Original Article

Prediction of Adulthood Obesity Using Genetic and Childhood Clinical Risk Factors in the Cardiovascular Risk in Young Finns Study

Fatemeh Seyednasrollah, MSc;* Johanna Mäkelä, PhD;* Niina Pitkänen, PhD; Markus Juonala, MD, PhD; Nina Hutri-Kähönen, PhD; Terho Lehtimäki, PhD; Jorma Viikari, MD, PhD; Tanika Kelly, PhD; Changwei Li, PhD; Lydia Bazzano, PhD; Laura L. Elo, PhD;* Olli T. Raitakari, MD, PhD*

Background—Obesity is a known risk factor for cardiovascular disease. Early prediction of obesity is essential for prevention. The aim of this study is to assess the use of childhood clinical factors and the genetic risk factors in predicting adulthood obesity using machine learning methods.

Methods and Results—A total of 2262 participants from the Cardiovascular Risk in YFS (Young Finns Study) were followed up from childhood (age 3–18 years) to adulthood for 31 years. The data were divided into training (n=1625) and validation (n=637) set. The effect of known genetic risk factors (97 single-nucleotide polymorphisms) was investigated as a weighted genetic risk score of all 97 single-nucleotide polymorphisms (WGRS97) or a subset of 19 most significant single-nucleotide polymorphisms (WGRS19) using boosting machine learning technique. WGRS97 and WGRS19 were validated using external data (n=369) from BHS (Bogalusa Heart Study). WGRS19 improved the accuracy of predicting adulthood obesity in training (area under the curve [AUC=0.787 versus AUC=0.744, $P<0.0001$) and validation data (AUC=0.769 versus AUC=0.747, $P=0.026$). WGRS97 improved the accuracy in training (AUC=0.782 versus AUC=0.744, $P<0.0001$) but not in validation data (AUC=0.749 versus AUC=0.747, $P=0.785$). Higher WGRS19 associated with higher body mass index at 9 years and WGRS97 at 6 years. Replication in BHS confirmed our findings that WGRS19 and WGRS97 are associated with body mass index.

Conclusions—WGRS19 improves prediction of adulthood obesity. Predictive accuracy is highest among young children (3–6 years), whereas among older children (9–18 years) the risk can be identified using childhood clinical factors. The model is helpful in screening children with high risk of developing obesity. **(*Circ Cardiovasc Genet*. 2017;10:e001554. DOI: 10.1161/CIRCGENETICS.116.001554.)**

Key Words: genetics ■ machine learning ■ obesity ■ risk factor ■ single-nucleotide polymorphism genetics ■ statistics

Obesity is a risk factor contributing to severe health problems including hypertension, cardiovascular diseases, type 2 diabetes mellitus and cancer.[1–4] Childhood clinical and environmental factors are known to influence obesity risk later in life. Environmental factors known to increase obesity risk include among others high childhood body mass index (BMI),[5] parental obesity,[6] maternal smoking,[7] chronic insufficient sleep during childhood,[8] and low socioeconomic class.[9] The familial clustering of obesity is due to both genetic factors and shared environmental factors.[10]

BMI and obesity are complex traits influenced by many environmental and genetic factors. Estimates of BMI heritability range from 40% to 70%,[11,12] and genomewide association studies have identified several loci for BMI and obesity.[13–17] For instance, a recent study reported 97 single-nucleotide polymorphisms (SNPs) associated with BMI in over 100 000 genotyped individuals.[14]

See Editorial by Belsky
See Clinical Perspective

Despite substantial heritability, the effects of individual variants on BMI and obesity risk are modest. One crucial need alongside the identification of new loci is to investigate whether these variants individually or together can be used in predicting

Received September 14, 2016; accepted December 06, 2016.

From the Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Finland (F.S., J.M., L.L.E.); Department of Mathematics and Statistics (F.S.), Research Centre of Applied and Preventive Cardiovascular Medicine (N.P., O.T.R.), and Department of Medicine (M.J., J.V.), University of Turku, Finland; Division of Medicine (M.J., J.V.) and Clinical Physiology and Nuclear Medicine (O.T.R.), Turku University Hospital, Finland; Department of Pediatrics (N.H.-K.) and School of Medicine (T.L.), University of Tampere, Finland; Tampere University Hospital, Finland (N.H.-K.); Department of Clinical Chemistry, Fimlab Laboratories, Tampere, Finland (T.L.); Tulane University Health Sciences Center, New Orleans, LA (T.K., L.B.); and Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens (C.L.).

*F. Seyednasrollah and Dr Mäkelä contributed equally to this work as first authors. Drs Elo and Raitakari contributed equally as senior authors.

The Data Supplement is available at http://circgenetics.ahajournals.org/lookup/suppl/doi:10.1161/CIRCGENETICS.116.001554/-/DC1.

Correspondence to Johanna Mäkelä, PhD, Turku Center for Biotechnology, Tykistökatu 6 A, 20520 Turku, Finland. E-mail jolepp@utu.fi

adulthood obesity. In particular, computational techniques are needed to fully integrate the importance of genetic risk factors in predictive analysis. This can be done through defining weighted genetic risk scores[18,19] or by applying computational techniques capable of capturing nonlinear effects of SNPs.[20] The main advantage of developing an early predictive approach for obesity risk is that it would enable early intervention and preventive measures to those individuals with highest risk.

In this study, we assess the use of recently identified genetic risk variants and childhood environmental and clinical factors in predicting adulthood obesity. Our objective is to create a predictive model for adulthood obesity using ensemble-based machine learning technique and validate the predictive accuracy of our model with an independent validation set. Previously, we have shown that constructing a multifactorial model is more efficient in predicting adulthood obesity than single risk factors.[21] The previous study found no clear support for the hypothesis that genetic risk variants would improve the prediction of obesity in adulthood above clinical childhood risk factors. However, the number of genetic risk variants was low compared with the number of loci currently known to be associated with obesity. Here we use the same clinical and environmental factors for prediction as in our previous study, combined with updated genetic information from the recent genomewide association studies.[14] In addition to updated genetic information, we have now longer follow-up and more detailed clinical data. In this study, we aim to reanalyze the value of genetic factors in obesity prediction. We hypothesize that the new genetic factors together with longer follow-up of the participants will introduce more power to our analysis.

Methods

Study Cohort

The Cardiovascular Risk in YFS (Young Finns) study is an on-going population-based follow-up study of cardiovascular risk factors. In 1980, altogether 3596 participants aged 3 to 18 years were examined in five Finnish cities and their surroundings. Subsequently, follow-up studies have been conducted regularly in year 2001 (n=2283), 2007 (n=2204), and 2011 to 2012 (n=2060). The follow-up time in this study was 31 or 32 years. The data used in this study for obesity risk prediction consists of participants from whom a complete set of baseline variables (age, sex, baseline BMI, maternal BMI, and family income), genotype data, and outcome (adult BMI) were available (n=2262, 62.9%). All participants gave written informed consent and the local ethics committees approved the study.

Experimental Design

We examined the association of genetic risk factors with childhood environmental and clinical factors, later referred as clinical factors, with adulthood obesity. The study cohort was randomly divided into a training data (n=1625, 72%) and an independent validation data (n=637, 28%). The model was built on the training data, and its performance was evaluated using the independent validation part. Different variable sets and predictive strategies were examined to capture the best performing model. Association between childhood clinical factors and adulthood obesity has been thoroughly investigated in our previous study.[21] Accordingly, in this study, we only focused on the already proved significant nongenetic, that is, clinical risk factors. More specifically, we selected the variables with P value ≤0.05 in the previously published predictive models. These covariates included childhood BMI SD-score, maternal BMI, family income and C-reactive protein (CRP) measured in childhood. While testing the model, we discovered that CRP did not influence the performance of the model. As the exclusion of CRP resulted in higher number of participants to be included in this study, it was therefore excluded from the final model.

Study Variables

The prediction outcome, adulthood obesity, was defined as BMI ≥ 30 (kg/m^2). Height and weight of the participants were measured during their visit in 2011, and BMI was calculated as weight (kg) divided by the square of height (m^2). For those participants with missing BMI information from year 2011, we used the information from study visits conducted in year 2007 or 2001 if available. Baseline clinical features are presented in Table 1.

Childhood BMI was measured during the first follow-up examination. Childhood BMI was adjusted for age and sex by calculating a BMI SD-score in statistical analysis. Family income and maternal BMI were collected from self-reported questionnaires at baseline. If the mother was pregnant, her prepregnancy BMI was used. For the genetic risk factors, we examined 97 SNPs derived from the previous study by Locke et al.[14] In our previous risk prediction study,[21] we used 31 SNPs from which 25 were included in the study by Locke et al,[14] and thus were included here as well. Genotyping was performed using a custom-built Illumina Human 670k BeadChip. Genotypes were called using Illumina's clustering algorithm.[22] Genotype imputation was performed using IMPUTE2 software[23] and the 1000G Phase I Integrated Release Version 3 as a reference panel.[24]

Statistical Analyses and Model Building

Baseline characteristics between obese and nonobese participants were studied using Wilcoxon rank-sum test for continuous and χ^2 test for categorical variables. Univariate regression analysis was performed for the 97 SNPs derived from the previous study by Locke et al.[14] (Table I in the Data Supplement). In the predictive analysis, genetic risk factors were used as a weighted genetic risk score, which was defined as the arithmetic sum of the SNP values x_i (effective alleles) weighted by their corresponding β scores β_i provided by the original study[14]:

$$\mathrm{WGRS97} = \sum_{i=1}^{97} \beta_i x_i$$

In addition to WGRS97, we calculated an obesity prediction score based on the most significant SNPs (P<0.1) in the univariate regression analysis (Table 2). The final predictive weighted genetic risk score consisted of 19 SNPs (WGRS19) from the total of 97 BMI-associated SNPs previously defined by Locke et al.[14] and was defined similarly as the WGRS97 using the corresponding β scores β_i provided by the original study[14]:

$$\mathrm{WGRS19} = \sum_{i=1}^{19} \beta_i x_i$$

The final clinical model included baseline childhood BMI SD-score, maternal BMI, and family income, whereas the final clinical and genetic model included the same variables and in addition WGRS97 or WGRS19.

In the model building, the outcome adulthood obesity (BMI≥30 kg/m^2) was considered as a binary outcome. Gradient boosting algorithm[25,26] implemented in the R package gbm was used to build the predictive model. Gradient boosting is an ensemble learning technique that builds a final strong model over several middle-step weak models. More specifically, it sequentially trains a series of weak models using a base learning algorithm so that in each successive model, it gives more weight to those observations that were misclassified in the previous models. The final strong model is then generated as a weighted combination of the base models. Several studies have demonstrated the robust performance of gradient boosting machines.[27,28] Boosting methods can account for nonlinear effects of used variables, incorporate variable interactions, and handle outliers and missing data values.[29] Regularizing parameters, including learning rate and

Table 1. Baseline Characteristics of Participants in the Cardiovascular Risk in YFS (Young Finns Study) and BHS (Bogalusa Heart Study)

	YFS			BHS		
	Obesity in Adulthood			Obesity in Adulthood		
Variable	No	Yes	*P*-Value*	No	Yes	*P*-Value*
n	1814	448		219	150	
Female, %	54.4	51.8	0.356	54.8	46.0	0.112
Baseline age, y	10.3 (5.1)	11.2 (4.7)	0.0006	9.7 (3.3)	9.3 (2.3)	0.166
Baseline BMI, kg/m^2	17.4 (2.7)	19.7 (3.7)	<0.0001	16.8 (2.5)	18.9 (3.7)	<0.0001
Baseline BMI SD-score	-0.18 (0.83)	0.68 (1.2)	<0.0001	-0.02 (0.97)	0.96 (1.4)	<0.0001
Baseline maternal BMI, kg/m^2	23.6 (3.6)	25. (4.4)	<0.0001	NA	NA	
Adult BMI, kg/m^2	24.4 (2.9)	34.2 (4.2)	<0.0001	25.2 (3.1)	36.0 (5.9)	<0.0001
WGRS19	0.50 (0.08)	0.54 (0.08)	<0.0001	0.50 (0.07)	0.53 (0.09)	0.002

Values are presented as mean (SD). BMI indicates body mass index.
*Wilcoxon rank-sum test for continuous and χ^2 test for categorical variables.

subsampling fraction, were fine-tuned to eliminate model sensitivity because of low-dimensional setting of the training data as suggested by previous studies.[25,30] Cross-validation was used to penalize model overfitting. Finally, the discriminative performance of the model was evaluated by the area under the receiver operating characteristic curve (AUC) in both the training and validation data. The AUC values between the receiver operating characteristic curves were compared with DeLong method.

To examine the effect of WGRS19 on BMI trajectories, the participants were initially divided into 4 groups based on their WGRS19 quartiles. The BMI trajectories represent the mean values of BMI from all participants with measured BMI at each given age in each quartile. For clarity, only the highest and lowest quartiles are presented. Statistical comparison between the highest and lowest quartiles was performed with Wilcoxon rank-sum test separately at each age.

Validation of the Results

External validation of the weighted genetic risk scores (WGRS19 and WGRS97) was performed using data from the BHS (Bogalusa Heart

Table 2. Univariate Logistic Regression Analysis Results for SNPs With *P*<0.1

SNP	Chr	Position	Gene	OR	*P*-Value	Ranking in Locke et al[14]
rs2207139	6	50 953 449	TFAP2B	1.326	0.001	6
rs16951275	15	65 864 222	MAP2K5	0.727	0.004	15
rs2112347	5	75 050 998	POC5	0.809	0.006	17
rs1558902	16	52 361 075	FTO	1.203	0.014	1
rs492400	2	219 057 996	USP37	0.835	0.015	56
rs11688816	2	62 906 552	EHBP1	0.889	0.019	73
rs7239883	18	38 401 669	LOC284260	1.321	0.035	69
rs9400239	6	109 084 356	FOXO3	1.171	0.042	71
rs16851483	3	142 758 126	RASA2	1.082	0.0491	44
rs9641123	7	93 035 668	CALCR	1.183	0.0762	42
rs17724992	19	18 315 825	PGPEP1	0.847	0.078	89
rs2245368	7	76 446 079	PMS2L11	0.855	0.079	87
rs1460676	2	164 275 935	FIGN	1.134	0.081	96
rs11727676	4	145 878 514	HHIP	0.845	0.0849	79
rs7164727	15	70 881 044	LOC100287559	1.197	0.088	53
rs13078960	3	85 890 280	CADM2	1.134	0.09	21
rs9374842	6	120 227 364	LOC285762	1.145	0.0906	81
rs11057405	12	121 347 850	CLIP1	1.05	0.0922	74
rs12885454	14	28 806 589	PRKD1	0.881	0.096	41

Adult obesity (BMI≥30 kg/m^2) was used as an outcome. BMI indicates body mass index; and SNP, single-nucleotide polymorphism.

Study).[31] From the total cohort with genetic data available (n=737), we included white subjects (n=369) because the YFS included only white participants. The follow-up time in BHS was 36 years. To examine the effect of WGRS19 and WGRS97 on BMI trajectories at different ages, the participants were divided into 4 groups based on their WGRS19 and WGRS97 quartiles. The analysis of BMI trajectories was performed similarly as in YFS.

Results

Baseline Characteristics

From all the Cardiovascular Risk in YFS participants (n=3596), participants with complete set of baseline features were selected (n=2262, 62.9%). The follow-up time for these participants was 21 to 31 years.

Table 1 shows the baseline characteristics of the participants. Obese participants differed from the other participants in regards to childhood baseline BMI, childhood baseline BMI SD-score, maternal BMI, adult BMI and WGRS19. Participants who were obese in adulthood had higher BMI and BMI SD-score in childhood than the other participants (P<0.0001 for both). Maternal BMI at baseline was higher for those participants who became obese in adulthood when compared with the other participants (P<0.0001). Moreover, obese participants had higher WGRS19 than the other participants (P<0.0001). The participants in the training set did not differ from the participants in the validation set (P>0.05 for all).

Results from the univariate regression analysis of the individual SNPs are shown in Table I in the Data Supplement. In the univariate regression analysis of the total of 97 SNPs, 19 SNPs had P values <0.1 and were selected to the final weighted genetic risk score (WGRS19, Table 2).

Associations of Childhood Clinical and Genetic Risk Factors With Adulthood Obesity

WGRS19 improved the prediction of obesity in the training (AUC=0.787 versus AUC=0.744, P<0.0001) and validation (AUC=0.769 versus AUC=0.747, P=0.026) data when compared with clinical factors alone (Table 3). When examining the performance of the model in different age groups, the genetic factors improved the prediction in the youngest (3–6 years) age group both in training (AUC=0.754 versus AUC=0.692, P<0.0001) and in validation data sets (AUC=0.771 versus AUC=0.700, P=0.002). For older children (9–12 years and 15–18 years) statistically significant differences in predictive accuracy between genetic and clinical factors were seen only in the training data (P<0.01 for both), but not in the validation data (P>0.293 for both). WGRS97 improved the prediction in the training data (P<0.001 for all), but not in the validation data except in the youngest 3 to 6-year-old children (P=0.020). In children aged 9 to 12 years, the prediction accuracy of clinical factors was better than WGRS97 (P=0.049). The receiver operating characteristic curves for all participants are presented in Figure 1 for the training and validation data separately. The receiver operating characteristic curves indicate prediction accuracy of clinical factors and combined genetic and clinical factors among training and validation set.

We further examined the association of the WGRS19 and WGRS97 with the development of BMI across different age points by dividing the participants into quartiles based on the WGRS19 or WGRS97 scores (Figure 2). Comparison between the highest and lowest quartiles revealed that the participants with the highest and lowest genetic risk of obesity had statistically significantly different BMI trajectories starting from the age of 9 years (P<0.05, Kruskal–Wallis test). At the age of 3 and 6 years, no differences between any of the groups were discovered (P>0.05 for all).

Replication of the Findings

To validate our results externally, we used the data from BHS white participants (n=369). Unfortunately, the model could not be used for the risk predictions because of different study designs with different baseline variables available. However, we examined the association of the WGRS19 and WGRS97 with the development of BMI across different age points by dividing the participants into quartiles based on the WGRS19 or WGRS97 scores. It was discovered that participants with higher genetic risk of obesity had higher BMI (Figure 3). These results support our findings that the weighted genetic

Table 3. Results From Boosting Analysis of Genetic Risk Scores (WGRS19 and WGRS97) and Clinical Factors Compared With Clinical Factors for the Training (n=1625) and Validation (n=637) Data

Age	WGRS19 and Clinical Factors AUC	WGRS97 and Clinical Factors AUC	Clinical Factors AUC	WGRS19 vs Clinical *P*-Value	WGRS97 vs Clinical *P*-Value
Training Data (n=1625)					
All	0.787	0.782	0.744	<0.0001*	<0.0001*
3–6 y	0.754	0.736	0.692	<0.0001*	<0.0001*
9–12 y	0.809	0.805	0.777	0.002*	<0.0001*
15–18 y	0.793	0.795	0.752	0.001*	<0.0001*
Validation data (n=637)					
All	0.769	0.749	0.747	0.026*	0.785
3–6 y	0.771	0.742	0.700	0.002*	0.020*
9–12 y	0.799	0.767	0.786	0.293	0.049*
15–18 y	0.734	0.738	0.740	0.743	0.922

AUC indicates area under the curve.

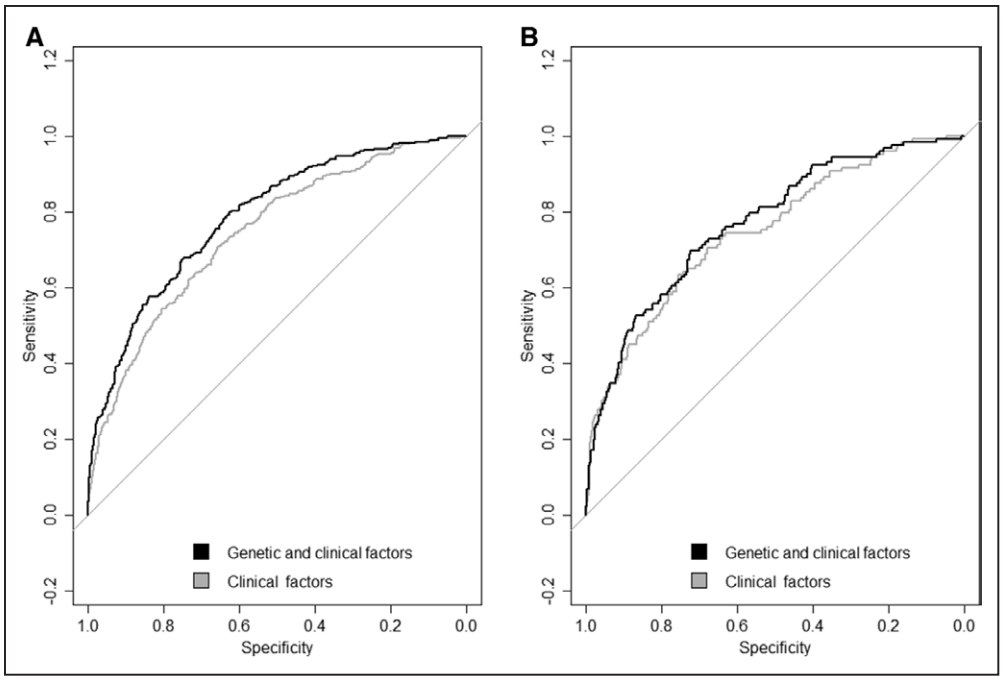

Figure 1. Receiver operating characteristic curves for all participants in the (**A**) training and (**B**) validation data representing prediction accuracy of WGRS19 and childhood clinical factors for obesity in adulthood.

risk scores calculated from BMI-associated SNPs are predicting later obesity.

Discussion

This longitudinal study demonstrates that the inclusion of both genetic and childhood clinical factors increases the accuracy of adulthood obesity prediction. Our results show that a model constructed from previously established genetic risk factors and childhood clinical factors predicts adulthood obesity. The impact of genetic factors on the prediction accuracy is highest in young children. In addition, we validated the weighted genetic risk scores using external validation data from the BHS.

Previously, it has been shown that constructing a multifactorial model is more efficient in predicting adulthood obesity than single risk factors.[21] Here we constructed a model using updated genetic information and childhood clinical factors in addition to longer follow-up than previously.[21] The measurement of genetic risk factors is easy and they remain unchanged during lifetime. However, despite the role of genetic factors in the risk of obesity, becoming obese requires a larger energy intake than expenditure which means that environmental factors also have a significant role. The clinical and environmental factors, however, change during life course and have varying effects on disease outcomes. Early life is known to influence health and disease risk later in life.[32] There is still limited data on how the childhood clinical and environmental factors associate with adulthood obesity risk. The childhood clinical and environmental factors studied here were chosen based on results from a previous study where childhood BMI, maternal BMI, family income and childhood CRP were significant predictors of obesity in adulthood.[21] The clinical and environmental factors included in the models were the same except for childhood CRP that was not significantly improving the predictive accuracy of our new model and was thus excluded from the final model. Also other previous studies have highlighted the importance of childhood BMI,[33] maternal BMI[6] and family income[9] in predicting obesity in adulthood.

The genetic factors included in our final model were defined by genomewide association studies.[13,14] Compared with our previous risk prediction study,[21] we now studied a higher number of SNPs (total 97 SNPs) and constructed genetic risk scores WGRS97 and WGRS19. The WGRS19 was calculated based on the 19 most significant SNPs in this cohort. We used advanced machine-learning techniques and performed careful internal validation in an independent subset of the data that was completely held out when building the model. The new model suggests the use of genetic factors in predicting adulthood obesity in children. The predictive accuracy of WGRS97 was lower than that of WGRS19. This may be because of the fact that some of the 97 SNPs that associate with adulthood BMI may affect BMI already during early childhood. Therefore, this predictive use over clinical factors decreases because the genetic effect on BMI is already manifested. Accordingly, here the best predictors were not the SNPs with strongest influence on BMI based on previous study[14] and the best predictors were the SNPs whose effect on phenotype had not yet manifested.

The main advantage of this study is the high-quality longitudinal data used together with advanced computational methods with careful cross-validation to avoid overfitting. We used AUC as statistical measure for evaluating the performance of our obesity risk prediction models. The AUC reflects the overall performance of the model and is regarded as a standard measure of the effect of a new marker in risk prediction. The validation with an independent subsample of the participants confirmed the use of our model and the genetic risk factors in predicting adulthood obesity in all children. The external

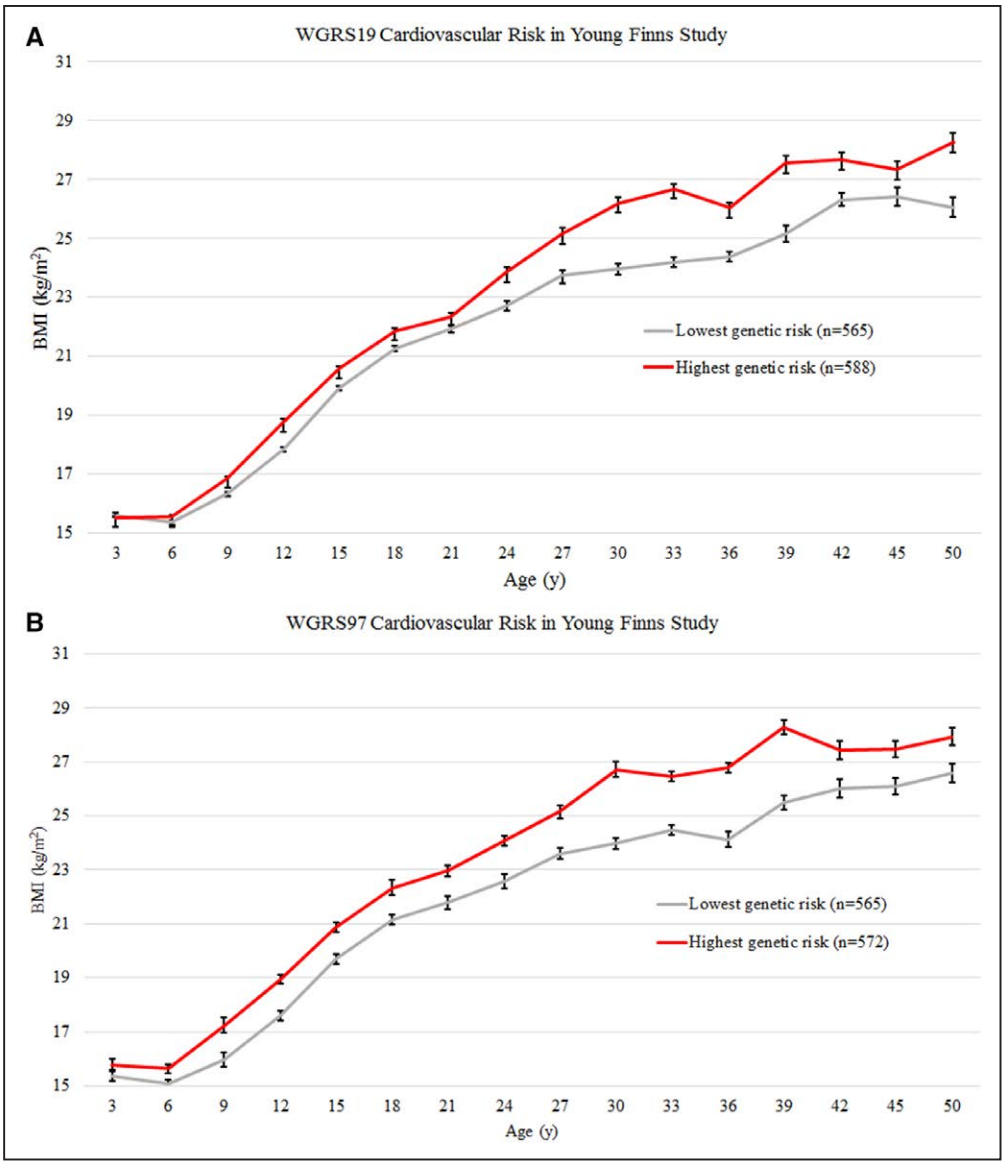

Figure 2. The body mass index (BMI) trajectories with mean value and SEM bars in participants of the Cardiovascular Risk in YFS (Young Finns Study) with low or high genetic risk of obesity according to lowest and highest weighted genetic risk score quartile from 3 to 50 years according to (**A**) WGRS19 or (**B**) WGRS97 quartiles. Statistically significant differences ($P<0.05$) were seen between participants with high genetic risk compared with participants with low genetic risk starting from the age of 9 and 6 years, respectively.

validation with BHS white participants confirmed that our findings are not restricted to our cohort only.

Although genetic variants are associated with the incidence of obesity, only limited added clinical value has been reported in the prediction of obesity in adulthood.[21] Genetic variants may be better predictors in younger individuals and even over longer follow-up periods. We have previously shown for other outcomes, such as adult hypertension,[34] dyslipidemia,[35] fatty liver[36] and type 2 diabetes mellitus,[37] that genetic information provides predictive information in addition to nongenetic childhood risk factors. In this study we confirm similar findings in adulthood obesity prediction. As the costs for determining genetic factors is rapidly decreasing, in the future specific genetic factors can be used even in clinical practice to determine the individuals most susceptible to later obesity. These highly susceptible individuals can then be addressed with more intensive obesity prevention measures.

It is likely that genetic information may help to identify individuals with high risk of obesity in early life when other risk factors have not yet developed. Our results are in line with this hypothesis by demonstrating the genetic factors that provide additional information over clinical risk factors in identifying children who are at risk of developing obesity in adulthood, although clinical factors show a substantial independent predictive role as well. Our results show that the prediction of adulthood obesity differs depending on the baseline age and that the effect of genetic factors on predictive accuracy is higher among young children. Adding WGRS19 to the model increased the overall AUC from 0.747 to 0.769. The SNPs that had the strongest influence on adulthood obesity in this study (WGRS19) may be SNPs whose effect into BMI or weight has not yet established in early childhood. In our study, the most obesity-predictive SNPs were not the same as the most obesity-associated SNPs because the most predictive

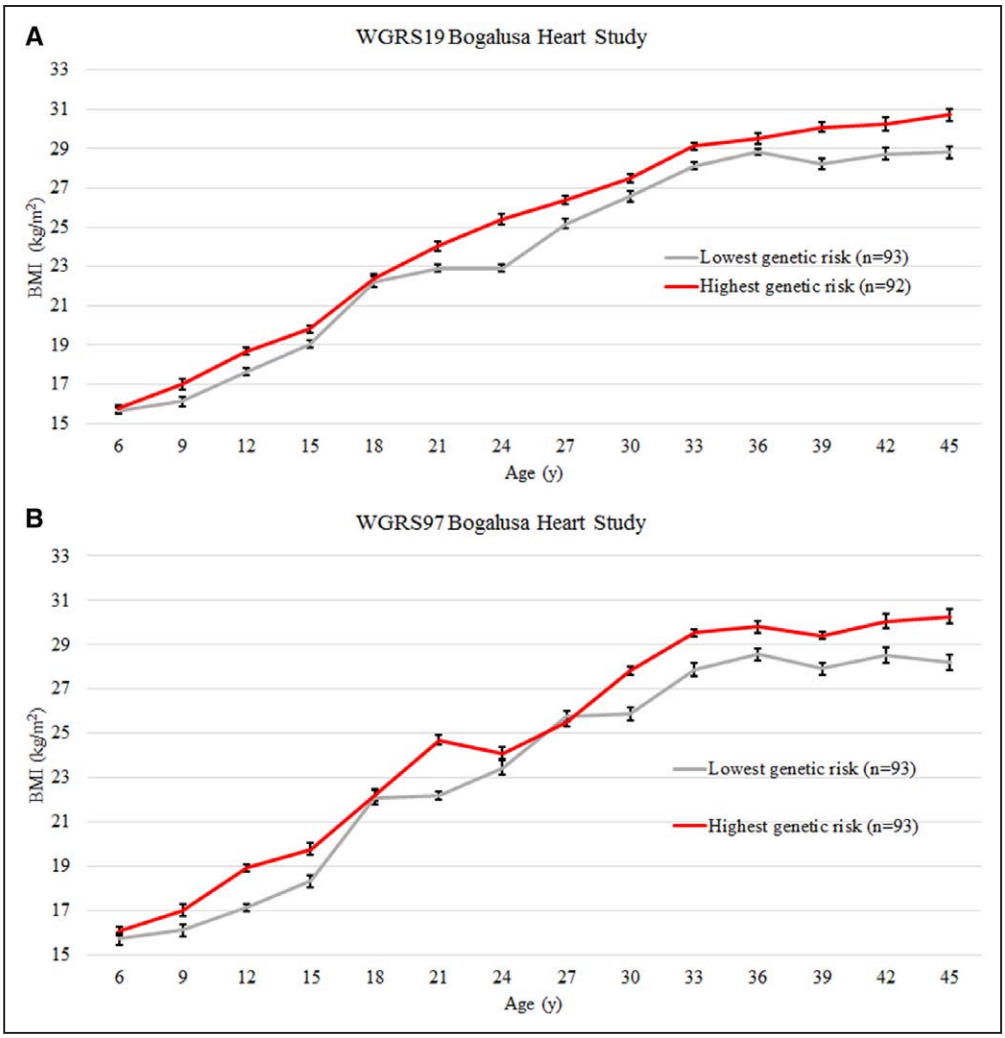

Figure 3. The body mass index (BMI) trajectories with mean value and SEM bars in participants of the BHS (Bogalusa Heart Study) with low or high genetic risk of obesity according to lowest and highest weighted genetic risk score quartile from 6 to 45 years according to (**A**) WGRS19 and (**B**) WGRS97 quartiles.

SNPs would be the ones whose influence on BMI has not yet manifested at childhood but rather manifest at older age. Our refined score WGRS19 was constructed to assess whether genetics can indeed provide enhanced predictive value over childhood clinical factors. We discovered that the genetic influence of WGRS19 on BMI was manifested between 9 and 12 years of age. Interestingly, the participants with highest genetic risk differed in their BMI from the participants with low genetic risk already at the age of 9 years. This finding was validated using BHS participants. For the WGRS97, the association with BMI was significant already at 6 years of age which is in line with the lower prediction accuracy. This provides explanation why the prediction accuracy for adulthood obesity was higher among the younger age groups, as in the older age groups the genetic effect on BMI had already manifested. Indeed, as highlighted by previous research, the influence of genetics on adult obesity risk is already manifested during childhood.[18,38–40] Therefore, genetics enhances obesity risk prediction in young children, but not in older ones. A previous study has shown that, for example, the impact of FTO on BMI manifests at the age of 7 years, which supports our findings.[41] The validation of the WGRS19 and WGRS97 in the BHS data further supports our findings, but other longitudinal studies are needed to validate the new predictive score (WGRS19).

Limitations of this study include the loss of original participants during the follow-up. However, we have previously shown that, although the nonparticipants were younger and more often male than the participants, the baseline risk factors were similar after adjusting them for age and sex.[21] Thus, the participants are considered to represent the original cohort. However, genetic data are available only for the participants and the comparison between participants and nonparticipants regarding their genetic factors could not be done. The study cohort consists of whites, and therefore the results are limited to this ethnic group only. Similarly, the validation with BHS data was performed in white participants.

In conclusion, our results demonstrate the effectiveness of our multifactorial model in predicting obesity. Our validation results confirm that WGRS19 and WGRS97 are associated with higher BMI trajectories starting from childhood. Genetic factors can improve early prediction of adulthood obesity in children, but clinical factors have a substantial

independent role as well. The effect of genetic factors on prediction accuracy is especially high among young children, whereas among older children (9–18 years), the risk of obesity can be identified using childhood clinical factors. The proposed model is anticipated to be helpful in screening children at young age with high risk of developing obesity in adulthood.

Acknowledgments

We wish to thank all the funders of this study (see Sources of Funding for details) and the participants of the Cardiovascular Risk in YFS (Young Finns Study) and the BHS (Bogalusa Heart Study).

Sources of Funding

The YFS (Young Finns Study) has been financially supported by the Academy of Finland: grants 286284 (Dr Lehtimäki), 134309 (Eye), 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi), and 41071 (Skidi); the Social Insurance Institution of Finland; Kuopio, Tampere and Turku University Hospital Medical Funds (grant X51001 for Dr Lehtimäki); Juho Vainio Foundation; Paavo Nurmi Foundation; Finnish Foundation of Cardiovascular Research; Finnish Cultural Foundation; Tampere Tuberculosis Foundation; Emil Aaltonen Foundation; and Yrjö Jahnsson Foundation. In addition, this study has been funded by University of Turku Graduate School (UTUGS), Sigrid Juselius Foundation, JDRF, Marie Sklodowska-Curie Innovative Training Network, European Research Council (ERC) Starting Grant no.677943, and the Academy of Finland Grant number 296801.

Disclosures

None.

References

1. Van Gaal LF, Mertens IL, De Block CE. Mechanisms linking obesity with cardiovascular disease. *Nature*. 2006;444:875–880. doi: 10.1038/nature05487.
2. Kahn SE, Hull RL, Utzschneider KM. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature*. 2006;444:840–846. doi: 10.1038/nature05482.
3. Kotsis V, Stabouli S, Papakatsika S, Rizos Z, Parati G. Mechanisms of obesity-induced hypertension. *Hypertens Res*. 2010;33:386–393. doi: 10.1038/hr.2010.9.
4. Renehan AG, Zwahlen M, Egger M. Adiposity and cancer risk: new mechanistic insights from epidemiology. *Nat Rev Cancer*. 2015;15:484–498. doi: 10.1038/nrc3967.
5. Singh AS, Mulder C, Twisk JW, van Mechelen W, Chinapaw MJ. Tracking of childhood overweight into adulthood: a systematic review of the literature. *Obes Rev*. 2008;9:474–488. doi: 10.1111/j.1467-789X.2008.00475.x.
6. Jääskeläinen A, Pussinen J, Nuutinen O, Schwab U, Pirkola J, Kolehmainen M, et al. Intergenerational transmission of overweight among Finnish adolescents and their parents: a 16-year follow-up study. *Int J Obes (Lond)*. 2011;35:1289–1294. doi: 10.1038/ijo.2011.150.
7. Oken E, Levitan EB, Gillman MW. Maternal smoking during pregnancy and child overweight: systematic review and meta-analysis. *Int J Obes (Lond)*. 2008;32:201–210. doi: 10.1038/sj.ijo.0803760.
8. Taveras EM, Gillman MW, Peña MM, Redline S, Rifas-Shiman SL. Chronic sleep curtailment and adiposity. *Pediatrics*. 2014;133:1013–1022. doi: 10.1542/peds.2013-3065.
9. Kleiser C, Schaffrath Rosario A, Mensink GB, Prinz-Langenohl R, Kurth BM. Potential determinants of obesity among children and adolescents in Germany: results from the cross-sectional KiGGS Study. *BMC Public Health*. 2009;9:46. doi: 10.1186/1471-2458-9-46.
10. Ritchie LD, Welk G, Styne D, Gerstein DE, Crawford PB. Family environment and pediatric overweight: what is a parent to do? *J Am Diet Assoc*. 2005;105(5 Suppl 1):S70–S79. doi: 10.1016/j.jada.2005.02.017.
11. Rokholm B, Silventoinen K, Tynelius P, Gamborg M, Sørensen TI, Rasmussen F. Increasing genetic variance of body mass index during the Swedish obesity epidemic. *PLoS One*. 2011;6:e27135. doi: 10.1371/journal.pone.0027135.
12. Wardle J, Carnell S, Haworth CM, Plomin R. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *Am J Clin Nutr*. 2008;87:398–404.
13. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al; MAGIC; Procardis Consortium. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*. 2010;42:937–948. doi: 10.1038/ng.686.
14. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; International Endogene Consortium. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015;518:197–206. doi: 10.1038/nature14177.
15. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM, et al; Wellcome Trust Case Control Consortium; Genetic Investigation of ANthropometric Traits Consortium. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet*. 2009;41:25–34. doi: 10.1038/ng.287.
16. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, et al; Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial; KORA; Nurses' Health Study; Diabetes Genetics Initiative; SardiNIA Study; Wellcome Trust Case Control Consortium; FUSION. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet*. 2008;40:768–775. doi: 10.1038/ng.140.
17. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*. 2007;316:889–894. doi: 10.1126/science.1141634.
18. Warrington NM, Howe LD, Wu YY, Timpson NJ, Tilling K, Pennell CE, et al. Association of a body mass index genetic risk score with growth throughout childhood and adolescence. *PLoS One*. 2013;8:e79547. doi: 10.1371/journal.pone.0079547.
19. González JR, Estévez MN, Giralt PS, Cáceres A, Pérez LM, González-Carpio M, et al. Genetic risk profiles for a childhood with severe overweight. *Pediatr Obes*. 2014;9:272–280. doi: 10.1111/j.2047-6310.2013.00166.x.
20. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;383:367–378.
21. Juonala M, Juhola J, Magnussen CG, Würtz P, Viikari JS, Thomson R, et al. Childhood environmental and genetic predictors of adulthood obesity: the cardiovascular risk in young Finns study. *J Clin Endocrinol Metab*. 2011;96:E1542–E1549. doi: 10.1210/jc.2011-1243.
22. Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP, et al. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*. 2007;23:2741–2746. doi: 10.1093/bioinformatics/btm443.
23. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009;5:e1000529. doi: 10.1371/journal.pgen.1000529.
24. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al; 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467:1061–1073.
25. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat*. 2000;28:407.
26. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann. Stat*. 2010;29:1189–1232.
27. Ogutu JO, Piepho HP, Schulz-Streeck T. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc*. 2011;5 Suppl 3:S11. doi: 10.1186/1753-6561-5-S3-S11.
28. Hirasawa H, Murata H, Mayama C, Araie M, Asaoka R. Evaluation of various machine learning methods to predict vision-related quality of life from visual field data and visual acuity in patients with glaucoma. *Br J Ophthalmol*. 2014;98:1230–1235. doi: 10.1136/bjophthalmol-2013-304319.
29. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol*. 2008;77:802–813. doi: 10.1111/j.1365-2656.2008.01390.x.
30. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *Systems, Man, and Cybernetics, Part C Appl Rev IEEE Trans*. 2012;42(4):463–484.
31. Smith EN, Chen W, Kähönen M, Kettunen J, Lehtimäki T, Peltonen L, et al. Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa Heart Study. *PLoS Genet*. 2010;6:e1001094. doi: 10.1371/journal.pgen.1001094.

32. Barker DJ, Winter PD, Osmond C, Margetts B, Simmonds SJ. Weight in infancy and death from ischaemic heart disease. *Lancet*. 1989;2:577–580.
33. Juonala M, Raitakari M, S A Viikari J, Raitakari OT. Obesity in youth is not an independent predictor of carotid IMT in adulthood. The Cardiovascular Risk in Young Finns Study. *Atherosclerosis*. 2006;185:388–393. doi: 10.1016/j.atherosclerosis.2005.06.016.
34. Oikonen M, Tikkanen E, Juhola J, Tuovinen T, Seppälä I, Juonala M, et al. Genetic variants and blood pressure in a population-based cohort: the Cardiovascular Risk in Young Finns study. *Hypertension*. 2011;58:1079–1085. doi: 10.1161/HYPERTENSIONAHA.111.179291.
35. Tikkanen E, Tuovinen T, Widén E, Lehtimäki T, Viikari J, Kähönen M, et al. Association of known loci with lipid levels among children and prediction of dyslipidemia in adults. *Circ Cardiovasc Genet*. 2011;4:673–680. doi: 10.1161/CIRCGENETICS.111.960369.
36. Suomela E, Oikonen M, Pitkänen N, Ahola-Olli A, Virtanen J, Parkkola R, et al. Childhood predictors of adult fatty liver. The Cardiovascular Risk in Young Finns Study. *J Hepatol*. 2016;65:784–790. doi: 10.1016/j.jhep.2016.05.020.
37. Pitkänen N, Juonala M, Rönnemaa T, Sabin MA, Hutri-Kähönen N, Kähönen M, et al. Role of conventional childhood risk factors versus genetic risk in the development of type 2 diabetes and impaired fasting glucose in adulthood: the cardiovascular risk in Young Finns Study. *Diabetes Care*. 2016;39:1393–1399. doi: 10.2337/dc16-0167.
38. Steinsbekk S, Belsky D, Guzey IC, Wardle J, Wichstrøm L. Polygenic Risk, Appetite traits, and weight gain in middle childhood: A Longitudinal Study. *JAMA Pediatr*. 2016;170:e154472. doi: 10.1001/jamapediatrics.2015.4472.
39. Belsky DW, Moffitt TE, Houts R, Bennett GG, Biddle AK, Blumenthal JA, et al. Polygenic risk, rapid childhood growth, and the development of obesity: evidence from a 4-decade longitudinal study. *Arch Pediatr Adolesc Med*. 2012;166:515–521. doi: 10.1001/archpediatrics.2012.131.
40. Elks CE, Loos RJ, Sharp SJ, Langenberg C, Ring SM, Timpson NJ, et al. Genetic markers of adult obesity risk are associated with greater early infancy weight gain and growth. *PLoS Med*. 2010;7:e1000284. doi: 10.1371/journal.pmed.1000284.
41. Hakanen M, Raitakari OT, Lehtimäki T, Peltonen N, Pahkala K, Sillanmäki L, et al. FTO genotype is associated with body mass index after the age of seven years but not with energy intake or leisure-time physical activity. *J Clin Endocrinol Metab*. 2009;94:1281–1287. doi: 10.1210/jc.2008-1199.

CLINICAL PERSPECTIVE

Obesity is a known risk factor for cardiovascular disease, and early prediction is essential for obesity prevention. The aim of this study was to assess the use of childhood clinical factors and the genetic risk factors in predicting adulthood obesity using machine learning methods. Using the data from YFS (Cardiovascular Risk in Young Finns Study n=2262) an BHS (Bogalusa Heart Study; n=369), we discovered that including genetic risk factors to the predictive model the accuracy of prediction increased statistically significantly ($P<0.05$). However, from clinical perspective the increase in accuracy was modest. The predictive accuracy was highest among young children (3–6 years), whereas among older children (9–18 years), the risk could be identified using childhood clinical factors. Our results suggest that the proposed model is helpful in screening children with high risk of developing obesity, especially at young ages where also inclusion of genetic factors showed increased prediction accuracy. Moreover, in this study, we demonstrated that the inclusion of only 19 predictive single-nucleotide polymorphisms was enough to increase the accuracy of prediction over clinical factors. Although the clinical obesity risk factors remain important factors in the prediction of obesity, the genetic factors may improve their accuracy and have potential in determining the individuals with high risk of later obesity. Our results demonstrate the genetic factors that provide additional information over clinical risk factors in identifying children who are at risk of developing obesity in adulthood, although clinical factors show a substantial independent predictive role as well.