DIABETES/METABOLISM RESEARCH AND REVIEWS
Diabetes Metab Res Rev 2014; **30**: 424–434.
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/dmrr.2510

RESEARCH ARTICLE

Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study

Hye-Seung Lee[1] Brant R. Burkhardt[2] Wendy McLeod[1] Susan Smith[1] Chris Eberhard[1] Kristian Lynch[1] David Hadley[1] Marian Rewers[3] Olli Simell[4] Jin-Xiong She[5] Bill Hagopian[6] Ake Lernmark[7] Beena Akolkar[8] Anette G. Ziegler[9] Jeffrey P. Krischer[1]* The TEDDY study group

[1]*Pediatrics Epidemiology Center, Department of Pediatrics, University of South Florida, Tampa, FL, USA*

[2]*Department of Cell Biology, Microbiology and Molecular Biology, University of South Florida, Tampa, FL, USA*

[3]*Barbara Davis Center for Childhood Diabetes, University of Colorado Denver, Aurora, CO, USA*

[4]*Department of Pediatrics, Turku University Central Hospital, Turku, Finland*

[5]*Center for Biotechnology and Genomic Medicine, Medical College of Georgia, Georgia Regents University, Augusta, GA, USA*

[6]*Pacific Northwest Diabetes Research Institute, Seattle, WA, USA*

[7]*Department of Clinical Sciences, Lund University, Malmo, Sweden*

[8]*National Institute of Diabetes & Digestive & Kidney Disorders, Bethesda, MD, USA*

[9]*Institute of Diabetes Research Helmholtz Zentrum München, and Klinikum rechts der Isar, Technische Universität München, and Forschergruppe Diabetes e.V. Neuherberg, Germany*

*Correspondence to: Jeffrey P. Krischer, Pediatrics Epidemiology Center, Department of Pediatrics, University of South Florida, Tampa, FL, USA.
E-mail: Jeffrey.Krischer@epi.usf.edu

Received: 12 August 2013
Revised: 29 October 2013
Accepted: 4 December 2013

Abstract

Aims The Environmental Determinants of Diabetes in the Young planned biomarker discovery studies on longitudinal samples for persistent confirmed islet cell autoantibodies and type 1 diabetes using dietary biomarkers, metabolomics, microbiome/viral metagenomics and gene expression.

Methods This article describes the details of planning The Environmental Determinants of Diabetes in the Young biomarker discovery studies using a nested case–control design that was chosen as an alternative to the full cohort analysis. In the frame of a nested case–control design, it guides the choice of matching factors, selection of controls, preparation of external quality control samples and reduction of batch effects along with proper sample allocation.

Results and conclusion Our design is to reduce potential bias and retain study power while reducing the costs by limiting the numbers of samples requiring laboratory analyses. It also covers two primary end points (the occurrence of diabetes-related autoantibodies and the diagnosis of type 1 diabetes). The resulting list of case–control matched samples for each laboratory was augmented with external quality control samples. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords batch effects; biomarker discovery; nested case–control design; TEDDY; type 1 diabetes

The Environmental Determinants of Diabetes in the Young (TEDDY) is designed as a prospective cohort study of 8677 children enrolled before 4.5 months of age and followed for 15 years to identify genetic and environmental triggers of type 1 diabetes (T1D). The TEDDY cohort consists of children identified to be of increased genetic risk who either had a parent or sibling with T1D (first-degree relative) or not (general population). TEDDY planned analyses include the comparison of the natural history and biomarkers of those children developing T1D to those who did not. The large cohort size and the high costs of these technologies make the full cohort analysis expensive and inefficient.

Epidemiological designs, such as a nested case–control and a case cohort, are available to improve efficiency in a large cohort study, while providing a

similar result that the full cohort analysis would have produced. The strengths and weaknesses of those available designs have been compared in great detail [1–3]. For biomarker studies, a nested case–control design has more advantages than others, stemming from the ability of matching cases and controls for potentially confounding variables [4–7], as well as the ability of saving more resources because information on time-dependent exposures in controls does not require samples or data to be collected beyond the time of follow-up of the case [8]. However, a nested case–control design requires careful planning to avoid bias and loss of generality while trying to improve efficiency [4]. Failure to select controls as nested in each risk set from the full cohort can produce biassed results [9,10]. Furthermore, a nested case–control design shares the general concerns in considering special sampling techniques, such as the choice of matching factors.

In this article, we present the details of planning the TEDDY biomarker discovery studies using a nested case–control design that was chosen as an alternative to the full cohort analysis. Our design reduces potential bias and retains study power while reducing the costs by limiting the numbers of samples requiring laboratory analyses. It also covers two primary end points (the occurrence of diabetes-related autoantibodies and the diagnosis of T1D). The resulting list of case–control matched samples for each laboratory was augmented with external quality control (QC) samples prepared by the data coordinating centre (DCC) QC laboratory. The external QC samples were masked so that the laboratories were unaware of whether the samples came from cases or controls. We first describe the TEDDY cohort and the application of a nested case–control design and then the steps taken to select controls. The definition of cases and controls are detailed, and the preparation of external QC samples is also described.

Materials and methods

Study population

The Environmental Determinants of Diabetes in the Young enrolled children younger than 4.5 months of age from December 2004 to July 2010 through newborn screening for high-risk HLA-DR-DQ genotypes at six centres: three in the United States at the Pacific Northwest Diabetes Research Institute, Seattle, Washington, the Barbara Davis Center, Denver, Colorado; a combined Georgia/Florida site at the Medical College of Georgia, Augusta, Georgia and the University of Florida, Gainesville, Florida, and three in Europe at University of Turku, (Turku, Oulu and Tampere, Finland); Lund University, Malmo, Sweden and the Diabetes Research Institute, Munich, Germany. Detailed study design and methods have been previously published [11,12]. Written informed consents were obtained for all study participants from a parent or primary caretaker, separately, for genetic screening and participation in prospective follow-up. The study was approved by local institutional review boards and is monitored by external evaluation committee formed by the National Institutes of Health.

The first primary endpoint in TEDDY is the appearance of persistent confirmed islet autoimmunity (IA). Persistent confirmed IA is defined as the presence of one confirmed autoantibody (GAD65A, IA-2A or IAA) on two or more consecutive samples. IAs are measured in two laboratories (Barbara Davis Center, Aurora, Colorado, and the University of Bristol Laboratory, Bristol, UK) depending upon the location of the clinical site. All samples identified as positive in one TEDDY laboratory are sent to the other laboratory for confirmation [13]. The second primary outcome is the clinical appearance of T1D as defined using the American Diabetes Association criteria [14].

The TEDDY study collects participants' stool, plasma, serum, red blood cells, peripheral blood mononuclear cells, along with extensive questionnaire data. Blood sample collection begins at the 3-month study visit and continues at a 3-month interval up to 4 years of age. If a subject develops persistent IA, then they continue on the 3-month interval schedule up to age 15 years; otherwise, they switch to a 6-month interval schedule. In addition, the child's parent collects at least 5 g of the child's stool each month (up until 48 months of age, then every 3 months until the age of 10 years and then biannually thereafter) into the three plastic stool containers provided by the clinical centre. All samples have been stored at a central TEDDY repository by following the centralized DCC instructions [15].

Biomarkers

The design of the TEDDY biomarker studies included the assessment of dietary biomarkers, metabolomics, gene expression and microbiome/viral metagenomics in plasma and stool samples collected at protocol-specified time points from the participating children. A different laboratory responsible for the analysis of each biomarker was selected after carefully reviewing applications received in response to a request for proposals developed by TEDDY investigators.

DOI: 10.1002/dmrr

The dietary biomarker laboratory (Disease Risk Unit, National Institute for Health and Welfare, Helsinki, Finland) was selected to analyse plasma 25-hydroxyvitamin D, vitamin C, alpha/gamma-tocopherol, carotenoid and cholesterol concentrations and erythrocyte fatty acid composition. The metabolomics laboratory (The NIH West Coast Metabolomics Center, University of California Davis, CA, USA) was selected to profile metabolomes using plasma samples. The micobiome/vial metagenomics laboratory (Baylor College of Medicine, Houston, Texas, USA) was selected to identify viral candidates and associated microbiome (bacterial, eukaryotes, viruses) using stool and plasma samples. The gene expression laboratory (Jinfiniti Biosciences LLC, Georgia Health Sciences University, GA, USA) was selected to identify gene expression profiles using mRNA samples.

Study design and application

A subject who developed one of the two primary outcomes (persistent confirmed IA and/or T1D) was defined as a case. The event time of persistent confirmed IA was the date of first blood draw of confirmed IA that was subsequently found to be persistent. The event time of T1D was the date of diagnosis. If the diagnosis was based on two oral glucose tolerance tests (OGTTs), then the date of diagnosis was the first OGTT that met the diagnostic criteria.

In a nested case–control design, controls should be randomly selected among cohort members who have not yet developed the disease at the time a case is diagnosed (risk set sampling or incidence density sampling) [16]. TEDDY defined potential controls for a case as subjects who were event-free within ±45 days of the case's event time, which corresponds to the midpoint between the 3-month interval-scheduled protocol visits at which the events were determined. A control for a case of persistent confirmed IA was a TEDDY participant who had not developed persistent confirmed IA by the time that the case to which it is matched developed IA, within ±45 days of the event time. Because two consecutive samples are involved to determine the persistency, the subject was counted as a potential control if his or her valid sample within ±45 days of the event time was not confirmed positive; or if the sample was confirmed positive, then the following sample had to be not confirmed positive with the available results. A control for a case of T1D was defined as a TEDDY subject who had not been diagnosed as T1D, within ±45 days of the event time. If a subject had an OGTT indicative of diabetes within ±45 days of the event time of the case to which it is matched, the subject was excluded to be a potential control if there was no following OGTT or if the following OGTT met the definition of diabetes.

Matching factors were chosen to be clinical centre, gender and family history of T1D to control the differences in genetic background and in sample/data handling between clinical centres. Although matching is often used to improve statistical efficiency, a minimum number of matching factors is recommended in biomarker studies to avoid overmatching [4,17]. Also, matching on risk factors in a nested case–control design may increase the likelihood that a control becomes a case later during the follow-up than the likelihood in the full cohort.

Because of sample assay costs, the selection of three controls per case was planned for the dietary biomarker and metabolomics samples (1 : 3 matched), and one control per case was planned for gene expression and metagenomics samples (1 : 1 matched). To plan on synergistic and comparative studies across biomarker studies, the same controls for each case were planned to be used for all analyses. All samples collected from TEDDY study visits up to the event time were to be processed. Thus, biospecimen availability at each study visit was also a consideration in the selection of potential controls because the number of samples varied with each type of sample and compliance with protocol visits. A random sample of matched controls from the pool of potential controls resulted in only limited success in finding controls with a high proportion of samples that matched the sample availability of the cases (Figure 1). This approach would have generated about a 40% loss of case–control pairs. To overcome this problem, six potential controls were randomly selected from the pool of controls, and then, three controls were selected on the basis of the best sample availability. The sample availability of a potential control was counted only when the case to which it is matched carried available sample. The best sample availability was based on the ratio of the number of available samples in a potential control to the number of available samples

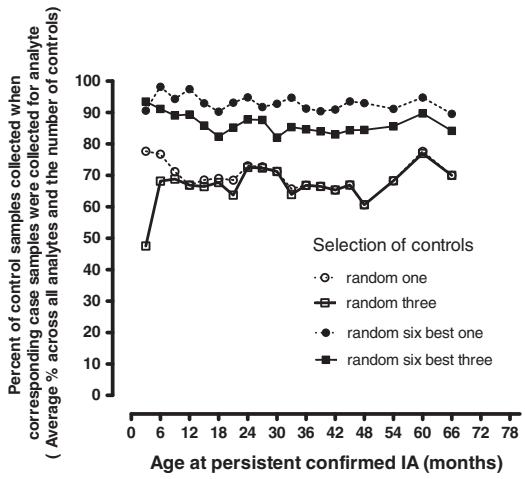

Figure 1. Average number of control samples collected

DOI: 10.1002/dmrr

in the case. Figure 2 summarizes this control selection procedure for TEDDY biomarker studies.

Each analyte was to be run in a 'batch', which was determined by each laboratory. The analytic batch in a laboratory would be a group of biological samples that are analysed together under a particular assay technology. Batch effects were considered because they could create bias in identifying a biomarker by generating a nonidentical random error distribution between cases and controls [18,19], if the corresponding samples were run in different batches. For example, batch effects can occur if a subset of experiments were run on 1 day and another set on the other day, if two technicians were responsible for different subsets of the experiments or if two different lots of reagents, chips or instruments were used [20–23]. To minimize batch effects in comparing case–control samples at the design stage, we arranged all samples collected from a case and matched controls to be run in the same analytic batch. When it was not feasible because of the limited number of samples that the batch can process, we arranged samples that were to be compared with each other (i.e. collected at the same visit) to be run in the same analytic batch.

External QC samples

External QC samples were used to measure batch to batch variability for the various biomarkers being assayed. These QC samples were prepared by the DCC QC laboratory located in the Biomedical Science Facility on the campus of the University of South Florida. Each QC sample was designed to be biologically identical to allow for true evaluation of inter assay variability. Handling/processing and sample volume/storage container appearance was identical to case and control specimens to allow for proper blinding of QC samples to each laboratory. QC samples for the dietary biomarkers, gene expression, metabolomics and microbiome/viral metagenomics are described in Table 1. Following preparation, QC samples were shipped to the TEDDY central repository for storage prior to dissemination among various laboratories for subsequent analysis.

Data analysis plan

Conditional logistic regression will examine the association between a candidate biomarker and becoming a case within a stratum. For high-throughput analyses, false discovery rate will be controlled to filter potential biomarkers, and penalized conditional logistic regression will be used for simultaneous selection [24].

Each analyte from birth to an event time (mostly 3 months apart) will be profiled for a subject. A marker will be analysed as a profile at a given age of interest or a subject specific change estimated from a mixed effects model.

Confounders other than matching factors may be adjusted in the analysis as identified. Biomarkers specific to a matching factor may be missed in this study.

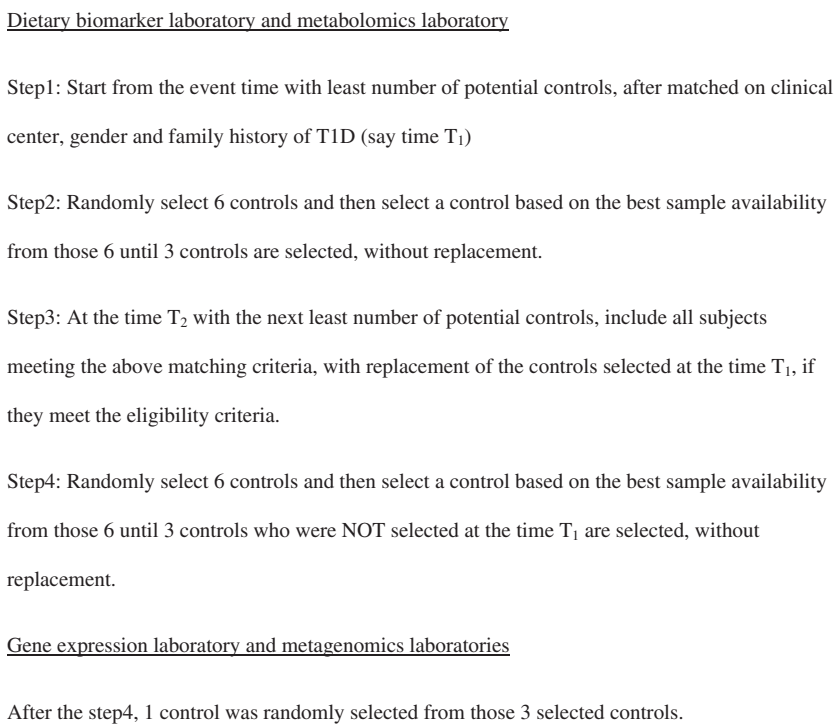

Figure 2. Control selection procedure

DOI: 10.1002/dmrr

Table 1. External quality control (QC) sample

Laboratory		QC sample preparation
Dietary biomarkers	Plasma	Human plasma aliquoted into case/control matched subject vials[1]
Gene expression	Whole blood	RNA isolated from whole blood aliquoted into case/control matched subject vials
Metabolomics	Plasma	Human plasma aliquoted into case/control matched subject vials[1]
Microbiome/viral metagenomics	Stool	Human stool aliquoted into case/control matched subject vials[2]
	Plasma	Viral collection spiked into human plasma and aliquoted into case/control matched subject vials[3]

[1]Human plasma commercially received from Rockland Immunochemicals in four independent lots.
[2]Human stool received from two non-T1D independent donors.
[3]Following viruses were spiked into donor plasma at the documented concentrations: poliovirus (PV 200 PFU/mL), rotavirus (20 000 PFU/mL), vesicular stomatitis virus (200 PFU/mL) and adenovirus (20 000 PFU/mL).

Results

Cases and controls

This nested case–control study was based on the data collected as of 31 May 2012. The median age of follow-up was 40 months with the first quartile (Q1) is 25 months and the third quartile (Q3) is 60 months. There were 114 T1D cases (median age of diagnosis 29 months, Q1 = 19 and Q3 = 41), and 419 persistent confirmed IA cases (median age 21 months, Q1 = 12 and Q3 = 33). However, one persistent confirmed IA case did not have a potential control after matching on clinical centre, gender and family history of T1D.

Two separate nested case-control studies were planned for persistent confirmed IA and T1D; 95 cases were identified for both T1D and persistent confirmed IA, 323 persistent confirmed IA cases were not diagnosed with T1D, and 19 cases developed T1D without previously meeting the criteria for a persistent confirmed IA. Over 50% of cases were from Sweden and Finland, and about 30% of the cases had a first degree of relative in T1D (Table 2).

A total 1253 controls were selected for the persistent confirmed IA studies for the dietary biomarker laboratory and metabolomics laboratory. Except for one case with only two potential controls available, 417 cases were matched with three controls. Of those 1253 controls, 418 controls were selected for persistent confirmed IA studies in the gene expression laboratory and the metagenomics laboratory. For T1D, 342 controls were for 114 cases for the studies in the dietary biomarker laboratory and the metabolomics laboratory. Of those 342 controls, 114 controls were selected for studies in the gene expression laboratory and the metagenomics laboratory. For a control, samples were to be processed only when the matched case had available sample at a corresponding visit.

On the basis of our design, because of the sample availability, there was about a 10% reduced number of pairs for 1 : 1 studies and a 20% reduced number of pairs for 1 : 3 studies, instead of the 40% reduction from the simple random control selection. That is, the numbers of available pairs for IA studies are 1002 pairs (1 : 3) and 376 pairs (1 : 1) and those for T1D studies are 273 pairs (1 : 3) and 102 pairs (1 : 1). As a result, the nested case–control study will have 80% or greater power at a significance level of 5% to detect $\geq$2.01 relative risk (RR) with 1002 pairs if the proportion of exposure was 5%, and it can detect $\geq$3.14 RR with 376 pairs. With 273 pairs, the study will have at least 80% power to detect $\geq$3.83 RR, and with 102 pairs, it will detect $\geq$8.99 RR [25].

Efficiency in the number of samples to be processed

Because of the nature of a nested case–control design, there were controls that subsequently became cases later in their follow-up. Among those 418 persistent confirmed IA cases, 42 (10%) subjects were selected as controls for another persistent confirmed IA cases prior to becoming IA, 23 (6%) were selected as controls for T1D cases and 8 (2%) were selected for both. Among those 114 T1D cases, six (5%) subjects were selected as controls for another T1D cases, six (5%) were selected as controls for persistent confirmed IA cases and one (1%) was selected for both. On the other hand, 116 (9%) controls for persistent confirmed IA cases were also selected as controls for T1D. This links one matched case and its controls to another matched case and its controls. A 'set' was created to include unique subjects from the linkages in persistent confirmed IA case controls and the T1D case controls. The number of subjects in a set increased depending on the complexity of the linkage.

For those persistent confirmed IA cases who also serve as controls for T1D, all samples collected up to the event time of persistent confirmed IA were to be processed, but samples collected from the visits after the time of persistent

DOI: 10.1002/dmrr

Table 2. Study subject characteristics: mean (SD) or *n* (%)

		Persistent confirmed islet autoimmunity		T1D	
		Case	Control	Case	Control
Design	1 : 1	418	418	114	114
	1 : 3	417	1251	114	342
		1*	2		
Age (months)		24 (15) (min = 2, max = 72)		32 (16) (min = 8, max = 75)	
Matching variables					
Clinical site	Colorado	57 (14%)	171 (14%)	16 (14%)	48 (14%)
	Georgia/Florida	29 (7%)	87 (7%)	6 (5%)	18 (5%)
	Washington	38 (9%)	113 (9%)	8 (7%)	24 (7%)
	Finland	114 (27%)	342 (27%)	36 (32%)	108 (32%)
	Germany	37 (9%)	111 (9%)	18 (16%)	54 (16%)
	Sweden	143 (34%)	429 (34%)	30 (26%)	90 (26%)
T1D family history	First-degree relative	95 (23%)	284 (23%)	41 (36%)	123 (36%)
	General population	323 (77%)	969 (77%)	73 (64%)	219 (64%)
Gender	Female	184 (44%)	551 (44%)	61 (54%)	183 (54%)
	Male	234 (56%)	702 (56%)	53 (46%)	159 (46%)

T1D, type 1 diabetes.
*There was one case with only two controls available.

confirmed IA were to be processed only for the visits when the matched T1D case had available sample. For those T1D cases who were preceded by persistent confirmed IA, as well as serve as controls for persistent confirmed IA cases, all samples collected until the T1D event time were to be processed. For those controls selected for both persistent confirmed IA and T1D, samples were to be processed if either case had available sample at the visit.

Although each analyte result derived from a sample will be utilized for all biomarker studies of persistent confirmed IA and/or T1D, samples will only need to be processed once for each analysis in each lab. Extracting unique samples from these sets reduced by about 10% the number of samples that need to be processed in each lab. The third column in Table 3 summarizes the number of samples to be processed per analyte in each biomarker laboratory. For example, persistent confirmed IA study will need 3060 ascorbic acid analysis results, and T1D study will need 1039 results. To do that, 3736 unique samples were identified from 253 sets for ascorbic acid analysis.

To reduce batch effects

As shown in Table 4, the number of samples to be shipped at a single time, as well as the number of samples that can be analysed together, was determined by each laboratory specific to the assay technology.

The dietary biomarker laboratory was capable of handling 56 samples in one batch for vitamin C analysis, 96 samples for vitamin D only analysis, 64 samples for tocopherol and additional vitamin D analysis and 20 samples for fatty acid analysis. After saving places for external QC samples, the remaining number of places was available for case–control samples. For example, 52 places were available for the case and control samples for vitamin C analysis, after leaving four places for external QC samples in the batch. The median number of samples per set was 10 (Q1 = 6 and Q3 = 16). If we planned to run all samples collected from all TEDDY visits in a set, 9 out of 253 sets (4%) would not be run in the same batch because of exceeding the maximum number of samples that the lab can run in the same batch for vitamin C analysis (i.e. 52). The metabolomics laboratory was capable to process 40 samples in one batch. After saving four places for external QC samples, 36 places were available for case–control samples. Although the first attempt was to run all samples collected from all TEDDY visits in the same set, the analytic batch sizes for fatty acid composition analysis in the dietary biomarker laboratory and analysis for metabolomics were very limited. For fatty acid, the median number of samples per set was 12 (Q1 = 8 and Q3 = 19), and about 26% of sets included more than the limit (i.e. 18). For the metabolomics laboratory, the median number of samples per set was 27 (Q1 = 15 and Q3 = 44), and about 35% of sets included the number of samples greater than the limit (i.e. 36). Hence, for 1 : 3 matched studies, a set was modified to include the samples collected at a specific visit in the set, and the allocation of those case–control samples between visits was left at random. For fatty acid analyses, the 266 sets were modified to be 907, and of those modified 907 sets, about 2% exceeded the limit. For the metabolomics laboratory, the 275 sets were modified to be 2308, and 0.1% of those 2308 modified sets exceeded the limit. Modified sets were randomly ordered, as well as the subjects within a modified set. Those modified sets exceeding the limit in each analyte were randomly divided and allocated in the consecutive batches.

DOI: 10.1002/dmrr

Table 3. Number of samples to be processed per analyte

		Persistent confirmed islet autoimmunity			Type 1 diabetes			Combined			
Laboratory	Analyte	Case + control	Control	Case	Case + Control	Control	Case	Case + Control	Control	Case	Number of sets to provide samples
Dietary biomarkers	Ascorbic acid	3 060	2 259	801	1 039	760	279	3 736	2 779	957	253
	Vitamin D (3 or 9 months)	2 214	1 604	610	619	447	172	2 468	1 825	643	246
	Alpha/gamma-tocopherol and vitamin D (other than 3 and 9 months)	3 069	2 268	801	1 029	755	274	3 736	2 780	956	253
	Fatty acid	4 106	3 004	1102	1 346	979	367	4 889	3 621	1268	266
Metabolomics	Plasma	9 394	6 877	2517	3 291	2 372	919	11 571	8 486	3085	275
Microbiome/viral metagenomics	Stool	10 446	4 593	5853	3 821	1 728	2093	13 073	6 033	7 040	399
	Plasma	4 948	2 372	2 576	1 799	864	935	6 230	3 078	3 152	404
Gene expression	mRNA	4 004	1 808	2 196	1 484	680	804	5 080	2 369	2 711	398

Table 4. Number of samples per batch after sample allocation

		Number of places allowed per batch		Number of case–control samples per set in one batch: median (Q1,Q3)			
Laboratory	Analyte	Case controls	External QC	Including all TEDDY visits	At a given TEDDY visit	Number of external QC samples (number of batches)	Number of samples per shipment
Dietary biomarkers	Ascorbic acid	52	4	10 (6,16)	4 (4,4)	312 (78)	~1200
	Vitamin D (3 or 9 months)	92	4	7 (4,8)	4 (4,6)	112 (28)	~2500
	Alpha/gamma-tocopherol and vitamin D (other than 3 and 9 months)	60	4	10 (5,16)	4 (4,4)	268 (67)	~1400
	Fatty acid	18	2	12 (8,19)	4 (4,4)	638 (319)	~1000
Metabolomics	Plasma	36	4	27 (15,44)	4 (4,4)	1388 (347)	~500
Microbiome/viral metagenomics	Stool	93	2	28 (15,44)	2 (2,2)	330 (165)	~1000
	Plasma	93	2	14 (8,21)	2 (2,2)	150 (75)	~1000
Gene expression	mRNA	94	2	11 (6,18)	2 (2,2)	120 (60)	~350

QC, quality control; TEDDY, The Environmental Determinants of Diabetes in the Young.

For the metagenomics and the gene expression studies that used 1 : 1 matching, the analytic batch size made it possible to include all samples collected from all TEDDY visits in a set. About 2% of sets (6 out of 399) exceeded the limit only for stool analysis, which led to a random divide into two consecutive batches.

Number of samples and batches to be processed

After decoding the linked complexity and arranging batches to minimize the effects, there were 3736 samples to be processed in 78 batches for vitamin C analysis, 2468 samples in 28 batches for vitamin D analysis, 3736 samples in 67 batches for tocopherol and vitamin D analysis, 4889 samples in 319 batches for fatty acid analysis, 11 571 samples in 347 batches for metabolomics analysis, 13 073 samples in 165 batches for metagenomic analysis using stool, 6230 samples in 75 batches for metagenomic analysis using plasma and 5080 samples in 60 batches for gene expression analysis.

Discussion

With the TEDDY experience as an example, we presented the implementation of a modified nested case–control design specific to multiple biomarker studies for IA and T1D. Major steps taken were the choice of an epidemiological design, the choice of handling sample availability and the choice to reduce batch effects.

In implementing a nested case–control design, a control selected to match one case is possibly selected to match another case, and if an individual selected as a control develops the disease later, this person can also serve as a case. In practice, this aspect brings up negative reaction when investigators seek to select controls from among those who remain disease free throughout the follow-up, which is typically used in biomarker discovery studies. But a random selection of controls from a clearly defined risk set is necessary to obtain unbiased results as it has been discussed in previous studies [4,10].

We chose an approach to selecting controls with the most available samples from a random sample of potential controls. This markedly improved the efficiency of the case–control study over a random selection of controls by increasing the number of samples available for each analysis. Although one option was to match on entirely cases' sample availability, concerns were raised from the possible spectra of bias because protocol compliance (i.e. sample availability) might be correlated with an environmental trigger of IA or diabetes. Hence, our choice was to randomly select six subjects from the pool of potential controls and then select the controls from them, on the basis of the best sample availability. This approach was intended to mediate the concern of bias, while saving the efficiency from the potential loss in selecting completely at random.

Additionally, the organization of the selected cases and controls and their samples to accommodate the varying batch sizes posed logistical challenges in order to minimize the batch effects when case–control comparisons are to be made. Although each laboratory strives to be certain that analytic results obtained from a given sample are not influenced by a particular batch, the variability of analytic results within a batch is smaller than the variability of analytic results between batches. We avoided potential batch effects by arranging those samples in a set of cases and their matched controls to be run in the same analytic batch. However, because of the large volume of samples, the process in each laboratory will take between 8 and15 months, and batch effects because of this aspect are unavoidable. For example, 60 batches will be processed in the gene expression lab over 15 months. In our setting, the external QC data results will provide useful sources to assess the batch effects.

Through this process, the careful setting of risk sets retained the advantages expected by a nested case–control design. The selection of controls resulted in less than a 20% loss of the case–control pairs because of sample availability, the selection of samples improved by 10% the efficiency of analysing the two primary TEDDY end points and also reduced potential laboratory variability because of batch effects.

Acknowledgements

The TEDDY study group (see appendix)

Funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836 and UC4 DK95300 and contract no. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Juvenile Diabetes Research Foundation (JDRF) and Centers for Disease Control and Prevention (CDC).

Conflicts of interest

None declared.

Appendix

The TEDDY study group

Colorado Clinical Center: Marian Rewers, MD, PhD, PI[1,4,6,10,11], Katherine Barriga[12], Kimberly Bautista[12], Judith Baxter[9,12,15], George Eisenbarth, MD, PhD, Nicole Frank[2], Patricia Gesualdo[2,6,12,14,15], Michelle Hoffman[12,13,14], Lisa Ide, Rachel Karban[12], Edwin Liu, MD[13], Jill Norris, PhD[2,3,12], Kathleen Waugh[6,7,12,15], Adela Samper-Imaz, Andrea Steck, MD[3]. University of Colorado, Anschutz Medical Campus, Barbara Davis Center for Childhood Diabetes.

Georgia/Florida Clinical Center: Jin-Xiong She, PhD, PI[1,3,4,11,†], Desmond Schatz, MD*[4,5,7,8], Diane Hopkins[12], Leigh Steed[12,13,14,15], Jamie Thomas*[6,12], Katherine Silvis[2], Michael Haller, MD*[14], Meena Shankar*[2], Eleni Sheehan*, Melissa Gardiner, Richard McIndoe, PhD, Haitao Liu, MD†, John Nechtman†, Ashok Sharma, Joshua Williams, Gabriela Foghis, Stephen W. Anderson, MD^. Medical College of Georgia, Georgia Regents University. *University of Florida, †Jinfiniti Biosciences LLC, Augusta, GA, ^Pediatric Endocrine Associates, Atlanta, GA.

Germany Clinical Center: Anette G. Ziegler, MD, PI[1,3,4,11], Andreas Beyerlein PhD[2], Ezio Bonifacio PhD*[5], Michael Hummel, MD[13], Sandra Hummel, PhD[2], Kristina Foterek[¥2], Mathilde Kersting, PhD[¥2], Annette Knopff[7], Sibylle Koletzko, MD[¶13], Claudia Peplow[12], Roswith Roth, PhD[9], Julia Schenkel[2,12], Joanna Stock[9,12], Elisabeth Strauss[12], Katharina Warncke, MD[14], Christiane Winkler, PhD[2,12,15]. Forschergruppe Diabetes e.V. at Helmholtz Zentrum München. *Center for Regenerative Therapies, TU Dresden, ¶Dr von Hauner Children's Hospital, Department of Gastroenterology, Ludwig Maximillians University Munich, ¥Research Institute for Child Nutrition, Dortmund.

Finland Clinical Center: Olli G. Simell, MD, PhD, PI[¥^1,4,11,13], Heikki Hyöty, MD, PhD*[±6], Jorma Ilonen, MD, PhD[¥3], Mikael Knip, MD, PhD*[±], Annika Koivu[¥^], Mirva Koreasalo*[±§2], Miia Kähönen[µ¤], Maria Lönnrot, MD, PhD*[±6], Katja Multasuo[µ¤], Elina Mäntymäki[¥^], Juha Mykkänen, PhD[^¥3], Kirsti Näntö-Salonen, MD, PhD[¥^12], Tiina Niininen[±*12], Mia Nyblom*[±], Jenna Rautanen[±§], Anne Riikonen*[±], Minna Romo[¥^], Aaro Simell[¥^], Barbara Simell[¥^9,12,15], Tuula Simell, PhD[¥^9,12], Ville Simell[^¥13], Maija Sjöberg[¥^12,14], Aino Stenius[µ¤12], Jorma Toppari, MD, PhD, Eeva Varjonen[¥^12], Riitta Veijola, MD, PhD[µ¤14], Suvi M. Virtanen, MD, PhD*[±§2], Mari Åkerlund*[±§]. ¥University of Turku, *University of Tampere, µUniversity of Oulu, ^Turku University Hospital, ±Tampere University Hospital, ¤Oulu University Hospital, §National Institute for Health and Welfare, Finland, ¶University of Kuopio.

Sweden Clinical Center: Åke Lernmark, PhD, PI[1,3,4,5,6,8,10,11,15], Daniel Agardh, MD, PhD[13], Carin Andrén-Aronsson[2,13], Maria Ask, Jenny Bremer, Ulla-Marie Carlsson, Corrado Cilio, PhD, MD[5], Emilie Ericson-Hallström[2], Lina Fransson, Thomas Gard, Joanna Gerardsson, Rasmus Håkansson, Monica Hansen, Gertie Hansson[12,14], Susanne Hyberg, Fredrik Johansen, Berglind Jonasdottir MD, Linda Jonsson, Helena Larsson MD, PhD [6,14], Barbro Lernmark, PhD[9,12], Maria Månsson-Martinez, Maria Markan, Theodosia Massadakis, Jessica Melin[12], Zeliha Mestan, Anita Nilsson, Emma Nilsson, Kobra Rahmati, Anna Rosenquist, Falastin Salami, Monica Sedig Järvirova, Sara Sibthorpe, Birgitta Sjöberg, Ulrica Swartling, PhD[9,12], Erika Trulsson, Carina Törn, PhD[3,15], Anne Wallin, Åsa Wimar[12], Sofie Åberg. Lund University.

Washington Clinical Center: William A. Hagopian, MD, PhD, PI[1,3,4, 5, 6,7,11,13, 14], Xiang Yan, MD, Michael Killian[6,7,12,13], Claire Cowen Crouch[12,14,15], Kristen M. Hay[2], Stephen Ayres, Carissa Adams, Brandi Bratrude, David Coughlin, Greer Fowler, Czarina Franco, Carla Hammar, Diana Heaney, Patrick Marcus, Arlene Meyer, Denise Mulenga, Elizabeth Scott, Jennifer Skidmore[2], Joshua Stabbert, Viktoria Stepitova, Nancy Williams. Pacific Northwest Diabetes Research Institute.

Pennsylvania Satellite Center: Dorothy Becker, MD, Margaret Franciscus[12], MaryEllen Dalmagro-Elias Smith[2], Ashi Daftary, MD, Mary Beth Klein. Children's Hospital of Pittsburgh of UPMC.

Data Coordinating Center: Jeffrey P. Krischer, PhD, PI[1,4,5,10,11], Michael Abbondondolo, Sarah Austin-Gonzalez, Rasheedah Brown[12,15], Brant Burkhardt, PhD[5,6], Martha Butterworth[2], David Cuthbertson, Christopher Eberhard, Steven Fiske[9], Veena Gowda, David Hadley, PhD[3,13], Page Lane, Hye-Seung Lee, PhD[1,2,13,15], Shu Liu, Xiang Liu, PhD[2,9,12], Kristian Lynch, PhD [5,6,9,15], Jamie Malloy, Cristina McCarthy[12,15], Wendy McLeod[2,5,6,13,15], Laura Smith, PhD[9,12], Susan Smith[12,15], Roy Tamura, PhD[1,2,13], Ulla Uusitalo, PhD[2,15], Kendra Vehik, PhD[4,5,6,14,15], Earnest Washington, Jimin Yang, PhD, RD[2,15]. University of South Florida.

Project scientist: Beena Akolkar, PhD[1,3,4,5, 6,7,10,11]. National Institutes of Diabetes and Digestive and Kidney Diseases.

Other contributors: Kasia Bourcier, PhD[5], National Institutes of Allergy and Infectious Diseases. Thomas Briese, PhD[6,15], Columbia University. Suzanne Bennett Johnson, PhD[9,12], Florida State University. Steve Oberste, PhD[6], Centers for Disease Control and Prevention. Eric Triplett, PhD[6], University of Florida.

Autoantibody Reference Laboratories: Liping Yu, MD^ [5], Dongmei Miao, MD^, Polly Bingley, MD, FRCP*[5], Alistair Williams*, Kyla Chandler*, Saba Rokni*, Anna Long PhD*, Joanna Boldison*, Jacob Butterly*, Jessica Broadhurst*, Gabriella Carreno*, Rachel Curnock*, Peter Easton*, Ivey Geoghan*, Julia Goode*, James Pearson*, Charles Reed*, Sophie Ridewood*, Rebecca Wyatt*. ^Barbara Davis Center for Childhood Diabetes, University of Colorado Denver, *School of Clinical Sciences, University of Bristol UK.

Cortisol Laboratory: Elisabeth Aardal Eriksson, MD, PhD, Ewa Lönn Karlsson. Department of Clinical Chemistry, Linköping University Hospital, Linköping, Sweden.
Dietary Biomarkers Laboratory: Iris Erlund, PhD[2], Irma Salminen, Jouko Sundvall, Jaana Leiviskä, Mari Lehtonen, PhD National Institute for Health and Welfare, Helsinki, Finland.
HbA1c Laboratory: Randie R. Little, PhD, Alethea L. Tennill. Diabetes Diagnostic Laboratory, Department of Pathology, University of Missouri School of Medicine.
HLA Reference Laboratory:: Henry Erlich, PhD[3], Teodorica Bugawan, Maria Alejandrino. Department of Human Genetics, Roche Molecular Systems.
Metabolomics Laboratory: Oliver Fiehn, PhD, Bill Wikoff, PhD, Tobias Kind, PhD, Mine Palazoglu, Joyce Wong, Gert Wohlgemuth. UC Davis Metabolomics Center.
Microbiome and Viral Metagenomics Laboratory: Joseph F. Petrosino, PhD[6]. Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine.
OGTT Laboratory:: Santica M. Marcovina, PhD, Sc.D. Northwest Lipid Metabolism and Diabetes Research Laboratories, University of Washington.
Repository: Heather Higgins, Sandra Ke. NIDDK Biosample Repository at Fisher BioServices.
RNA Laboratory and Gene Expression Laboratory: Jin-Xiong She, PhD, PI[1,3,4,11], Richard McIndoe, PhD, Haitao Liu, MD, John Nechtman, Yansheng Zhao, Na Jiang, MD Jinfiniti Biosciences, LLC.
SNP Laboratory: Stephen S. Rich, PhD[3], Wei-Min Chen, PhD[3], Suna Onengut-Gumuscu, PhD[3], Emily Farber, Rebecca Roche Pickin, PhD, Jordan Davis, Dan Gallo. Center for Public Health Genomics, University of Virginia.

Committees

[1]Ancillary Studies, [2]Diet, [3]Genetics, [4]Human Subjects/Publicity/Publications, [5]Immune Markers, [6]Infectious Agents, [7]Laboratory Implementation, [8]Maternal Studies, [9]Psychosocial, [10]Quality Assurance, [11]Steering, [12]Study Coordinators, [13]Celiac Disease, [14]Clinical Implementation, [15]Quality Assurance Subcommittee on Data Quality.

References

1. Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. *J Clin Epidemiol* 1999; **52**(12): 1165–1172. Epub 1999/12/02.
2. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *Am J Epidemiol* 1990; **131**(1): 169–176. Epub 1990/01/01.
3. Wacholder S, Gail M, Pee D. Selecting an efficient design for assessing exposure-disease relationships in an assembled cohort. *Biometrics* 1991; **47**(1): 63–76. Epub 1991/03/01.
4. Rundle A, Ahsan H, Vineis P. Better cancer biomarker discovery through better study design. *Eur J Clin Invest* 2012; **42**(12): 1350–1359.
5. Baker SG. Improving the biomarker pipeline to develop and evaluate cancer screening tests. *J Natl Cancer Inst* 2009; **101**(16): 1116–1119. Epub 2009/07/04.
6. Marshall E. Getting the noise out of gene arrays. *Science* 2004; **306**(5696): 630–631. Epub 2004/10/23.
7. Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. *Cancer Epidemiol Biomarkers Prev* 2005; **14**(8): 1899–1907. Epub 2005/08/17.
8. Wacholder S. Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology* 1991; **2**(2): 155–158. Epub 1991/03/01.
9. Lubin JH, Gail MH. Biased selection of controls for case-control analyses of cohort studies. *Biometrics* 1984; **40**(1): 63–75. Epub 1984/03/01.
10. Niccolai LM, Ogden LG, Muehlenbein CE, Dziura JD, Vazquez M, Shapiro ED. Methodological issues in design and analysis of a matched case-control study of a vaccine's effectiveness. *J Clin Epidemiol* 2007; **60**(11): 1127–1131. Epub 2007/10/17.
11. The Environmental Determinants of Diabetes in the Young (TEDDY) study: study design. *Pediatr Diabetes* 2007; **8**(5): 286–298. Epub 2007/09/14.
12. The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Ann N Y Acad Sci* 2008; **1150**: 1–13. Epub 2009/01/06.
13. Bonifacio E, Yu L, Williams AK, *et al*. Harmonization of glutamic acid decarboxylase and islet antigen-2 autoantibody assays for national institute of diabetes and digestive and kidney diseases consortia. *J Clin Endocrinol Metabol* 2010; **95**(7): 3360–3367. Epub 2010/05/07.
14. Puavilai G, Chanprasertyotin S, Sriphrapradaeng A. Diagnostic criteria for diabetes mellitus and other categories of glucose intolerance: 1997 criteria by the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (ADA), 1998 WHO consultation criteria, and 1985 WHO criteria. World Health Organization. *Diabetes Res Clin Pract* 1999; **44**(1): 21–26. Epub 1999/07/22.
15. Vehik K, Fiske SW, Logan CA, *et al*. Methods, quality control and specimen management in an international multicenter investigation of type 1 diabetes: TEDDY. *Diabetes Metab Res Rev* 2013; **29**(7): 557–567. Epub 2013/05/16.
16. Rothman K. Modern Epidemiology. Little, Brown and Company: Boston, 1986.
17. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. III. Design options. *Am J Epidemiol* 1992; **135**(9): 1042–1050. Epub 1992/05/01.
18. Rundle A. Environmental Health Sciences New York: Mailman School of Public Health, 2000; 152.
19. Schulte P, Perera F. Molecular Epidemiology: Principles and Practices. Academic Press: San Diego (CA), 1993.
20. Leek JT, Scharpf RB, Bravo HC, *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010; **11**(10): 733–739. Epub 2010/09/15.
21. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 2000; **97**(18): 10101–10106. Epub 2000/08/30.
22. Benito M, Parker J, Du Q, *et al*. Adjustment of systematic microarray data biases. *Bioinformatics* 2004; **20**(1): 105–114. Epub 2003/12/25.

23. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007; **8**(1): 118–127. Epub 2006/04/25.

24. Sun H, Wang S. Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data. *Stat Med* 2013; **32**(12): 2127–2139. Epub 2012/12/06.

25. Lachin JM. Sample size evaluation for a multiply matched case-control study using the score test from a conditional logistic (discrete Cox PH) regression model. *Stat Med* 2008; **27**(14): 2509–2523. Epub 2007/09/22.