Disease as collider: a new case-only method to discover environmental factors in complex diseases with genetic risk estimation

Félix Balazard[1,2]*, Sophie Le Fur[2,3], Pierre Bougnères[2,3], Alain-Jacques Valleron[2], the Isis-Diab collaborative group

*: Corresponding author. Email: felix.balazard@inserm.fr

1: Sorbonne Universités, UPMC Univ Paris 06, CNRS, Paris, France.

2: INSERM U1169, Hôpital Bicêtre, Université Paris-Sud, Kremlin-Bicêtre, France.

3: Department of pediatric endocrinology, Hôpital Bicêtre, Kremlin-Bicêtre, France

Abstract

Background: Genetic risk scores can quantify part of the predisposition of an individual to a disease. The identification of environmental factors is more challenging. Collider bias appears between two causes (e.g. gene and environment) when conditioning on a shared consequence (the collider, disease).

Methods: We introduce Disease As Collider (DAC), a new case-only methodology to validate environmental factors using genetic risk. A complex disease is a collider between genetic and environmental factors. Under reasonable assumptions, a negative correlation between genetic risk and environment in cases provides a signature of a genuine environmental risk factor. Simulation of disease occurrence in a source population allows to estimate the statistical power of DAC as a function of prevalence of the disease, predictive accuracy of genetic risk and sample size. We illustrate DAC in 831 type 1 diabetes (T1D) patients.

Results: The power of DAC increases with sample size, prevalence and accuracy of genetic risk estimation. For a prevalence of 1% and a realistic genetic risk estimation, power of 80% is reached for a sample size under 3000. Power was low in our case study as the prevalence of T1D in children is low (0.2%).

Conclusions: DAC could provide a new line of evidence for discovering which environmental factors play a role in complex diseases, or validating results obtained in case-control studies. We discuss the circumstances needed for DAC to participate in the triangulation of environmental causes of disease. We highlight the link with the case-only design for gene environment interaction.

Key-words: collider bias, graphical models, genetic risk estimation, environmental factors.

Key messages :

- Disease is a collider between genetic risk and environmental factors.
- This can be used to discover or validate the association between a disease and an environmental factor in a case-only setting.
- Statistical power of this approach depends strongly on the prevalence of the disease as well as on sample size and genetic risk prediction accuracy.

Introduction

The dissection of environmental determinants of complex diseases is difficult. For diseases with a prevalence under 1% such as T1D, celiac disease and inflammatory bowel disease, prospective cohort studies imply to follow tens of thousand of participants for many years. Screening a population for genetic predisposition is a way of making prospective studies more tractable such as in the ongoing TEDDY[1] study for T1D. In comparison, the case-control design allows to obtain a large population of cases at a reduced cost. It is however sensitive to the choice of controls. All in all, progress in identifying environmental determinants of these diseases is slow, vexing and expensive[2–4].

On the other hand, the genetics of diseases have become better understood in the past decade. Genome-wide association studies (GWAS) have resulted in over a thousand validated associations between disease and loci[5]. Another use of GWAS datasets has been to estimate at an individual level the genetic risk of developing the disease using statistical learning techniques[8–13]. This provides a

one dimensional summary of genetic data relevant to epidemiology. Genetic risk scores have been obtained for T1D, Crohn's disease and celiac disease. For those three diseases, the prediction accuracies of the scores were comparable, achieving area under the receiver operating curve (AUC) around 0.85[8,9,11].

GWAS has allowed epidemiology to flourish at the crossroad between gene and environment. Such efforts have focused mainly on identifying GxE interactions[6] and on Mendelian randomization[7]. In this article, we examine the interplay between genetic risk estimation and environmental risk factors and we do so by considering the concept of collider bias.

Collider bias is the negative correlation that appears between two causes when conditioning on their shared consequence (the collider)[14]. It can mislead epidemiological investigation[15,16]. A classic example is Berkson's bias in which two diseases are negatively associated in a hospitalized population even though they are independent in the general population[17,18]. In this example, the collider is hospitalization, the shared consequence of both diseases. By looking only at patients in the hospital, i.e. by conditioning on hospitalization, a negative correlation appears between the two diseases. Collider bias has been suggested as an explanation for the birth weight paradox, the observation that neonatal mortality is higher in low birth weight infants whose mother did not smoke[19].

In this paper, we propose a change of viewpoint. Instead of considering collider bias as a nuisance, we try to harness it in service of epidemiology. Indeed, disease is a collider between genetic risk and environmental factors. By conditioning on disease, i.e. by considering only cases, a negative association found between the genetic risk and an environmental candidate will signal a possible causal link between the environmental factor and the disease. This idea is summarized in Fig 1. We refer to this methodology as Disease As Collider (DAC). To sum up, DAC allows to detect or validate a putative environmental factor by looking for an association of this factor with genetic risk in case-only data.

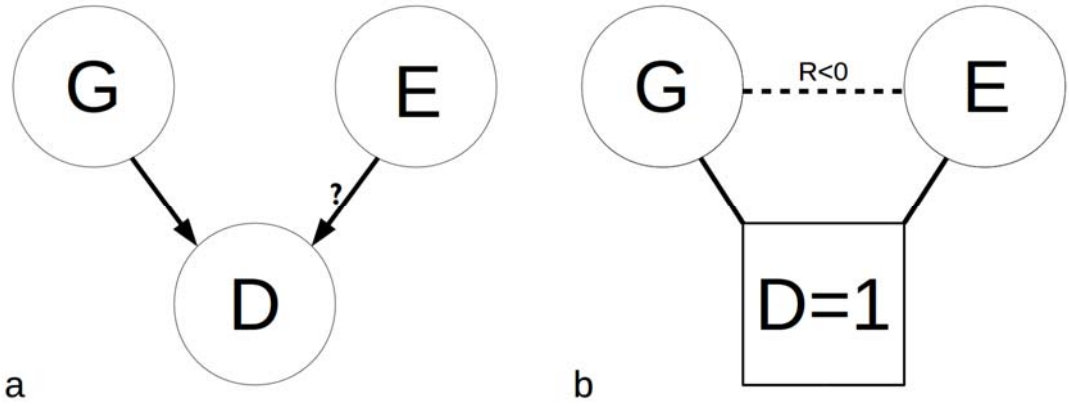

Figure 1: DAC methodology. a: Disease is a consequence of both genetic and environmental causes. Those are often independent in the general population. The environmental factor's association with the disease requires confirmation. b: When conditioning on the disease and if there is a genuine association between the environmental factor and the disease, a negative association appears between genetic risk and the environmental factor. This confirms that the environmental factor is associated with the disease.

In the next section, we describe the method and its assumptions. We present a simulation framework of disease occurrence as a function of the individual genetic susceptibility and environmental risk that allows to estimate the power of DAC. Then, we illustrate our methodology on a subset of genotyped patients from the Isis-Diab case-control study of T1D focused on the search of environmental factors[20]. Finally, we estimate power for our example and in generic scenarii by relying on our simulation framework and the genetic risk distribution in our illustration data. We evaluate the influence on power of prevalence, prediction accuracy of the genetic risk estimation and sample size.

Methods

Model assumptions

Genetic risk and environmental factor independence

In order to attribute to collider bias the responsibility for an association between genetic risk and an environmental factor in cases, we need to assume that genetic risk and the environmental factor are independent in the general population.

Multiplicative combination of odd ratios

We need to make an assumption on how genetic risk and environmental risk combine in order to evaluate the power of DAC. In order to express our assumption, we need a few definitions.

The logistic model transfers probabilities in [0,1] to log odd-ratios in the real line thanks to the logit function $logit(x)=log(x/(1-x))$. We refer to the target set of the logit function as the logit scale.

For an individual with genome G (e.g. genotyping) and environment E (e.g. responses to an environmental questionnaire), we can define its genetic risk of developing a disease D by $R_g(G)=logit(P(D=1|G))$ and its environmental risk $R_e(E)=logit(P(D=1|E))$. R_g and R_e are actually the logit of an absolute risk. As most of our calculations are made on the logit scale, we will write risk instead of risk on the logit scale throughout.

The simplest way to define the total risk $R(G,E)=logit(P(disease|G,E))$ is to assume that environmental and genetic odd ratios combine multiplicatively, i.e. environmental and genetic risk combine linearly on the logit scale. We therefore have that:

$R(G,E)=logit(P(disease|G,E))=R_g(G)+R_e(E)$.

This is an assumption of absence of interaction between the genetic risk and the environmental risk.

Description of DAC

Under the assumptions described above, there is an association between genetic risk and environmental factors in cases due to collider bias. Our method consists simply in estimating the genetic risk in cases and then on testing for association between genetic risk and environmental factors using standard tests (such as a linear regression t-test) while controlling for potential confounders.

The association that appears because of collider bias is a negative association. Therefore, DAC predicts that the cases the most at risk genetically are the least at risk because of environment. Therefore, when a putative direction of association has been established, one can perform one-sided tests. This is the case, when DAC is applied to validate findings from a case-control association.

Framework for estimating power of DAC

In order to estimate the power of our approach, we simulate disease occurrence in a source population according to our model assumptions. For a sample size of N patients, the source population consists of N/K individuals, K being the prevalence of the disease.

To allocate disease status in this synthetic population, we need to define a genetic risk distribution and an environmental risk distribution. We describe the choice of both distributions used for our illustration below. Once both distributions are defined, we then attribute to each individual in the population its genetic risk and its environmental risk by drawing independently from those distributions. This uses the assumption of independence of genetic risk and environmental factor in the population.

Once both a genetic risk and an environmental risk are defined for each individual in our source population, we define the total risk as the sum of the two risks in accordance with our assumption of multiplicative combination of odd-ratios.
To decide if an individual with genes *G* and environment *E* has the disease, we draw a uniform variable *U* on [0,1] and we could then define the disease variable *D*:

$$\begin{cases} D=1 \text{ if } U \leq P(disease|G,E) \\ D=0 \text{ if } U > P(disease|G,E) \end{cases}$$

This approach would yield a different number of cases in each simulation. To have a fixed number of patients, we compute *R(G,E)-logit(U)* and define the top *N* individuals for that sum as the patients (*D=1*). The distribution of *logit(U)* where *U* is uniformly distributed over [0,1] is called the Laplace distribution.
We then perform a regression of the environmental factor on the genetic risk in the patients and obtain a one-sided p-value. We repeat the procedure the desired number of times (100 000 times in our illustration). Our estimator of power is then the proportion of p-values under the threshold 0.05.

Application of DAC to the Isis-Diab study
The Isis-Diab cohort is a multi-centric study of T1D patients in France which recruitment started in 2007. Criteria for entering the study were insulin-dependent diabetes mellitus with positivity for anti-GAD, or anti-insulin, or anti-IA2 antibodies. As the genetic risk estimator described below was trained on patients and controls of European descent, we excluded other ethnicities in the following analyses.

Genome wide genotyping and imputation
Among the 1491 Isis-Diab patients for whom genotype data were available, 817 patients were genotyped with Illumina Human610-Quadv1_B (610 000 SNPs) microarrays, and 673 patients with Illumina Human Omni 5 Exome microarrays (4 500 000 SNPs). Genome-wide genotyping was performed on bar-coded LIMS (Laboratory Information Management System) tracked samples using two different Illumina microarrays (Human610-Quadv1_B and HumanOMNI5-4v1_B). BeadChips were processed within an automated BeadLab at the Centre National de Genotypage as per the manufacturer's instructions. Samples were subject to strict quality control criteria including assessment of concentration, fragmentation and response to PCR. A total of 20 µl of DNA aliquoted to a concentration of 50 ng/µl was used for each array. In the discovery phase, genome-wide genotypes were used for controlling the quality of the samples. First individuals with call rates <95% or duplicates and individuals who were possibly non-European were removed. By using this filtered sample set, we calculated quality control statistics, and SNPs with call rates <98% or SNPs with a Hardy-Weinberg equilibrium test p-value $<10^{-6}$ or SNPs with a minor allele frequency <1% were excluded.
Finally, 517 864 SNPs (Human610-Quadv1_B) and 3 309 261 SNPs (HumanOMNI5-4v1_B) were used for imputation analysis. Imputation was done using IMPUTE v2[22], following the instructions provided by the author. Full sequence data from the phase I 1000 Genomes Project was used[23]. Only SNPs with an info metric over 0.8 for both chips were kept for analysis.

Environmental data
The long questionnaire used in the Isis-Diab case-control study has been described before[20]. When patients did not respond to the long questionnaire used in the case-control study, a shorter questionnaire of 49 questions was sent. This questionnaire was designed while the study was underway and partial results were used to choose the included questions. As a result, 7 of the 22 variables deemed significant in the analysis of the case-control study are among the 49 questions of

the shorter questionnaire. We applied DAC to those 7 variables. The 7 variables are the answers to the following questions concerning the period before diagnosis:
-"As a baby, did the patient like baby food jar containing sweet foods more than the ones without sweet foods?",
-"Had the patient gone to the dentist?",
-"How many times a day did the patient brush his teeth?",
-"Did the patient experience severe diarrhea accompanied by vomiting?",
-"Did the patient eat hazelnut cocoa spread?",
-"Had the patient been to winter sports?" and
-"Did the patient attend a club with other children (sports, music,…)?".
The associations of those variables with T1D in the case-control data were all negative.
A total of 2959 patients filled a questionnaire: 1713 patients filled the long questionnaire and 1246 the short questionnaire. Finally, 831 patients of European descent had both genetic data and environmental data from a questionnaire. This subset constitutes the dataset on which we apply DAC.

Genetic risk estimation
To define a genetic risk score for each patient, we used a genetic risk estimator as close as possible to the best one in Wei et al.[8]. It has obtained an AUC of 0.84 on a Canadian and an American dataset[8]. Our estimator was trained on the same data: the Wellcome Trust Case Control Consortium 1 (WTCCC1)[24] T1D study (1963 cases and 2938 controls). Cases from the WTCCC1 studies on the non-autoimmune diseases type 2 diabetes, hypertension, coronary artery disease and bipolar disorder totaling 7670 individuals were used as validation controls. We refer to this group as cases of non-autoimmune diseases (CNAD).
The same quality control was used as in the original paper to filter SNPs: missing rate<5%, Hardy-Weinberg Equilibrium p-value $>10^{-3}$ and minor allele frequency>5% in the training set. Additionally, SNPs had to have missing rate<5% in the Isis-Diab study. SNPs were then selected if their training set association p-value was under 10^{-5} which was the tied best-performing threshold in the original paper. This resulted in a set of 505 SNPs.
The remaining missing data in the training set, the CNAD controls and the Isis-Diab patients were imputed by sampling randomly the training set.
The machine learning method achieving the best performance in the original paper is Support Vector Machine (SVM). It is part of a family of methods called kernel methods[25].
SVM with the default radial kernel was trained on the training set. The estimator was then used to predict on the validation set: Isis-Diab patients and the CNAD. AUC on the validation set was computed. The e1071 package implementation of SVM was used.
A machine learning prediction maximizes the separation between the two classes as measured by AUC. However, AUC depends only on the ranking of the predictions and not on its numerical value. For our purpose, in particular for the estimation of power, we need the genetic risk to encode probabilities. Therefore, we perform calibration of our risk estimate on the validation set: the SVM estimated probabilities are replaced by probabilities estimated by a logistic regression of disease outcome on the logit of the SVM estimated probabilities.

Analysis
A potential source of dependence between genetic risk and environmental factors is age at diagnosis. For T1D, the MHC region, the region that affects genetic risk the most, is also associated with age at diagnosis[26,27]. Of course, age at diagnosis has a strong impact on the experiences that a child has had before diagnosis and therefore the environment of patients as measured by a questionnaire. Consequently, we assessed association between genetic risk and age at diagnosis

using linear regression on the 1491 Isis-Diab patients of European descent for whom genetic risk was available.

Our main analysis is testing for association between environmental factors and genetic risk. This association was assessed while controlling for age at diagnosis. We used a generalized additive model (GAM) in order for the dependence between environmental factor and age to be captured by a smooth function. The environmental factor was regressed on the genetic risk and a smooth function of age. Association with genetic risk was tested using the standard Student test provided by the fitted GAM. The tests were one sided as explained above. In our case, the associations with disease are negative and therefore the association between genetic risk and environmental factor is expected to be positive: the most at risk genetically were more exposed to protective factors. The mgcv package was used for GAM[28].

Empirical power estimation

We estimate power for the precise setting of each variable tested in the Isis-Diab data. We also consider more generic scenarios in order to evaluate the influence of prevalence, predictive accuracy and sample size on the power of DAC. The simulations performed for each purpose follow the general framework described above but differ on the values of parameters and the definition of environmental risk.

Genetic risk distribution

The distribution of genetic risk in the general population is a mixture of the distribution of genetic risk in the controls and in the patients. If we denote $D(X)$ the distribution of X, we have that:

$D(R_g) = (1-K)D(R_g^{controls}) + KD(R_g^{cases})$.

It should be noted that an individual whose genetic risk comes from the distribution of genetic risk for cases does not necessarily have the disease. In practice, we sample N genetic risks from the genetic risks of Isis-Diab patients and we sample the rest, i.e. the majority, from the genetic risks of the CNAD controls.

It should be noted that the model used for simulation is similar but slightly different from the threshold liability model used in the quantitative genetics literature[29]. The assumption of normality for genetic risk in the threshold liability model is replaced by the actual predicted risk distribution in the population and the noise is Laplace-distributed instead of normally distributed.

Power estimation for Isis-Diab

For the power estimation on the Isis-Diab data, the prevalence K was set to 0.2%, the prevalence of T1D in France, and the sample size was set to the actual sample size available after excluding missing data for each variable.

To define the environmental risk for the Isis-Diab data, we need to choose an effect size for each environmental factor. In order to do this, we took into account the results of the case-control study. Two analyses were performed on the case-control data. Between the two resulting effect size estimate for each variable, we chose the most extreme, i.e. the most favorable scenario for power.

To have a power estimate as precise as possible, we need to obtain in the simulated case population approximately the observed distribution of the environmental factor in cases. To achieve this goal, the distribution in the source population is the observed distribution in cases weighted by the inverse of the effect size.

Power estimation in generic scenarios

For the estimation of power in the generic scenarios, we evaluated the influence of prevalence and then of prediction accuracy of genetic risk. To do this for prevalence, genetic risk was left untouched, we set prevalence to 0.2%, 0.6% or 1% and sample size to 500, 1500, 3000 or 5000. The three prevalences correspond to the prevalence of T1D in France for the lowest, T1D in Finland for the intermediate value and high estimation of prevalence of celiac disease for the highest[21].

Concerning the influence of prediction accuracy of the genetic risk, we set prevalence at 0.2%, we modified the genetic risk estimate to have an AUC of 0.88, 0.90 or 0.92 and we set the sample size to 500, 1500, 3000 or 5000. The genetic risk distribution with modified AUC was obtained by adding to the risk of patients a constant chosen to obtain the desired AUC. The distribution of risk in patients and controls was then calibrated again.
Concerning the definition of the environmental risk, we chose an effect size of 3 which is a large but plausible effect size for epidemiology and we chose the most favorable distribution of the environmental factor in the patients, i.e the one with the most variance: an evenly split binary variable. In the same fashion as above, the distribution in the source population was chosen to obtain the desired distribution in cases.

Code

Code used for analysis and power estimation is available at github.com/FelBalazard/DAC .

Results

Genetic risk

The genetic risk estimation trained on the WTCCC1 yielded an AUC of 0.86 when evaluated on Isis-Diab patients and CNAD controls. This value is intermediate between the AUC of 0.89 obtained in cross-validation on the WTCCC1 data and the AUC of 0.84 obtained on North-American cohorts. This may be due to the use of controls from the same study as the training set. The ROC curve, calibration plot and density plot of the genetic risk are presented in figure 2. The estimator is well calibrated except for the highest intervals. Given the larger proportion of controls compared with cases, those intervals also contain the least observations.

Genetic risk and age at diagnosis

Regression of age on genetic risk yielded a negative association. In average, patients with a genetic risk increased by one standard deviation were younger at diagnosis by 3.5 months (CI=[-5.7,-1.3], $p=2x10^{-3}$). This underscores the importance of controlling for age at diagnosis when looking for an association between genetic risk and environmental factors.

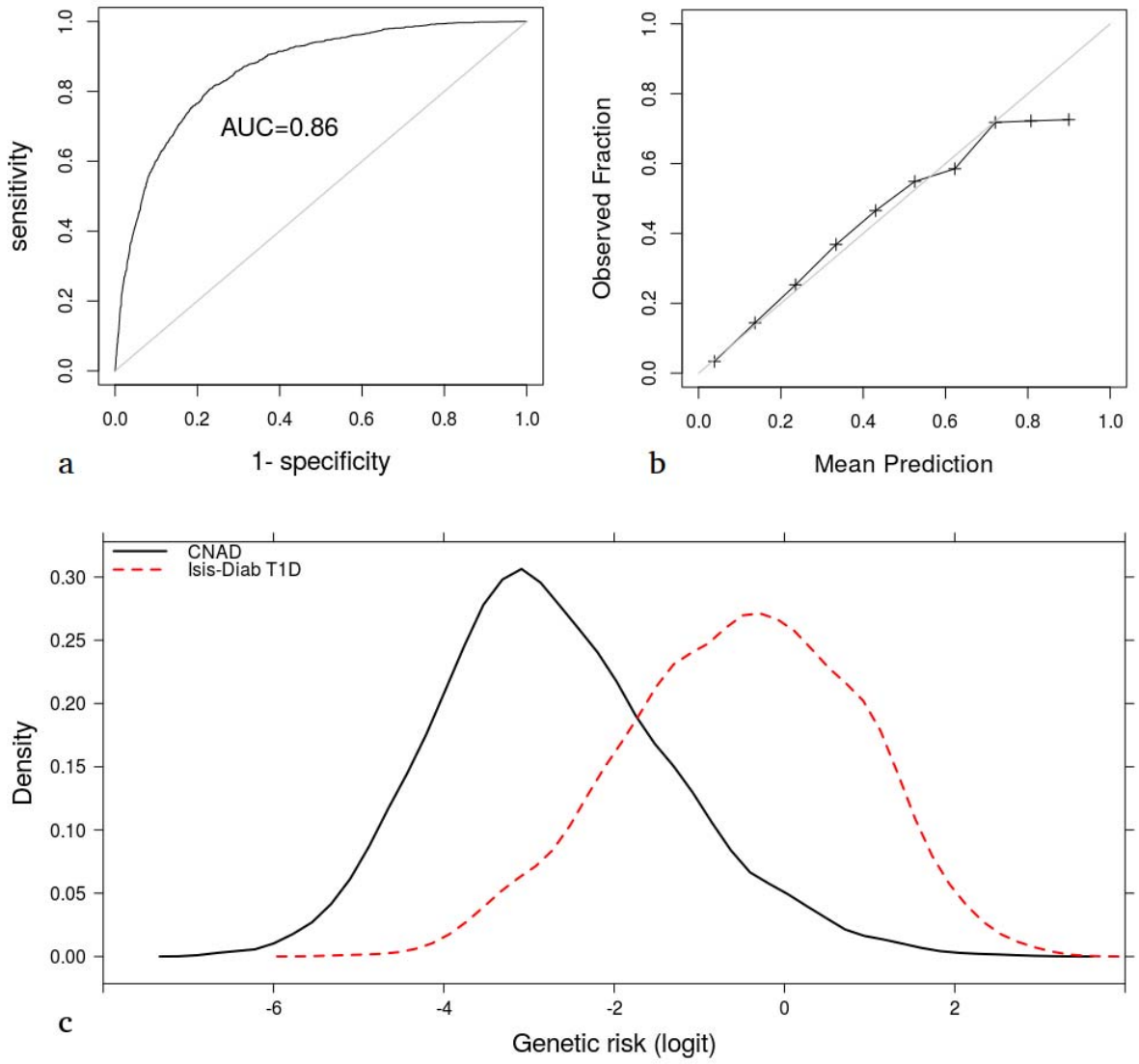

Figure 2: Genetic risk estimation on the CNAD and Isis-Diab patients. a: Receiver operating curve (ROC) of the estimator. The AUC is given. b: Calibration plot after calibration of the risk estimator. The range of values taken by the estimator is divided in 10 bins of equal length. The average prediction is plotted against the actual proportion of cases in each bin. c: Density plot on the logit scale of the risk estimate of the CNAD and Isis-Diab patients.

Results and power of DAC on the Isis-Diab patients

Results are summarized in Table 1. Oral hygiene is nominally associated with genetic risk in the expected direction in Isis-Diab patients. However, this does not take into account correction for multiple testing. Other variables do not show association with genetic risk.

The estimation of power for DAC in the Isis-Diab data showed that DAC had power under 10% for every variable. Given the low power of DAC herein, the nominally significant result for oral hygiene is almost as unlikely under the alternative than under the null. This low power estimate shows that DAC is not informative in the Isis-Diab data.

Variable	Missing data	DAC p-value	Effect size in simulation	Power
Taste for sugar (baby)	4%	0.09	0.59	7%
Dentist	4%	0.80	0.37	10%
Oral hygiene	2%	0.032	0.39	8%
Diarrhea	6%	0.66	0.56	7%
Cocoa spread	0.4%	0.77	0.33	7%
Winter sports	1%	0.22	0.49	8%
Club	1%	0.60	0.49	8%

Table 1 :Results of the DAC method for confirmation of 7 variables from the Isis-Diab case-control study. The effect size used in simulations is the farthest from 1 in the two analysis made for the case-control study.

Power estimation

The results of the power estimation in generic scenarios are presented in figure 3. Power increases with sample size, prediction accuracy of the genetic risk and prevalence of the disease.

With a prevalence of 0.2% and an AUC of 0.86, power was very limited. Even if our sample size had been 5000 cases and despite the favorable assumptions made on the effect size and the distribution of the environmental factor in cases, power would be only 26%.

Our estimation show that power depends strongly on prevalence of the disease. For a disease with prevalence of 1%, 80% power is attained for a sample size under 3000.

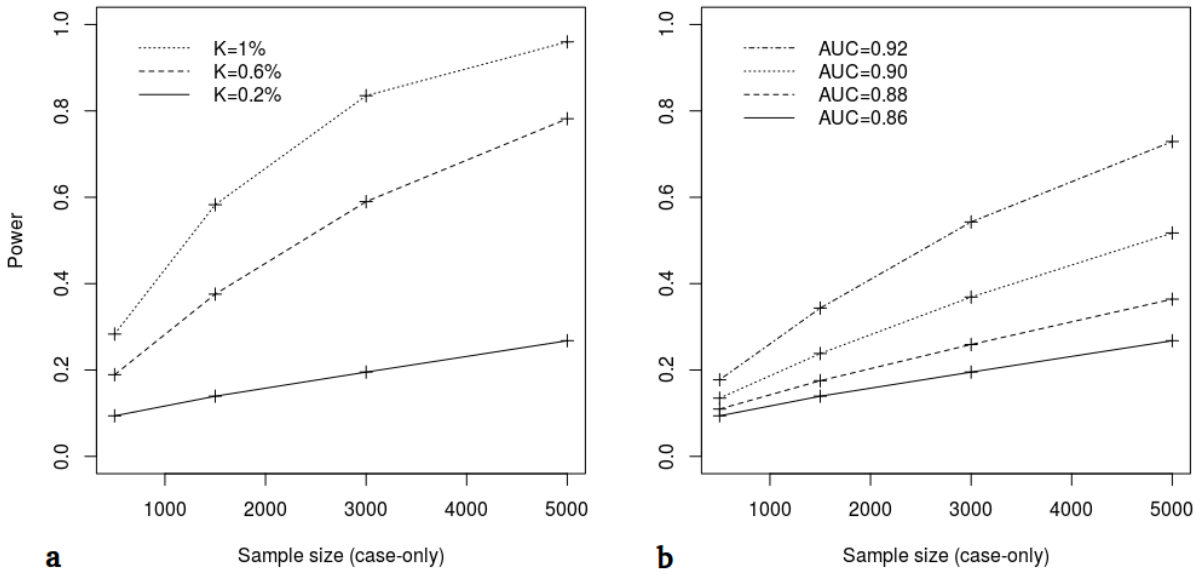

Figure 3: Power of the DAC methodology in different settings. The effect size is set at 3 and the environmental factor in cases is evenly split. a: Influence of the prevalence of the disease on power. The AUC of the genetic risk estimator remains at 0.86. b: Influence of the genetic risk accuracy (AUC) on power. The prevalence of the disease remains at 0.2%

Discussion

Studies of environmental factors of disease are long, difficult and expensive. While the recent progress in the genetic study of disease is in itself a success, it can also be leveraged to inform the environmental determinants of disease. The most successful methodology in this direction is Mendelian randomization but all possibilities have not been explored.

As such a new possibility, we propose DAC to discover or validate associations of environmental factors with disease using genetic risk and environmental data in a case-only setting. Its main strength is that the information it provides is independent from the choice of controls, an important source of confusion in the case-control design. Ideally, DAC should be used after a standard environmental case-control study to validate findings. A practical advantage is that it requires genetic data only on cases.

DAC leverages collider bias. It assumes independence between genetic risk and the environmental factors in the general population as well as lack of interaction between them.

The assumption of independence is reasonable but deviation from it should be kept in mind as an alternative explanation for a positive result. While certain genes affect certain exposures such as alcohol consumption, coffee consumption or smoking[30–32], there is some a priori for independence between most genes and most environmental factors. We stress that the only independence needed is between the aggregated genetic risk score and the environmental factor: DAC does not require independence between each SNP and the environmental factor.

DAC is sensitive to the assumption of multiplicative combination of odd-ratios. Indeed, interactions between genetic risk and environmental factor are problematic. A negative interaction strengthens the negative association that DAC tries to uncover but makes the findings less actionable as the people at highest genetic risk would respond less to intervention on the environmental factor. A positive interaction cancels the negative association that DAC tries to uncover despite increasing the prevention potential of the factor. This is a notable caveat to DAC as interactions between genetic risk and environmental factors have been detected in relation to obesity[33] and must be present in other settings as well.

The epidemiological design closest to DAC is the case-only design for gene-environment interaction (CODGEI)[34]. Like DAC, CODGEI relies on case-only data, crosses genetic data and environmental data and relies on the assumption that gene and environment are independent in the general population. The goal of CODGEI differs from DAC: it aims to discover interactions between gene and environment. Therefore, there is no assumption of absence of interaction between gene and environment in CODGEI.

A second difference is that CODGEI needs a rare disease assumption. This is not limited to the prevalence being negligible compared to 1. If we denote the value taken by a variant i by G_i with values in a set J, the precise assumption is:

For all j in J, $P(D=1|G_i=j)$ is negligible compared to 1.

Under that assumption, an association between a gene and an environmental factor in cases is proof of an interaction between the two. In particular, this tells us that under the rare disease assumption, there will be no effect due to collider bias. Therefore, this shows that DAC has low power for low prevalences of the disease. As a consequence genetics are considered differently in the two methods. DAC needs to maximise the distance to the rare disease assumption and therefore uses an aggregated risk estimation. On the other hand, in CODGEI, each variant is considered on its own.

This theoretical argument for absence of collider bias at low prevalences is in accordance with the results of power estimation. Those power estimations show that DAC can be successful in higher prevalence situations, with large sample sizes and better genetic risk estimation.

In particular, DAC is quite sensitive to the prevalence of the disease. However, in more common diseases, genetic risk estimation typically obtains sensibly weaker results and the prospective cohort design is more feasible. Nevertheless, DAC needs stronger prevalences of the disease to achieve reasonable power. This can mean being applied in a country with high prevalence of T1D such as Finland or on more frequent diseases such as celiac disease.

Given the prevalence of T1D in France, DAC is underpowered in the setting of the Isis-Diab study. Nevertheless, we decided to present the application of our method on this data to illustrate the practical considerations that go into applying DAC such as the problem of confounding by age at diagnosis. Furthermore, it allowed to base our power estimations on the actual predicted genetic risk distribution.

DAC underscores the importance for epidemiology of having a genetic risk estimation as predictive as possible. There has been limited access to the largest consortium datasets for this goal and a consequent turn to methods that use only summary statistics[12,35]. In the case of Crohn's disease and ulcerative colitis, the International Inflammatory Bowel Disease Genetic Consortium dataset was used for this purpose and significant improvement was obtained [9]. Methodology to adapt machine-learning methods to GWAS datasets is also a promising avenue of research[13].

The power estimations we performed for DAC can also be seen as a way to test the validity of the rare disease assumption for CODGEI. This assumption has attracted less scrutiny than the gene-environment independence assumption. It is not always respected in practice[36] and is not mentioned in the review by Thomas of methods for identification for gene-environment interaction[6]. As our power estimations show, this is problematic.

Ethics statement

The research protocol of the Isis-Diab study was approved by the Ethics committee of Ile de France (DC-2008-693) and the Commission Nationale Informatique et Libertés (DR-2010-0035). The ClinicalTrial.gov identifier was NCT02212522. All patients provided written informed consent for participation in the study and donation of samples.

Acknowledgements

We thank Gérard Biau for his comments on the manuscript. We thank Mark Lathrop for the Isis Diab genetic data. We thank Yoichiro Kamatani for performing imputation on the Isis-Diab genetic data. We thank Sophie Valtat for rationalizing the questionnaires. We thank Alain Fourreau, Adeline Guégan, Gaël Leprun and Valérie Jauffret for technical assistance in the epidemiological investigation. We acknowledge the collective effort of the Isis diab collaborative group. We thank the participants and their parents for their time.

We thank the Wellcome Trust Case Control Consortium (WTCCC) for making the Affymetrix data available for our analysis. The WTCCC is funded by Wellcome Trust award 076113, and a full list of the investigators who contributed to the generation of the data are available from http://www.wtccc.org.uk.

Funding

FB acknowledges a PhD grant from Ecole Normale Supérieure. The Isis-Diab study was funded by an ongoing institutional grant from NovoNordisk France and Lilly France, a specific Inserm support lasting from 2008 to 2014, the Programme Hospitalier de Recherche Clinique (PHRC ISIS-DIAB and ISIS-VIRUS), the Agence Nationale de la Recherche (ANR ENVIROGENEPIG), the Association pour la Recherche sur le Diabète (ARD). The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. They also did not have access to the database, nor any opportunity to review the manuscript. The corresponding author had full access to all the data in the study and final responsibility for the decision to submit for publication.

References

1. Hagopian WA, Lernmark A, Rewers MJ, et al. TEDDY--The Environmental Determinants of Diabetes in the Young: an observational clinical trial. *Ann N Y Acad Sci*. 2006 Oct;**1079**:320–326.

2. Rewers M, Ludvigsson J. Environmental risk factors for type 1 diabetes. *The Lancet*. 2016 Jun 4;**387**(10035):2340–2348.

3. Ludvigsson JF, Green PHR. The missing environmental factor in celiac disease. *N Engl J Med*. 2014 Oct 2;**371**(14):1341–1343.

4. Molodecky NA, Kaplan GG. Environmental Risk Factors for Inflammatory Bowel Disease. *Gastroenterol Hepatol*. 2010 May;**6**(5):339–346.

5. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. 2014 Jan 1;**42**(D1):D1001–D1006.

6. Thomas D. Gene–environment-wide association studies: emerging approaches. *Nat Rev Genet*. 2010 Sep 3;**11**(4):259–272.

7. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*. 2014 Sep 15;**23**(R1):R89-98.

8. Wei Z, Wang K, Qu H-Q, et al. From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. *PLOS Genet*. 2009 Oct 9;**5**(10):e1000678.

9. 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 Jun 6;**92**(6):1008–1012.

10. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res*. 2014 Jun 24;gr.169375.113.

11. Abraham G, Tye-Din JA, Bhalala OG, Kowalczyk A, Zobel J, Inouye M. Accurate and Robust Genomic Prediction of Celiac Disease Using Statistical Learning. *PLOS Genet*. 2014 Feb 13;**10**(2):e1004137.

12. Abraham G, Havulinna AS, Bhalala OG, et al. Genomic prediction of coronary heart disease. *Eur Heart J*. 2016 Sep 20;ehw450.

13. Botta V, Louppe G, Geurts P, Wehenkel L. Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies. *PLOS ONE*. 2014 Apr 2;**9**(4):e93379.

14. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010 Apr;**39**(2):417–420.

15. Greenland S. Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias: *Epidemiology*. 2003 May;**14**(3):300–306.

16. Gage SH, Smith GD, Ware JJ, Flint J, Munafò MR. G = E: What GWAS Can Tell Us about the Environment. *PLOS Genet*. 2016 Feb 11;**12**(2):e1005765.

17. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics*. 1946 Jun;**2**(3):47–53.

18. Snoep JD, Morabia A, Hernández-Díaz S, Hernán MA, Vandenbroucke JP. Commentary: A structural approach to Berkson's fallacy and a guide to a history of opinions about it. *Int J Epidemiol*. 2014 Apr;**43**(2):515–521.

19. Whitcomb BW, Schisterman EF, Perkins NJ, Platt RW. Quantification of collider-stratification bias and the birthweight paradox. *Paediatr Perinat Epidemiol*. 2009 Sep;**23**(5):394–402.

20. Balazard F, Le Fur S, Valtat S, et al. Association of environmental markers with childhood type 1 diabetes mellitus revealed by a long questionnaire on early life exposures and lifestyle in a case-control study. *BMC Public Health*. 2016 Sep 29;**16**(1):1021.

21. Gujral N, Freeman HJ, Thomson AB. Celiac disease: Prevalence, diagnosis, pathogenesis and treatment. *World J Gastroenterol WJG*. 2012 Nov 14;**18**(42):6036–6059.

22. 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 Jun 19;**5**(6):e1000529.

23. Consortium T 1000 GP. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012 Nov 1;**491**(7422):56–65.

24. Burton PR, Clayton DG, Cardon LR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007 Jun 7;**447**(7145):661–678.

25. Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge University Press; 2004.

26. Howson JMM, Cooper JD, Smyth DJ, et al. Evidence of Gene-Gene Interaction and Age-at-Diagnosis Effects in Type 1 Diabetes. *Diabetes*. 2012 Nov 1;**61**(11):3012–3017.

27. Caillat-Zucman S, Garchon HJ, Timsit J, et al. Age-dependent HLA genetic heterogeneity of type 1 insulin-dependent diabetes mellitus. *J Clin Invest*. 1992 Dec;**90**(6):2242–2250.

28. Wood S. mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation [Internet]. 2016 [cited 2016 Nov 30]. Available from: https://cran.r-project.org/web/packages/mgcv/index.html

29. Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet*. 1965 Aug 1;**29**(1):51–76.

30. Adkins DE, Clark SL, Copeland WE, et al. Genome-wide meta-analysis of longitudinal alcohol consumption across youth and early adulthood. *Twin Res Hum Genet Off J Int Soc Twin Stud*. 2015 Aug;**18**(4):335–347.

31. Cornelis MC, Byrne EM, Esko T, et al. Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Mol Psychiatry*. 2015 May;**20**(5):647–656.

32. Consortium TT and G. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*. 2010 May;**42**(5):441–447.

33. Tyrrell J, Wood AR, Ames RM, et al. Gene–obesogenic environment interactions in the UK Biobank study. *Int J Epidemiol* [Internet]. [cited 2017 Jan 30]; Available from:

https://academic.oup.com/ije/article/doi/10.1093/ije/dyw337/2886194/Gene-obesogenic-environment-interactions-in-the-UK

34. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994 Jan 30;**13**(2):153–162.

35. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet*. 2016 Nov;**online**.

36. Broeks A, Braaf LM, Huseinovic A, et al. Identification of women with an increased risk of developing radiation-induced breast cancer: a case only study. *Breast Cancer Res*. 2007;**9**(2):R26.