

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

Hye-Seung Lee¹, Brant R. Burkhardt², Wendy McLeod¹, Susan Smith¹, Chris Eberhard¹,
Kristian Lynch¹, David Hadley¹, Marian Rewers⁴, Olli Simell⁷, Jin-Xiong She⁸, Bill
Hagopian⁹, Ake Lernmark¹⁰, Beena Akolkar¹¹, Anette-G. Ziegler³, Jeffrey P. Krischer¹, and
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 33620
3. Institute of Diabetes Research, Helmholtz Zentrum München, and Klinikum rechts der
Isar, Technische Universität München, and Forschergruppe Diabetes e.V. Neuherberg,
Germany
4. Barbara Davis Center for Childhood Diabetes, University of Colorado Denver, Aurora CO,
USA
7. Department of Pediatrics, Turku University Central Hospital, Turku, Finland
8. Center for Biotechnology and Genomic Medicine, Medical College of Georgia, Georgia
Regents University, Augusta GA, USA
9. Pacific Northwest Diabetes Research Institute, Seattle WA, USA
10. Department of Clinical Sciences, Lund University, Malmö, Sweden
11. National Institute of Diabetes & Digestive & Kidney Disorders, Bethesda MD, USA

Corresponding author:

Jeffrey P. Krischer, Ph.D.

Pediatrics Epidemiology Center, Department of Pediatrics, University of South Florida,
Tampa FL, USA

This article has been accepted for publication and undergone full peer review but has not
been through the copyediting, typesetting, pagination and proofreading process which may
lead to differences between this version and the Version of Record. Please cite this article
as doi: 10.1002/dmrr.2510

Abstract

Aims: The Environmental Determinants of Diabetes in the Young (TEDDY) planned biomarker discovery studies on longitudinal samples for persistent confirmed islet cell autoantibodies and type 1 diabetes (T1D) using dietary biomarkers, metabolomics, microbiome/viral metagenomics and gene expression.

Methods: This paper describes 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. 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 reduce 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.

Key words: batch effects, biomarker discovery, nested case-control design, TEDDY, type 1 diabetes

Word count: 4000, Tables: 4, Figures: 2

The Environmental Determinants of Diabetes in the Young (TEDDY) is designed as a prospective cohort study of 8,677 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 since 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 biased 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 paper, 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 reduces 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 center (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

TEDDY 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 centers: three in the US 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 (PBMCs), 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 5g 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 center. 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.

The dietary biomarker laboratory (Disease Risk Unit, National Institute for Health and Welfare, Helsinki, Finland) was selected to analyze 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 microbiome/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. Since two consecutive samples are involved to determine the persistency, the subject was counted as a potential control if his/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 center, gender and family history of T1D to control the differences in genetic background and in sample/data handling between clinical centers. 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.

Due to 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, bio-specimen availability at each study visit was also a consideration in the selection of potential controls since 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, 6 potential controls were randomly selected from the pool of controls, and then 3 controls were selected based on 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 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 analyzed together under a particular assay technology. Batch effects were considered because they could create bias in identifying a biomarker by generating a non-identical 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 was run on one 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 due 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 analyzed 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.

RESULTS

Cases and controls

This nested case-control study was based on the data collected as of May 31, 2012. The median age of follow-up was 40 months with the first quartile (Q1) =25 months and the third quartile (Q3) =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 center, gender and family history of T1D.

While 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 1 case with only 2 potential controls available, 417 cases were matched with 3 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 selected 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.

Based upon our design, due to 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 ≥ 2.01 relative risk (RR)

with 1002 pairs if the proportion of exposure was 5%, and it can detect ≥ 3.14 RR with 376 pairs. With 273 pairs, the study will have at least 80% power to detect ≥ 3.83 RR, and with 102 pairs, it will detect ≥ 8.99 RR (25).

Efficiency in the number of samples to be processed

Due to 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, 6 (5%) subjects were selected as controls for another T1D cases, 6 (5%) were selected as controls for persistent confirmed IA cases, and 1 (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 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 3,060 ascorbic acid analysis results, and T1D study will need 1,039 results. To do that, 3,736 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 analyzed 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 4 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 due to 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 4 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.

For the metagenomics and the gene expression studies which 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 2 consecutive batches.

Number of samples and batches to be processed

After decoding the linked complexity and arranging batches to minimize the effects, there were 3,736 samples to be processed in 78 batches for vitamin C analysis, 2,468 samples in 28 batches for vitamin D analysis, 3,736 samples in 67 batches for tocopherol and vitamin D analysis, 4,889 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, 6,230 samples in 75 batches for metagenomic analysis using plasma, 5,080 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 since protocol compliance (i.e., sample availability) might be correlated with an environmental trigger of IA or diabetes. Hence, our choice was to randomly select 6 subjects from the pool of potential controls and then select the controls from them, based on 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, due to the large volume of samples, the process in each laboratory will take between 8 and 15 months, and batch effects due to 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 due to sample availability, the selection of samples improved by 10% the efficiency of analyzing the two primary TEDDY end points, and also reduced potential laboratory variability due to batch effects.

TEDDY Study 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).

Appendix

The Teddy Study Group

Colorado Clinical Center: Marian Rewers, M.D., Ph.D., PI^{1,4,6,10,11}, Katherine Barriga¹², Kimberly Bautista¹², Judith Baxter^{9,12,15}, George Eisenbarth, M.D., Ph.D., Nicole Frank², Patricia Gesualdo^{2,6,12,14,15}, Michelle Hoffman^{12,13,14}, Lisa Ide, Rachel Karban¹², Edwin Liu, M.D.¹³, Jill Norris, Ph.D.^{2,3,12}, Kathleen Waugh^{6,7,12,15}, Adela Samper-Imaz, Andrea Steck, M.D.³. University of Colorado, Anschutz Medical Campus, Barbara Davis Center for Childhood Diabetes.

Georgia/Florida Clinical Center: Jin-Xiong She, Ph.D., PI^{1,3,4,11}, †, Desmond Schatz, M.D.*^{4,5,7,8}, Diane Hopkins¹², Leigh Steed^{12,13,14,15}, Jamie Thomas*^{6,12}, Katherine Silvis², Michael Haller, M.D.*¹⁴, Meena Shankar*², Eleni Sheehan*, Melissa Gardiner, Richard McIndoe, Ph.D., Haitao Liu, M.D.†, John Nechtman†, Ashok Sharma, Joshua Williams, Gabriela Foghis, Stephen W. Anderson, M.D.^ . 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, M.D., PI^{1,3,4,11}, Andreas Beyerlein Ph.D.², Ezio Bonifacio Ph.D.*⁵, Michael Hummel, M.D.¹³, Sandra Hummel, Ph.D.², Kristina Foterek^{‡2}, Mathilde Kersting, Ph.D.^{‡2}, Annette Knopff⁷, Sibylle Koletzko, M.D.^{¶13}, Claudia Peplow¹², Roswith Roth, Ph.D.⁹, Julia Schenkel^{2,12}, Joanna Stock^{9,12}, Elisabeth Strauss¹², Katharina Warncke, M.D.¹⁴, Christiane Winkler, Ph.D.^{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 Maximilians University Munich, [¥]Research Institute for Child Nutrition, Dortmund.

Finland Clinical Center: Olli G. Simell, M.D., Ph.D., PI^{¥^1,4,11,13}, Heikki Hyöty, M.D., Ph.D.*^{±6}, Jorma Ilonen, M.D., Ph.D. [¥] [¶] ³, Mikael Knip, M.D., Ph.D.*[±], Annika Koivu^{¥^}, Mirva Koreasalo*^{±§2}, Miiä Kähönen^{μ*}, Maria Lönnrot, M.D., Ph.D.*^{±6}, Katja Multasuo^{μ*}, Elina Mäntymäki^{¥^}, Juha Mykkänen, Ph.D. ^{^¥} ³, Kirsti Näntö-Salonen, M.D., Ph.D. ^{¥^12}, Tiina Niininen^{±*12}, Mia Nyblom*[±], Jenna Rautanen^{±§}, Anne Riikonen*[±], Minna Romo^{¥^}, Aaro Simell^{¥^}, Barbara Simell^{¥^9,12,15}, Tuula Simell, Ph.D. ^{¥^9,12}, Ville Simell^{^¥13}, Maija Sjöberg^{¥^12,14}, Aino Stenius^{μ*12}, Jorma Toppari, M.D., Ph.D., Eeva Varjonen^{¥^12}, Riitta Veijola, M.D., Ph.D. ^{μ*14}, Suvi M. Virtanen, M.D., Ph.D.*^{±§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, Ph.D., PI^{1,3,4,5,6,8,10,11,15}, Daniel Agardh, M.D., Ph.D.¹³, Carin Andrén-Aronsson^{2,13}, Maria Ask, Jenny Bremer, Ulla-Marie Carlsson, Corrado Cilio, Ph.D., M.D.⁵, Emilie Ericson-Hallström², Lina Fransson, Thomas Gard, Joanna Gerardsson, Rasmus Håkansson, Monica Hansen, Gertie Hansson^{12,14}, Susanne Hyberg, Fredrik Johansen, Berglind Jonasdottir M.D., Linda Jonsson, Helena Larsson M.D., Ph.D. ^{6,14}, Barbro Lernmark, Ph.D. ^{9,12}, Maria Månsson-Martinez, Maria Markan, Theodosia Massadakis, Jessica Melin¹², Zeliha Mestan, Anita Nilsson, Emma Nilsson, Kobra Rahmati, Anna Rosenquist, Falastin Salami, Monica Sedig Järvirova, Sara Sibthorpe, Birgitta Sjöberg, Ulrica Swartling, Ph.D. ^{9,12}, Erika Trulsson, Carina Törn, Ph.D. ^{3,15}, Anne Wallin, Åsa Wimar¹², Sofie Åberg. Lund University.

Washington Clinical Center: William A. Hagopian, M.D., Ph.D., PI^{1,3,4,5,6,7,11,13,14}, Xiang Yan, M.D., Michael Killian^{6,7,12,13}, Claire Cowen Crouch^{12,14,15}, Kristen M. Hay², 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 Skidmore², Joshua Stabbert, Viktoria Stepitova, Nancy Williams. Pacific Northwest Diabetes Research Institute.

Pennsylvania Satellite Center: Dorothy Becker, M.D., Margaret Franciscus¹², MaryEllen Dalmagro-Elias Smith², Ashi Daftary, M.D., Mary Beth Klein. Children's Hospital of Pittsburgh of UPMC.

Data Coordinating Center: Jeffrey P. Krischer, Ph.D., PI^{1,4,5,10,11}, Michael Abbondandolo, Sarah Austin-Gonzalez, Rasheedah Brown^{12,15}, Brant Burkhardt, Ph.D. ^{5,6}, Martha Butterworth², David Cuthbertson, Christopher Eberhard, Steven Fiske⁹, Veena Gowda, David Hadley, Ph.D. ^{3,13}, Page Lane, Hye-Seung Lee, Ph.D. ^{1,2,13,15}, Shu Liu, Xiang Liu, Ph.D. ^{2,9,12}, Kristian Lynch, Ph.D. ^{5,6,9,15}, Jamie Malloy, Cristina McCarthy^{12,15}, Wendy McLeod^{2,5,6,13,15}, Laura Smith, Ph.D. ^{9,12}, Susan Smith^{12,15}, Roy Tamura, Ph.D. ^{1,2,13}, Ulla

Uusitalo, Ph.D.^{2,15}, Kendra Vehik, Ph.D.^{4,5,6,14,15}, Earnest Washington, Jimin Yang, Ph.D., R.D.^{2,15}. University of South Florida.

Project scientist: Beena Akolkar, Ph.D.^{1,3,4,5, 6,7,10,11}. National Institutes of Diabetes and Digestive and Kidney Diseases.

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

Autoantibody Reference Laboratories: Liping Yu, M.D.^{^5}, Dongmei Miao, M.D.[^], Polly Bingley, M.D., FRCP*⁵, Alistair Williams*, Kyla Chandler*, Saba Rokni*, Anna Long Ph.D.*⁵, 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, M.D., Ph.D., Ewa Lönn Karlsson. Department of Clinical Chemistry, Linköping University Hospital, Linköping, Sweden.

Dietary Biomarkers Laboratory: Iris Erlund, Ph.D.², Irma Salminen, Jouko Sundvall, Jaana Leiviskä, Mari Lehtonen, Ph.D. National Institute for Health and Welfare, Helsinki, Finland.

HbA1c Laboratory: Randie R. Little, Ph.D., Alethea L. Tennill. Diabetes Diagnostic Laboratory, Dept. of Pathology, University of Missouri School of Medicine.

HLA Reference Laboratory: Henry Erlich, Ph.D.³, Teodorica Bugawan, Maria Alejandrino. Department of Human Genetics, Roche Molecular Systems.

Metabolomics Laboratory: Oliver Fiehn, Ph.D., Bill Wikoff, Ph.D., Tobias Kind, Ph.D., Mine Palazoglu, Joyce Wong, Gert Wohlgenuth. UC Davis Metabolomics Center.

Microbiome and Viral Metagenomics Laboratory: Joseph F. Petrosino, Ph.D.⁶. Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine.

OGTT Laboratory: Santica M. Marcovina, Ph.D., 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, Ph.D., PI^{1,3,4,11}, Richard McIndoe, Ph.D., Haitao Liu, M.D., John Nechtman, Yansheng Zhao, Na Jiang, M.D. Jinfiniti Biosciences, LLC.

SNP Laboratory: Stephen S. Rich, Ph.D.³, Wei-Min Chen, Ph.D.³, Suna Onengut-Gumuscu, Ph.D.³, Emily Farber, Rebecca Roche Pickin, Ph.D., Jordan Davis, Dan Gallo. Center for Public Health Genomics, University of Virginia.

Committees:

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

Reference

1. Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. *Journal of clinical epidemiology*. 1999;52(12):1165-72. Epub 1999/12/02.
2. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *American journal of epidemiology*. 1990;131(1):169-76. 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):1350-9.
5. Baker SG. Improving the biomarker pipeline to develop and evaluate cancer screening tests. *Journal of the National Cancer Institute*. 2009;101(16):1116-9. Epub 2009/07/04.
6. Marshall E. Getting the noise out of gene arrays. *Science*. 2004;306(5696):630-1. Epub 2004/10/23.
7. Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*. 2005;14(8):1899-907. Epub 2005/08/17.
8. Wacholder S. Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*. 1991;2(2):155-8. 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. *Journal of clinical epidemiology*. 2007;60(11):1127-31. Epub 2007/10/17.

11. The Environmental Determinants of Diabetes in the Young (TEDDY) study: study design. *Pediatr Diabetes*. 2007;8(5):286-98. 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, Eisenbarth GS, Bingley PJ, Marcovina SM, et al. Harmonization of glutamic acid decarboxylase and islet antigen-2 autoantibody assays for national institute of diabetes and digestive and kidney diseases consortia. *The Journal of clinical endocrinology and metabolism*. 2010;95(7):3360-7. 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 research and clinical practice*. 1999;44(1):21-6. Epub 1999/07/22.
15. Vehik K, Fiske SW, Logan CA, Agardh D, Cilio CM, Hagopian W, et al. Methods, quality control and specimen management in an international multi-center investigation of type 1 diabetes: TEDDY. *Diabetes/metabolism research and reviews*. 2013. Epub 2013/05/16.
16. Rothman K. *Modern Epidemiology*. Boston: Little, Brown and Company; 1986.
17. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. III. Design options. *American journal of epidemiology*. 1992;135(9):1042-50. Epub 1992/05/01.
18. Rundle A. *Environmental Health Sciences* New York: Mailman School of Public Health; 2000. p.152 p.
19. Schulte P, Perera, F. *Molecular epidemiology: principles and practices*. San Diego (CA): Academic Press; 1993.

20. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews Genetics*. 2010;11(10):733-9. Epub 2010/09/15.
21. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*. 2000;97(18):10101-6. Epub 2000/08/30.
22. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of systematic microarray data biases. *Bioinformatics*. 2004;20(1):105-14. 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):118-27. 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):2127-39. 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):2509-23. Epub 2007/09/22.

Percent of control samples collected when corresponding case samples were collected for analyte (Average % across all analytes and the number of controls)

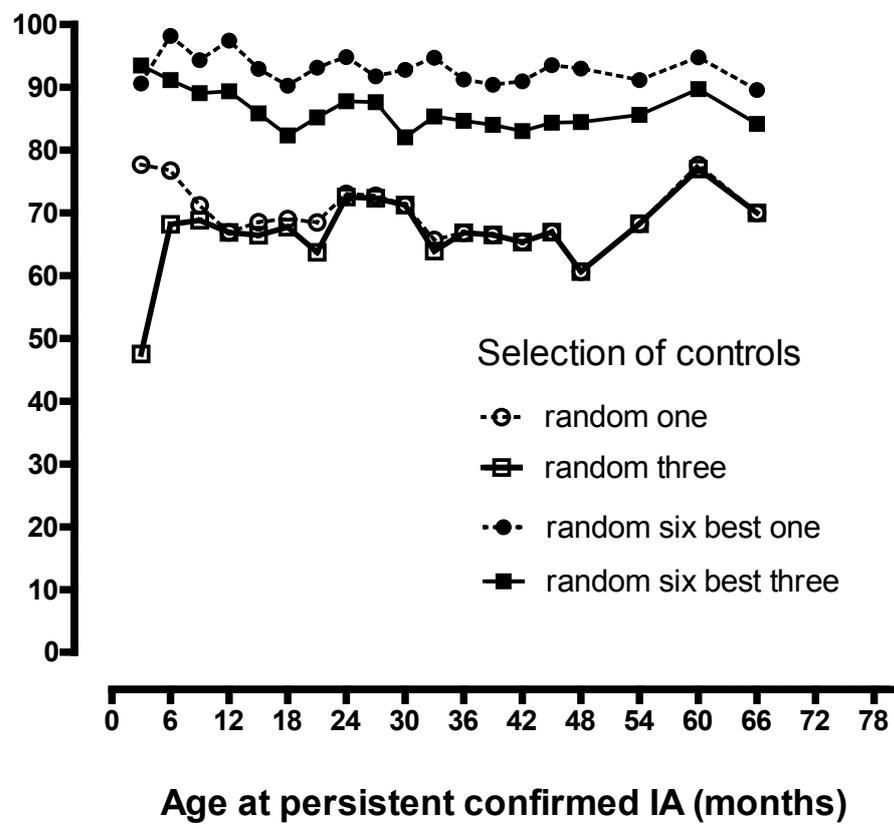


Figure 1. Average number of control samples collected

Dietary biomarker laboratory and metabolomics laboratory

Step1: Start from the event time with least number of potential controls, after matched on clinical center, gender and family history of T1D (say time T_1)

Step2: Randomly select 6 controls and then select a control based on the best sample availability from those 6 until 3 controls are selected, without replacement.

Step3: At the time T_2 with the next least number of potential controls, include all subjects meeting the above matching criteria, with replacement of the controls selected at the time T_1 , if they meet the eligibility criteria.

Step4: Randomly select 6 controls and then select a control based on the best sample availability from those 6 until 3 controls who were NOT selected at the time T_1 are selected, without replacement.

Gene expression laboratory and metagenomics laboratories

After the step4, 1 control was randomly selected from those 3 selected controls.

Figure 2. Control selection procedure

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

¹Human plasma commercially received from Rockland Immunochemicals in 4 independent lots.

²Human stool received from 2 non-T1D independent donors.

³Following viruses were spiked into donor plasma at the documented concentrations: Poliovirus (PV 200 PFU/ml), Rotavirus (20000 PFU/ml), Vesicular Stomatitis Virus (200 PFU/ml) and Adenovirus (20000 PFU/ml).

Table 2. Study subject characteristics: Mean (SD) or N (%)

		Persistent confirmed IA		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%)

* There was 1 case with only 2 controls available

Table 3. Number of samples to be processed per analyte

Laboratory	Analyte	Persistent confirmed IA			T1D			Combined			Number of sets to provide samples
		Case + Control	Control	Case	Case + Control	Control	Case	Case + Control	Control	Case	
Dietary biomarkers	Ascorbic Acid	3060	2259	801	1039	760	279	3736	2779	957	253
	Vitamin D (3 or 9 months)	2214	1604	610	619	447	172	2468	1825	643	246
	Alpha/Gamma-tocopherol and Vitamin D (Other than 3 and 9 months)	3069	2268	801	1029	755	274	3736	2780	956	253
	Fatty Acid	4106	3004	1102	1346	979	367	4889	3621	1268	266
Metabolomics	Plasma	9394	6877	2517	3291	2372	919	11571	8486	3085	275
Microbiome/Viral metagenomics	Stool	10446	4593	5853	3821	1728	2093	13073	6033	7040	399
	Plasma	4948	2372	2576	1799	864	935	6230	3078	3152	404
Gene expression	mRNA	4004	1808	2196	1484	680	804	5080	2369	2711	398

Table 4. Number of samples per batch after sample allocation

Laboratory	Analyte	Number of places allowed per batch		Number of case-control samples per set in one batch: Median (Q1,Q3)		Number of external QC samples (Number of batches)	Number of samples per shipment
		Case-Controls	External QC	Including all TEDDY visits	At a given TEDDY visit		
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/Vial 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