-L, 2005. Functional Data Analysis for Sparse Longitudinal Data. J. Am. Stat. Assoc. 100, 577–590.
Zhang, T., et al., 2019. Rate of change in body mass index at different ages during childhood and adult obesity risk. Pediatr. Obes. 14, e12513.
Zhou, J., et al., 2016. Rapid Infancy Weight Gain and 7- to 9-year Childhood Obesity Risk: A Prospective Cohort Study in Rural Western China. Medicine 95, e3425.
Zou, H., 2006. The Adaptive Lasso and Its Oracle Properties. J. Am. Stat. Assoc. 101, 1418–1429.