Genetic prediction of type 2 diabetes using deep neural network

Jongoh Kim,[1,2] Junmo Kim,[3] Min Ji Kwak,[4] Mandeep Bajaj[1,2]

[1] Division of Diabetes, Metabolism, and Endocrinology, Department of Medicine, Baylor College of Medicine, Houston, TX , USA

[2] Department of Medicine, Baylor St Luke's Medical Center, Houston, TX , USA

[3] Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

[4] Department of Medicine, The University of Texas McGovern Medical School, Houston, TX, USA

Conflict of interest: None.

Correspondence: Jongoh Kim, MD, Division of Diabetes, Metabolism, Endocrinology, Department of Medicine, Baylor College of Medicine

Address: 6620 Main Street, Baylor Clinic Endocrinology, Houston, TX 7730

E-mail: dr.jongoh.kim@gmail.com

Phone: 267-625-4231 Fax: 713-798-4593

Acknowledgments

We do not have any funding to report for the study. All of the authors made critical intellectual contributions to drafting and/or revising the manuscript and approved the final version. Jongoh Kim contributed to design of the study, analysis and interpretation of the data, and drafting of the manuscript. Junmo Kim contributed to design of the study and analysis and interpretation of the data. Min Ji Kwak and Mandeep Bajaj contributed to interpretation of the data and editing of the manuscript. Jongoh Kim is the guarantor of this work and, as such, had full access to all the data

in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

The datasets used for the analyses described in this manuscript were obtained from dbGaP through accession number phs000091.v2.p1.The human subjects derive from the NHS and HPFS and these studies are supported by National Institutes of Health grants CA87969, CA55075, and DK58845. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies, GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH GEI (U01HG004424).

Abstract

Type 2 diabetes (T2DM) has strong heritability but genetic models to explain heritability have been challenging. We tested deep neural network (DNN) to predict T2DM using the nested case-control study of Nurses' Health Study (3,326 females, 45.6% T2DM) and Health Professionals Follow-up Study (2,502 males, 46.5% T2DM). We selected 96, 214, 399, and 678 SNPs through Fisher's exact test and L1-penalized logistic regression. We split each dataset randomly in 4:1 to train prediction models and test their performance. DNN and logistic regressions showed better AUC of ROC curves than the clinical model when 399 or more SNPs included. DNN was superior to logistic regressions in AUC with 399 or more SNPs in male and 678 SNPs in female. Addition of clinical factors consistently increased AUC of DNN but failed to improve logistic regressions with 214 or more SNPs. In conclusion, we show that DNN can be a versatile tool to predict T2DM incorporating large numbers of SNPs and clinical information. Limitations include a relatively small number of the subjects mostly of European ethnicity. Further studies are warranted to confirm and improve performance of genetic prediction models using DNN in different ethnic groups.

Key Words: type 2 diabetes, deep neural network, genomic microarray, prediction model

Introduction

Type 2 diabetes has become a major epidemic. In the US, the number of adults diagnosed with diabetes (90-95% being type 2 diabetes), quadrupled from 1980 to 2012 with 1.7 million new adult cases every year (1). The complications of diabetes include death, cardiovascular disease, kidney failure, blindness, and lower limb amputations. The estimated cost of diabetes care amounted to 245 billion dollars in 2012 (2).

Type 2 diabetes develops through complex interactions of genetic and environmental factors. Even though rapid increase in diabetes suggests the role of environmental factors, differential susceptibility to diabetes and heritability of insulin sensitivity and secretion support a strong genetic basis of diabetes (3). Multiple twin and familial studies showed that up to 70% of variability of having type 2 diabetes is explained by genetic effects (4-7). A strong family history of type 2 diabetes such as two or more first degree relatives or one first degree relative and two or more second degree relatives with diabetes increases risk of diabetes by 5.5 fold even after adjustment for demographic and clinical factors including obesity (8). Adolescents with family history of diabetes have defects in beta cell function and are insulin resistant as compared to those without family history (9).

However, the genetic basis of type 2 diabetes is not fully understood. In Europeans, relative risk of diabetes attributed to 40 loci explained only about 10% of observed familial clustering (10). More genetic loci were identified in subsequent genome wide association studies (GWAS) (11-13) but each genetic locus was only weakly associated with diabetes. Genetic risk scores (GRS) using combination of 10-60 variants only modestly improved risk prediction (AUC 0.552-0.680) and were not superior to clinical prediction models (AUC 0.614-920) (14). Combinations of GRS and clinical factors (AUC 0.631-0.963) were minimally better than clinical models (14, 15). It was hypothesized that rare variants with strong effects could explain the residual heritability.

However, this hypothesis was disputed in a recent study using larger scale whole genome and exome sequencing examining 27 and 3 million sequence variants respectively (13).

In this study, we used deep neural network to analyze nonlinear multilayered interactions of large numbers of genetic variants up to ~700 single nucleotide polymorphisms (SNPs) and predict type 2 diabetes. We have shown for the first time that deep neural network models incorporating large numbers of SNPs and clinical information can predict type 2 diabetes with very high accuracy.

Materials and Methods

Design

The phenotype and genotype data of the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS) were obtained from the National Health Institute (NIH) dbGaP repository (phs000091.v2.p1). The study was approved by the data access committee of NIH and Baylor College of Medicine. As previously described, a nested case-control study using NHS and HPFS was done with diabetes cases followed through 2002 and controls matched on age, month and year of blood draw, and fasting status. Age and other clinical factors were collected from the most recent questionnaire before the blood draw (NHS in 1988, HPFS in 1990). Diabetes status was self reported and then confirmed with questionnaire. Diabetes was diagnosed by the National Diabetes Data Group criteria before 1998 and the American Diabetes Association criteria in 1998-2002 (16-18).

Genotyping

Informed consent for GENEVA (Genes and Environment Initiatives in Type 2 Diabetes) was obtained from participants eligible for the nest case control study in 2007-2008. Genotyping was done using the Affymetrix Genome-Wide Human Array 6.0 (Santa Clara, CA). The details of data processing and quality control and assurance have been described elsewhere (17, 18). We obtained genotypes of the subjects available as PLINK files from the dbGaP repository (phs000091.v2.p1) that passed quality control at the genotyping center and remained after exclusion of samples with identity and quality problems at the coordination center. Genotypes with large chromosome anomalies were preset to missing.

Analysis

3,326 females from NHS and 2,502 males from HPFS (after excluding uncertain types of diabetes) were included in the analysis. Genotypes were recorded as the number of minor alleles (0 for MM,1 for Mm, 2 for mm). Candidate genetic loci were selected through following two steps considering computational limitation. Initially, 43,785 out of 909,622 genetic loci with p value <0.05 on Fisher's exact test were selected using PLINK 1.9. Afterwards, all of the analyses were performed using R 3.3.2 (www.r-project.org). Genetic variants were further narrowed down using L1 penalized logistic regression. L1 penalized logistic regression helps select one candidate SNP among a group of SNPs in linkage disequilibrium (19). Only a few values of lambda allowed computation of L1 penalized logistic regression. 96, 214, 399, and 678 SNPs (supplementary file 1) were selected using lambda 0.24, 0.22, 0.2, and 0.18 respectively. All of the selected SNPs met the quality control and assurance criteria (17, 18). The NHS and HPFS dataset was randomly split in 4:1 respectively to a training dataset and a testing dataset. Prediction models were built from the training dataset using deep neural network versus multiple logistic regressions (H2O for R, www.h2o.ai). Clinical factors included age, body mass index

(BMI), family history of diabetes in first degree relatives, hypertension, hypercholesterolemia, physical activity, smoking, and alcohol drinking. Deep neural network is a recently developed machine learning algorithm with many potentials in biomedicine (20). We chose Tanh with dropout as an activation function based on performance tests and the concept of molecular signaling where signal transduction can be positively or negatively regulated. We simplified deep neural network models to improve computational efficiency to those with 2 hidden layers of 50 nodes based on performance tests with limit of 50 epochs and stopping parameters to prevent overfitting (Figure 1). The performance was evaluated from the testing dataset by area under the curve (AUC) of receiver operating characteristic (ROC) curves, categorical net reclassification improvement (NRI), and integrated discrimination improvement (IDI). NRI measures improvement in classification between two models by sum of the percentage of increased predicted risk in cases and the percentage of decreased predicted risk in controls. We used NRI with five categories with cutoff 0.2, 0.4, 0.6, and 0.8 to avoid overestimated reclassification from continuous NRI (21). IDI measures improvement in discrimination slopes between two models by absolute difference in average predictions for cases minus average predictions for controls (14, 22, 23). P value <0.05 was considered significant.

Results

Type 2 diabetes cases comprised 46.5% of 2,502 males in HPFS and 45.6% of 3,326 females in NHS. Age was matched between cases and controls and majority of the subjects were white. Diabetic subjects had higher BMI and lower physical activity. The prevalence of hypertension, hypercholesterolemia, and smoking was higher and the prevalence of alcohol drinking was lower

in the diabetic group. Family history of diabetes in first degree relatives was more common in the diabetic group (Table 1).

We developed deep neural network and multiple logistic regression models from the training dataset using different numbers of SNPs and applied them to the testing dataset to predict type 2 diabetes in male (HPFS) and female (NHS). As in figure 2a and 2b, the distribution of predicted risk of type 2 diabetes is much more clearly separated between cases and controls in deep neural network than in logistic regressions. The larger the number of SNPs that were included, the better the separation between cases and controls was observed.

We calculated AUC of ROC curves (Figure 3, Table 2) followed by categorical reclassification (NRI), and discrimination slopes (IDI) (Table 3) to compare performance of deep neural network and logistic regression models using different numbers of SNPs without and with clinical factors. Clinical factors alone in logistic regression predicted type 2 diabetes quite well as previously reported based on AUC (0.764 in male, 0.803 in female) (14, 15). When only 96 and 214 SNPs were included, neither deep neural network nor logistic regression models did outperform the clinical model in AUC. When more SNPs were included (399 and 678 SNPs), both logistic regression and deep neural network models showed better AUC than the clinical model ($p < 0.001$ to 0.023). Deep neural network became superior to logistic regression in AUC when 399 or more SNPs were included in male and 678 SNPs were included in female (all $p<0.001$). The deep neural network using 678 SNPs achieved AUC 0.931 in male and 0.928 in female. Deep neural network showed improved reclassification (NRI) and discrimination slopes (IDI) compared to logistic regression when 96 or more SNPs were included in male and 214 or more SNPs were included in female, consistent with the distribution of predicted risk in Figure 2a and 2b.

Next, we included clinical factors – age, BMI, family history of diabetes, hypertension, hypercholesterolemia, physical activity, smoking, and alcohol drinking - in deep neural network and logistic regression models. Clinical factors significantly improved AUC of deep neural network models across the board ($p < 0.001$ to 0.036). The impact of adding clinical factors was more prominent when lower numbers of SNPs were included. With 96 SNPs, clinical factors improved AUC of deep neural network models from 0.765 to 0.831 in male ($p=0.001$) and 0.718 to 0.865 in female ($p < 0.001$), whereas with 678 SNPs, from 0.931 to 0.948 in male ($p=0.025$) and from 0.928 to 0.946 in female ($p=0.014$). Reclassification (NRI) and discrimination slopes (IDI) did not further improve with addition of clinical factors when 399 or more SNPs were included in male and 678 SNPs were included in female. Interestingly, clinical factors did not improve AUC of logistic regression when 214 or more SNPs were included. Clinical factors even decreased AUC of logistic regression models in female with 399 and 678 SNPs. Accordingly, deep neural network showed better AUC, reclassification (NRI) and discrimination (IDI) than logistic regression models when clinical factors were added to 214 or more SNPs.

Discussion

In the present study, we have shown that deep neural network predicts type 2 diabetes robustly incorporating large numbers of genotypes and clinical factors as compared to logistic regression models. The number of included genotypes had a direct impact on the performance of genetic prediction models. Clinical factors consistently improved performance of deep neural network models, reflecting variability that is not captured even by large numbers of SNPs.

Our genetic prediction models clearly contrast with the previous studies where combination of SNPs captured only 10% of variability attributed to genetic factors and did not significantly improve clinical models. As compared to previously published studies, our study incorporated much larger number of SNPs. In addition, we used cutting edge machine learning tools available that enabled construction of larger non-linear models with up to ~700 variables.

Type 2 diabetes has been so called geneticists' nightmare (24). Since the introduction of GWAS, many previous studies have tried to find the genetic factors leading to type 2 diabetes. However, very strict criteria have been used to identify associated genetic loci in GWAS to ensure discovery of biologically meaningful genetic variants. This has also been driven by the model in which rare low frequency variants with strong effect explain majority of heritability. To increase power to detect rare variants, more participants in more diverse ethnicities have been included and analyzed on deeper levels of genome. As of now, approximately 90 genetic loci are identified but each of these loci is weakly associated with type 2 diabetes (12, 13). In a recently published study with large scale genome and exome sequencing, the number of rare variants discovered did not match the simulated result based on the rare low frequency variant model (13). Therefore, common variants should explain majority of heritability but no study has captured combined contribution of common variants to type 2 diabetes so far.

Our data of predicting type 2 diabetes using approximately 700 SNPs from the conventional microarray data of only 2,502 males (HPFS) and 3,326 females (NHS) supports the common polygenic model of type 2 diabetes. Importantly, the number of SNPs included in the analysis had a significant impact on performance. The models with 96 SNPs showed poor predictive performance similar to the previous reports (14, 15). However, the model performance dramatically improved with 399 and 678 SNPs, much larger than ~90 variants identified in

GWAS. This implies that much of genetic information related to type 2 diabetes was lost through strict selection of SNPs in the previous GWAS. Type 2 diabetes is a complex disorder involving multiple interacting signaling pathways. The effect of single protein can depend on presence or absence of other proteins. Therefore, the effect of some common variants can vary in each individual depending on presence or absence of other variants. These common variants cannot meet the strict selection criteria applied in the previous GWAS. We used a two step strategy to select candidate SNPs, first by Fisher's exact test followed by L1 penalized logistic regression. The whole genome data requires enormous memory and processing power to handle. Therefore, initial selection method should be computationally very simple. Our strategy was developed in consideration of limited computational resources. There is no agreement in how to select genetic variables to enter into prediction models (25). Further research is required to guide feature selection of genomic data.

We used deep neural network to construct diabetes prediction models for the first time. Deep neural network has already been used in everyday lives from image processing to voice recognition. Deep neural network of genetic information is at a very early stage (20). In our study, deep neural network models were more robust and expandable using clinical variables than multiple logistic regressions. Deep neural network can capture nonlinear multi-level interactions among many factors. Another remarkable strength of deep neural network is that the model can evolve when more cases and controls are provided. Our analysis was limited to mostly European ethnicity available from the two cohort studies. Expansion of deep neural network models to many other ethnicities will further improve performance.

Pathogenesis of type 2 diabetes is mediated by gene-environment interaction (26). For example, life style modification reduces risk of diabetes associated with TCF7L2 polymorphisms, the most

well known gene for type 2 diabetes (27). Our data showed that the clinical factors improve performance of genetic prediction models. The clinical factors included hypertension and hypercholesterolemia, body mass index, and behavioral factors such as smoking, alcohol drinking, and physical activity that can reflect the effect of environment. Interestingly, logistic regression models did not incorporate the clinical factors as well as deep neural networks, further supporting versatility of deep neural network in genetic prediction models.

Our study has several limitations. As mentioned earlier, we had a relatively small number of participants and that they were mostly of European ethnicity. Our result has not been validated outside the study population. The methodology for feature selection is not established and we developed an economic and efficient selection strategy. We did not look into molecular pathways and mechanisms due to very large numbers of SNPs in the prediction models.

In conclusion, we showed that deep neural network is a very promising machine learning tool to analyze genomic data and enables robust accurate models to predict type 2 diabetes.

References

1. Centers for Disease Control and Prevention. Diabetes Report Card 2014. Atlanta, GA: Centers for Disease Control and Prevention, US Dept of Health and Human Services, 2015.

2. Centers for Disease Control and Prevention. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. Atlanta, GA: Department of Health and Human Services, 2014.

3. Stumvoll M, Goldstein BJ, van Haeften TW. Type 2 diabetes: principles of pathogenesis and therapy. Lancet 2005: 365 (9467): 1333-1346.

4. Lehtovirta M, Pietiläinen K, Levälahti E et al. Evidence that BMI and type 2 diabetes share only a minor fraction of genetic variance: a follow-up study of 23,585 monozygotic and dizygotic twins from the Finnish Twin Cohort Study. Diabetologia 2010: 53 (7): 1314-1321.

5. Almgren P, Lehtovirta M, Isomaa B et al. Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia 2011: 54 (11): 2811-2819.

6. Carlsson S, Ahlbom A, Lichtenstein P et al. Shared genetic influence of BMI, physical activity and type 2 diabetes: a twin study. Diabetologia 2013: 56 (5): 1031-1035.

7. Willemsen G, Ward KJ, Bell CG et al. The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. Twin Res Hum Genet 2015: 18 (6): 762-771.

8. Valdez R, Yoon PW, Liu T et al. Family History and Prevalence of Diabetes in the U.S. Population. The 6-year results from the National Health and Nutrition Examination Survey (1999-2004). Diabetes Care 2007: 30 (10): 2517-2522.

9. Arslanian SA, Bacha F, Saad R et al. Family History of Type 2 Diabetes Is Associated With Decreased Insulin Sensitivity and an Impaired Balance Between Insulin Sensitivity and Insulin Secretion in White Youth. Diabetes Care 2005: 28 (1): 115-119.

10. Voight BF, Scott LJ, Steinthorsdottir V et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 2010: 42 (7): 579-589.

11. Morris AP, Voight BF, Teslovich TM et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 2012: 44 (9): 981-990.

12. DIAbetes Genetics Replication And Meta-analysis Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, South Asian Type 2 Diabetes Consortium et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 2014: 46 (3): 234-244.

13. Fuchsberger C, Flannick J, Teslovich TM et al. The genetic architecture of type 2 diabetes. Nature 2016: 536 (7614): 41-47.

14. Wang X, Strizich G, Hu Y et al. Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction. J Diabetes 2016: 8 (1): 24-35.

15. Keating BJ. Advances in Risk Prediction of Type 2 Diabetes: Integrating Genetic Scores With Framingham Risk Models. Diabetes 2015: 64 (5): 1495-1497.

16. Cornelis MC, Qi L, Zhang C et al. Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. Ann Intern Med 2009: 150 (8): 541-550.

17. Qi L, Cornelis MC, Kraft P et al. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet 2010: 19 (13): 2706-2715.

18. Laurie CC, Doheny KF, Mirel DB et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 2010: 34 (6): 591-602.

19. Wei Z, Wang W, Bradfield J et al. Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease. Am J Hum Genet 2013: 92 (6): 1008-1012.

20. Mamoshina P, Vieira A, Putin E et al. Applications of Deep Learning in Biomedicine. Mol Pharm 2016: 13 (5): 1445-1454.

21. Cook NR. Clinically Relevant Measures of Fit? A Note of Caution. Am J Epidemiol 2012: 176(6): 488–491.

22. Steyerberg EW, Vickers AJ, Cook NR et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology 2010: 21 (1): 128-138.

23. Pickering JW, Endre ZH. New Metrics for Assessing Diagnostic Potential of Candidate Biomarkers. Clin J Am Soc Nephrol 2012: 7 (8): 1355-1364.

24. Rich SS. Diabetes: Still a geneticist's nightmare. Nature 2016: 536 (7614): 37-38.

25. Ang JC, Mirzal A, Haron H et al. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. IEEE/ACM Trans Comput Biol Bioinform 2016: 13 (5): 971-989.

26. Franks PW, Pearson E, Florez JC. Gene-Environment and Gene-Treatment Interactions in Type 2 Diabetes, Progress, pitfalls, and prospects. Diabetes Care. 2013: 36(5): 1413–1421.

27. Florez JC, Jablonski KA, Bayley N et al. Diabetes Prevention Program Research Group. TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program. N Engl J Med. 2006: 355(3): 241-250.

Table 1. Characteristics of the participants

	Male - HPFS			Female –NHS		
	No diabetes	Diabetes	P-value	No diabetes	Diabetes	P-value
N	1338 (53.5%)	1164 (46.5%)		1810 (54.4%)	1516 (45.6%)	
Age	59.1	59.3	0.651	55.6	56.0	0.072
White race	95.9%	96.1%	0.755	98.3%	98.0%	0.569
Height (cm)	178.6	178.6	0.971	163.9	163.8	0.605

Weight (Kg)	80.5	89.0	<0.001	68.3	80.2	<0.001
BMI	25.2	27.8	<0.001	25.4	29.9	<0.001
Activity (MET, hours/week)	40.8	29.7	<0.001	15.5	12.9	<0.001
Family History of diabetes -1st degree	21.4%	42.5%	<0.001	22.1%	49.5%	<0.001
Hypertension	21.8%	41.2%	<0.001	19.9%	48.5%	<0.001
High Cholesterol	29.1%	40.5%	<0.001	10.5%	23.4%	<0.001
Current Smoking	5.8%	9.0%	0.002	10.4%	14.1%	0.001
Alcohol (>2 for men, >1 for women)	72.8%	65.5%	<0.001	59.3%	42.5%	<0.001

Abbreviation: BMI, body mass index; MET, metabolic equivalent

Table 2. Prediction of type 2 diabetes by deep neural network versus logistic regression using different numbers of SNPs without and with clinical factors.

Variables	Model	Male				Female			
		AUC	P vs. clinical	P vs. logistic	P vs. without clinical	AUC	P vs. clinical	P vs. logistic	P vs. without clinical
Clinical†	Logistic	0.764				0.803			

factors									
96 SNPs	Logistic	0.781	0.568			0.762	0.099		
	DNN	0.765	0.960	0.053		0.718	0.001	<0.001	
96 SNPs +clinical	Logistic	0.856	<0.001		<0.001	0.864	<0.001		<0.001
	DNN	0.831	<0.001	0.002	0.001	0.865	<0.001	0.962	<0.001
214 SNPs	Logistic	0.801	0.209			0.828	0.283		
	DNN	0.790	0.384	0.339		0.820	0.480	0.350	
214 SNPs +clinical	Logistic	0.832	<0.001		0.126	0.821	0.001		0.746
	DNN	0.867	<0.001	0.006	<0.001	0.891	<0.001	<0.001	<0.001
399 SNPs	Logistic	0.830	0.023			0.867	0.003		
	DNN	0.877	<0.001	<0.001		0.863	0.006	0.591	
399 SNPs +clinical	Logistic	0.850	<0.001		0.289	0.828	<0.001		0.038
	DNN	0.900	<0.001	<0.001	0.036	0.903	<0.001	<0.001	<0.001
678 SNPs	Logistic	0.857	0.001			0.902	<0.001		
	DNN	0.931	<0.001	<0.001		0.928	<0.001	<0.001	
678 SNPs +clinical	Logistic	0.847	<0.001		0.577	0.833	<0.001		<0.001
	DNN	0.948	<0.001	<0.001	0.025	0.946	<0.001	<0.001	0.014

Abbreviations: DNN, deep neural network; AUC, area under curve of receiver operating characteristic curves

†Clinical factors include age, body mass index, family history of diabetes in first degree relatives, hypertension, hypercholesterolemia, physical activity, smoking, and alcohol drinking

Table 3. Performance measured by net reclassification improvement (NRI) and integrated discrimination index (IDI) of deep neural network (DNN) versus logistic regression using different numbers of SNPs without and with clinical factors.

Models	Male		Female	
	NRI	IDI	NRI	IDI
96 SNPs				
L vs. Cl	7.6[-7.9,23.0]	3.3[-2.2,8.8]	-12.7[-26.0,0.6]	-4.9[-9.6,-0.2]*
DNN vs. Cl	37.2[[21.6,52.8]***	8.4[1.9,14.8]*	-30.2[-43.1,-17.3]***	-8.4[-12.9,-3.8]***
DNN vs. L	32.8[20.4,45.1]***	5.1[2.7,7.5]***	-33.7[-44.3,-23.1]***	-3.5[-5.4,-1.6]***
L+Cl vs. L	51.2[38.7,63.8]***	14.7[11.2,18.1]***	56.1[45.1,67.0]***	18.6[15.6,21.7]***
DNN+Cl vs. DNN	30.4[18.5,42.3]***	14.0[8.5,19.5]***	68.3[58.4,78.2]***	13.7[10.4,17.1]***
DNN+Cl vs. L+Cl	17.6[6.3,28.8]**	4.5[1.8,7.1]***	-18.2[-27.8,-8.7]***	-8.4[-10.4,-6.4]***
214 SNPs				
L vs. Cl	2.1[-13.2,17.4]	0.1[-4.8,5.0]	5.1[-8.1,18.4]	1.9[-2.8,6.6]
DNN vs. Cl	43.1[28.1,58.0]***	15.1[8.3,21.9]***	38.9[26.1,51.8]***	14.2[8.4,20.0]***
DNN vs. L	46.1[32.3,60.0]***	15.0[11.3,18.7]***	41.5[30.8,52.2]***	12.3[9.7,15.0]***
L+Cl vs. L	7.4[-6.8,21.6]	3.1[0.0-6.2]	-18.8[-31.5,-6.2]**	-5.4[-9.2,-1.6]**
DNN+Cl vs. DNN	24.1[13.0,35.2]***	14.7[9.3,20.1]***	26.4[16.2,36.5]***	14.4[9.9,19.0]***
DNN+Cl vs. L+Cl	76.9[63.4,90.4]***	26.6[22.5,30.8]***	85.5[74.5,96.6]***	32.2[28.6,35.8]***
399 SNPs				
L vs. Cl	7.1[-8.1,22.2]	4.4[-0.7,9.4]	28.1[15.1,41.1]***	11.9[7.0,16.8]***
DNN vs. Cl	89.1[75.2,103.1]***	39.5[32.2,46.9]***	60.5[48.0,72.9]***	26.2[20.1,32.3]***
DNN vs. L	86.3[73.2,99.4]***	35.2[30.4,39.9]***	35.5[25.4,45.6]***	14.3[11.2,17.5]***
L+Cl vs. L	4.6[-9.0,18.1]	0.4[-2.8,3.5]	-46.1[-58.4,-33.7]***	-14.7[-18.8,-10.7]***
DNN+Cl vs. DNN	-2.1[-11.2,6.9]	-1.4[-6.5,3.8]	15.8[7.8,23.9]***	10.1[5.9,14.2]***
DNN+Cl vs. L+Cl	94.3[81.5,107.1]***	33.5[29.0,37.9]***	100.6[89.5,111.8]***	39.1[34.8,43.4]***
678 SNPs				
L vs. Cl	19.9[5.1,34.8]**	8.5[3.5-13.5]***	47.2[34.9,59.6]***	20.4[15.4,25.3]***
DNN vs. Cl	100.9[88.0,113.9]***	46.4[39.8,52.9]***	91.7[80.7,102.6]***	44.3[38.5,50.1]***
DNN vs. L	90.2[77.7,102.6]***	37.9[33.6,42.1]***	50.9[42.1,59.7]***	23.9[20.9,26.9]***
L+Cl vs. L	-12.7[-25.7,0.4]	-3.2[-6.3,0.0]*	-66.0[-77.8,-54.1]***	-22.5[-26.6,-18.4]***
DNN+Cl vs. DNN	5.6[-2.4, 13.6]	4.5[-0.2,9.2]	0.6[-6.7,7.9]	1.1[-3.0,5.1]
DNN+Cl vs. L+Cl	112.4[100.3,124.4]***	45.5[40.9,50.1]***	111.9[101.5,122.4]***	47.5[43.3,51.6]***

Abbreviations: Cl, clinical factors; DNN, deep neural network using indicated SNPs; DNN+Cl, deep neural network using indicated SNPs plus clinical factors; L, logistic regressions using indicated SNPs; L+Cl, logistic regression using indicated SNPs plus clinical factors.

*p<0.05; **p<0.01;***p<0.001

Figure Legends

Figure 1. The architecture of the deep neural network analysis

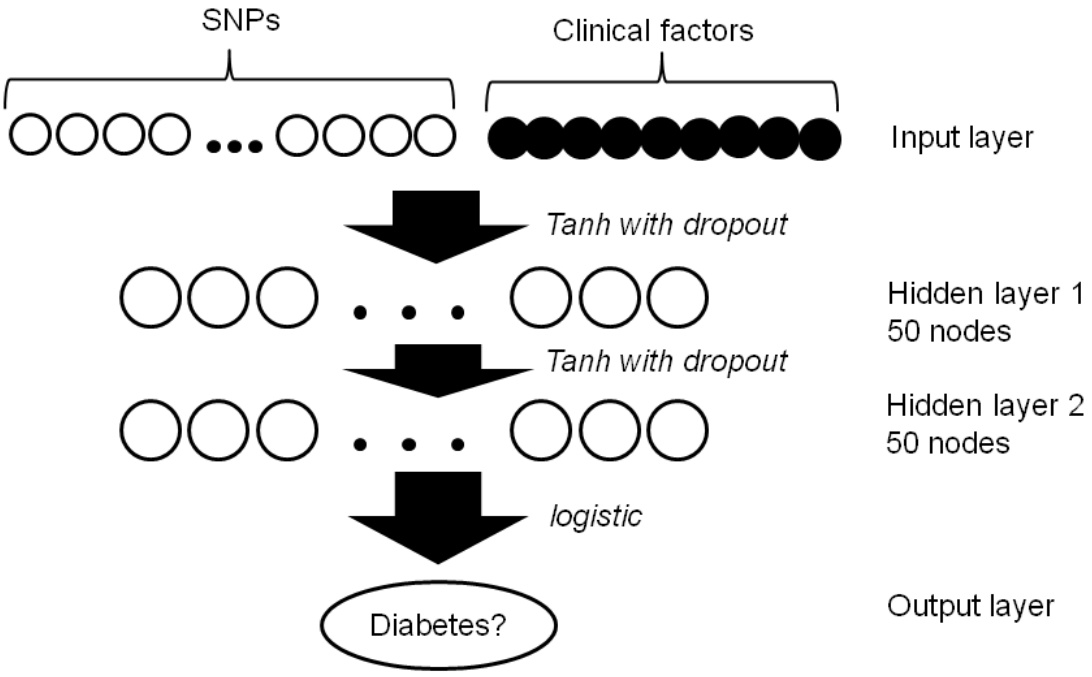

Figure 2a. Distribution of predicted risk by status of type 2 diabetes in males (from HPFS) using (a)~(d) deep neural network and (e)~(h) logistic regression. (a) and (e) 96 SNPs, (b) and (f) 214 SNPs, (c) and (g) 399 SNPs, (d) and (h) 678 SNPs included in the models.

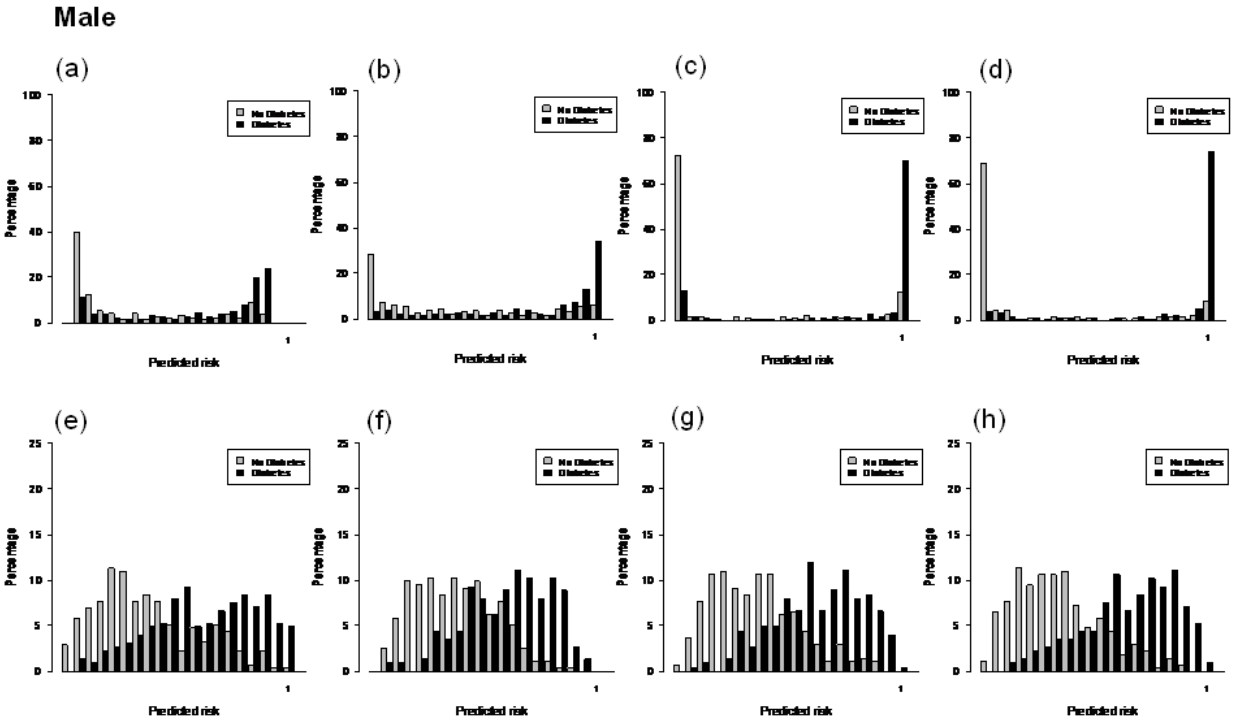

Figure 2b. Distribution of predicted risk by status of type 2 diabetes in Females (from NHS) using (a)~(d) deep neural network and (e)~(h) logistic regression. (a) and (e) 96 SNPs, (b) and (f) 214 SNPs, (c) and (g) 399 SNPs, (d) and (h) 678 SNPs included in the models.

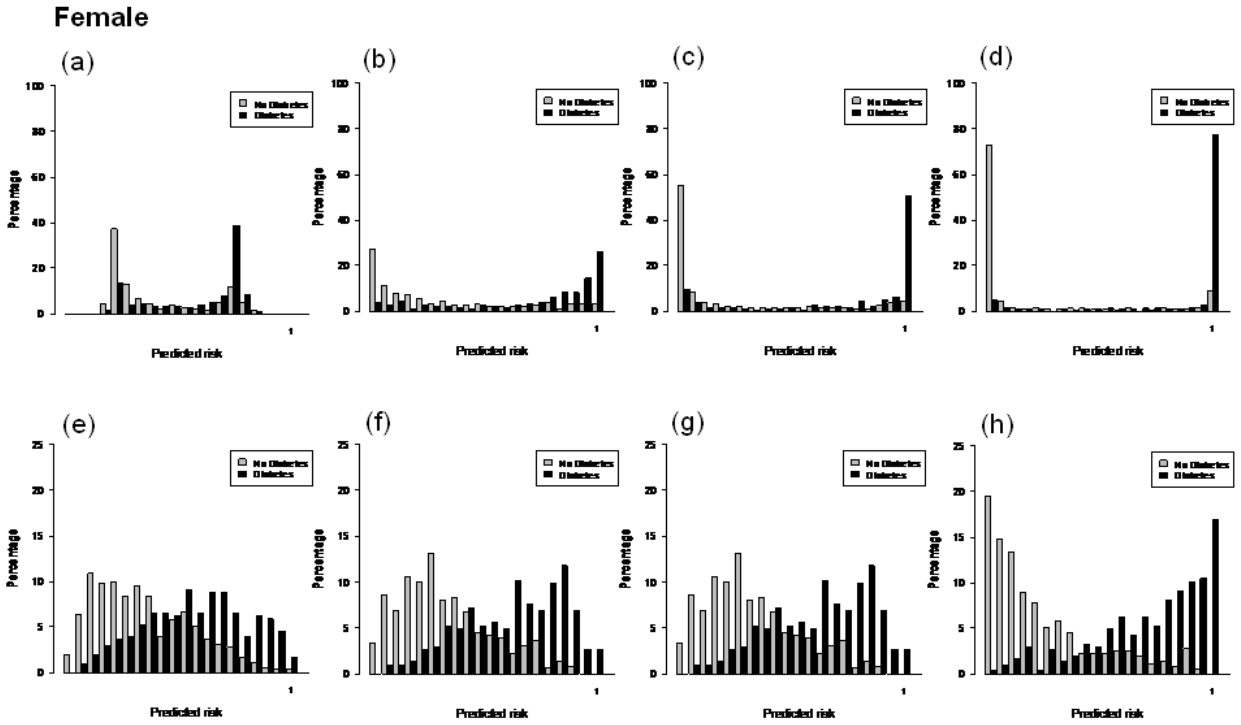

Figure 3. Receiver operating characteristic (ROC) curves of various models predicting type 2 diabetes compared with the clinical model using logistic regression (black line). (a)-(c) for male, (d)-(f) for female; (a) and (d) deep neural network (DNN) using 96, 214, 399, 678 SNPs. (b) and (e) logistic regression using 96, 214, 399, 678 SNPs. (c) and (f) DNN using 399 and 678 SNPs without and with clinical factors