

**Global Collaborative
For Changing Diabetes in
Children**

**Global Cohort Study for
Type 1 Diabetes**

**Using Data to Create the
Largest Cohort Study of
Individuals with Type 1
Diabetes Worldwide**

HEALTH SYSTEMS INNOVATION LAB
AT HARVARD UNIVERSITY



HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

HEALTH SYSTEMS
INNOVATION LAB

This page is intentionally left blank

About this Report

a. Purpose of report.

This report will outline the main steps needed to set up, and challenges to anticipate, in the development of a virtual global cohort of individuals with Type 1 diabetes (T1D). This virtual cohort study design should not be limited to those with T1D and can be developed and set up for any chronic medical condition. This report can be used to form the building blocks of any future virtual global cohort study.

b. Overview of the cohort study.

We set out to create the largest cohort study of children and adolescents with T1D. To do this, we advised on, and helped facilitate, the introduction of a data collection system in six ‘phase one’ countries across Africa and South Asia. This data system will subsequently be used to pool data and create country level registries comprising adolescents (≤ 20 years old at the time of first data point) with T1D. Data from these registries will then be modified by each country to make it suitable for international sharing with collaborators for analysis.

c. Benefits of a ‘virtual’ cohort.

A conventional cohort study requires significant on-the-ground resources to enroll vast numbers of patients. This can be time-intensive, costly and may not even capture a truly representative patient population. The benefit of virtual cohort formation is that a wide and diverse group of patients can be enrolled from a broader geographic area with lower costs and resourcing needs. Primary researchers are not required to be on-the-ground at all research sites, or even necessarily in the country where data is being collected. The patient enrollment and data collection process can be led by country experts and practitioners familiar with the context in which the T1D patient population is situated.

Authors:**Jake Figi**

Postdoctoral Research Fellow, Health Systems Innovation Lab, Harvard University

Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Harvard University

Che L. Reddy

Associate Director, Innovation and Translation, Health Systems Innovation Lab, Harvard University

Research Associate, Global Health Systems, Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Harvard University

Rifat Atun

Professor of Global Health Systems, Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Harvard University

Director, Health Systems Innovation Lab, Harvard University

Suggested citation:

J. T. Figi, C. L. Reddy, R. Atun. *Global Collaborative for Changing Diabetes in Children: Using Data to Create the Largest Cohort Study of Individuals with Type 1 Diabetes Worldwide*. Health Systems Innovation Lab, Harvard University, August 2024.

Table of Contents

Title Page.....	
About this Report.....	2
Authors and Citation.....	3
Executive Summary.....	5
1.....	Introduction
.....	6
2.....	Data Assets Available in CDiC Countries for the Study of T1D
.....	8
3.....	Data Acquisition and Management for T1D Research
.....	14
4.....	Using Data to Conduct T1D Research
.....	20
5.....	Creating an Enabling Environment for a CDiC Research Data Pipeline
.....	25
6.....	References
.....	28
7.....	Appendix
.....	32

Executive Summary

This report presents the foundational steps and anticipated challenges in establishing a virtual global cohort study focused on individuals with T1D. Although the initial study targets T1D, the design and methodologies outlined here are adaptable to any chronic medical condition, providing a blueprint for future virtual cohort studies.

The goal of this initiative is to create the largest cohort study of children and adolescents with T1D by implementing a data collection system across six Phase 1 countries in Africa and South Asia. This system will facilitate the formation of national registries of adolescents with T1D, which will then be adapted for international data sharing and collaborative analysis.

Leveraging a virtual cohort design allows for the inclusion of a diverse patient population from a wide geographic area with reduced costs and resource demands, enabling contextually relevant data collection led by local experts. The cohort seeks to address the gap in the diagnosis and management of T1D in low- and middle-income countries (LMICs), where many cases remain undiagnosed and untreated.

In order to form the cohort, a data collection system was developed by Dure Technologies, comprising electronic health records (EHRs) at the clinic level which feed into national registries. This system enables the further pooling of data from the registries into the global cohort for in-depth analysis led by the Harvard research team in conjunction with key clinicians from each research country. Four high-priority research questions were identified for the first phase analysis in the cohort study, focusing on disease incidence and prevalence, factors influencing disease control, mortality rates and causes, and demographic influences on disease progression.

Throughout the study, compliance with international data protection laws and institutional agreements is critical. The report outlines data use agreements (DUAs) and data security measures to protect personal information and maintain data integrity.

This report sets the stage for a groundbreaking approach to studying T1D and other chronic conditions through virtual cohort studies. By leveraging technology and international collaboration, it aims to enhance understanding, optimize health systems, and improve health outcomes for individuals with T1D worldwide.

1. Introduction

In 2021, there were an estimated 360,000 new cases of T1D among children and adolescents globally. This number is expected to continue to increase, reaching as high as 500,000 new cases of T1D in those under 20 by 2050.¹ Scientific advancements have substantially improved health outcomes for individuals with T1D, who depend on exogenous insulin to manage blood glucose levels, prevent severe complications and ensure long-term survival. Despite these breakthroughs, many individuals in low- and middle-income countries (LMICs) and certain high-income countries (HICs) face inadequate access to insulin. High-value health services are needed to improve access to insulin and ensure optimal health outcomes. This deficiency in insulin access and high-value health servicesⁱ for T1D at primary and secondary levels of the healthcare system places individuals at risk of premature disability and death.²

In a recent study conducted by the GC-CDiC, an estimated one in two cases of T1D in adults and two out of three cases in children and adolescents across West Africa, South and Southeastern Asia, and Melanesia go undiagnosed, highlighting a significant failure of country health systems.³ In addition to insulin, comprehensive health services at primary and secondary care levels are essential for effectively managing T1D in children and adolescents, ensuring continuity of care, preventing associated life-threatening complications such as Diabetic Ketoacidosis (DKA) and improving the life chances for children and adolescents with T1D.

A substantial gap persists in T1D healthcare services across most LMICs due to a lack of critical data needed to understand health system performance in relation to T1D, understand variation in care delivery and outcomes, and develop targeted solutions for improving access to care.⁴ The substantial underdiagnoses of T1D in LMICs has important implications for human capital, affecting families and households, and hindering these nations' ability to harness demographic advantages and advance economically. A staggering 50% of individuals are not diagnosed by health systems.⁵ There is a major opportunity in both LMICs and certain HICs to optimize data utilization within health systems to better understand the management of individuals with T1D

ⁱ A high-value health service, for instance those relating to diagnosis and management, is defined as a health service provided for T1D that is delivered effectively, efficiently, equitably and responsively.

across care pathways and to optimize provision of health services and innovations to enable the delivery of high-value health services for T1D.

This research will build on earlier collaborative activities conducted by the GC-CDiC. Initiated in 2009, CDiC addresses T1D care inequities in LMICs by providing insulin and necessary health resources through designated centers. According to the 2022 report, the program has benefited over 41,000 children across 26 countries through more than 360 clinics, distributing over 3.8 million vials of insulin.⁶ In 2021, CDiC and Harvard established the Global Collaborative for Changing Diabetes in Children (GC-CDiC). Consultation with leading clinicians in 16 CDiC collaborating countries led to the co-development of the four interrelated streams of work comprising GC-CDiC: (I) research, (II) data systems, (III) innovation, and (IV) translation. The design and implementation of a cohort study for children and adolescents with T1D was identified as a major priority within the GC-CDiC to build on prior collaborative work. CDiC has continued to expand and is now comprised of 30 partner countries as of 2024 (Figure 1).



Figure 1: CDiC Partnership Countries⁷

The cohort study was a specific output of these consultative discussions and will advance all four workstreams. It will not only support the mission of GC-CDiC, to improve T1D health systems and services, but will contribute to fundamental knowledge about the etiology of T1D, its

associations, disease progression, optimal management of care, how best health systems should adapt and harness policies, and innovations to improve outcomes for patients with T1D at the population level for large-scale impact. It will respond to a critical clinical and therapeutic observation raised by nearly all clinicians: there is great variation in presentation, pathogenesis, response to therapy, and outcomes among patients with T1D.

2. Data Assets Available in CDiC Countries for the Study of T1D

a. Characterizing the data.

1) Source

Data for clinical research of this type can be either publicly available data or private patient data. Publicly available data is easier to source and can come from a country's government or various global research partners. Private data generally will need to be sourced from providers of care to individuals with T1D. Both sources have their benefits, in particular around granularity of data, ease of obtainment and data security.

Publicly accessible data within the government will commonly be sourced from the Ministry of Health or the Ministry of Finance in the form of nationally representative reports. Research partners include institutions such as the WHO, USAID, and university affiliated or independent researchers and can include reports and peer-reviewed research publications. This data will present a broad image of health in a country or region, be easy to obtain if published on government or research websites and require no additional data security processes.

There is limited publicly available data which can be used to study T1D generally. There is a particular lack of information in CDiC countries where even current estimates of incidence and prevalence are of uncertain accuracy. One possible source of information to guide health systems research is through the national health accountsⁱⁱ of each country; however, the majority of these will provide a very limited insight into T1D-specific funding and expenditure.

ⁱⁱ National health accounts (NHAs) are statistical reference manuals which are standardized across countries to provide information on national health care expenses and funding. While some contain disease-specific financial data, this is not commonly so granular that T1D data is provided.

Private patient data must be sourced directly from healthcare providers and, depending on the level of data, may require patient consent. This can be a much more difficult process as data at this level is not commonly collected for research purposes. With continuously improving electronic health records (EHR) systems, this may be able to be processed relatively quickly; however, EHRs are not the standard of practice in CDiC countries, with most clinics relying on paper notes. Additionally, if personal identifiable information is collected it must be collected, stored and transferred securely to protect patient privacy. While more difficult and time consuming, sourcing data in this manner will provide much more granular information with wider research uses.

2) Level and type.

Level	Type	Location
Clinic (primary)	Data on individuals in the form of Health Records (most commonly paper but may be electronic or mixed); aggregate reports on clinic and patient outcomes	Paper health records commonly stored on-site; EHR either on a local server, cloud-based or mixed; reports stored locally or shared publicly on internet
Hospital (secondary and higher)	Data on individuals in the form of Health Records (most commonly paper but may be electronic or mixed); aggregate reports on clinic and patient outcomes	Paper health records commonly stored on-site; EHR either on a local server, cloud-based or mixed; reports stored locally or shared publicly on internet
National or regional (in large countries)	Aggregate in the form of registries, reports and studies	Stored locally as publicly available on-request or available online

Table 1: Levels and types of data – where data is and in what form.

b. Defining the data needed to conduct prioritized research.

Development of the variable list for data collection for the national registries was a fundamental step and required careful consideration such that useful outcomes and results could be produced. This process involved conducting a scoping review of global T1D registries, mapping of data collected in CDiC countries, data variable grouping and organization, priority research question determination, and finally mapping of variables to the research questions, after which the variable set was finalized.

In the scoping review of global T1D registries, it was found that 81.4% (n=114) of registries identified were based in high income countries (HICs), with only 15.7% based in upper-middle-

low-middle- and low-income countries.⁸ From all the registries identified, only seven were within a CDiC country: Rwanda, Uganda, India (x2), Sudan, Tunisia and Malaysia. Of these, only Uganda's is known to be currently active with registries in Sudan, Tunisia and both in India closed and the status of those in Rwanda and Malaysia unknown. This work provided a clear need for registry formation in the CDiC countries.

Following this, a mapping exercise was conducted to outline the data variables that were collected in CDiC countries at the clinic level. These variables were seen as the possible data points – and were organized into four modules in relation to the ease and frequency of collection.

The modules were termed Core, Core+, Core++ and Core+++. These ranged from the simplest variables in Core, which would be collected by nearly all clinics, up to expensive and less-commonly ordered tests that only advanced centers may perform in Core+++. These variables were additionally compared against the SWEET registry ('Better control in Pediatric and Adolescent diabetes: Working to Create Centers of Reference') – an international T1D registry used in clinics across LMICs and HICs – to identify any potential missing variables for inclusion.

The modules were established in this progressive manner, from Core to Core+++, so they could be easily and appropriately implemented in a variety of clinical settings. Clinics which had fewer testing facilities or available staff for data input could be started on the earlier modules and not have the added burden of unnecessary variables which would not be able to be completed.

During these preparations, the prioritized research questions were developed. These were determined first through detailed consultation with key opinion leaders (KOLs) who are experienced clinicians actively working with T1D patients in each country from CDiC countries to form a list of 35 important health challenges and questions around T1D, ranging from epidemiological questions to questions around the impact of T1D on pregnancy and maternal health (Appendix I, pp 32-33). The research questions were subsequently ranked in importance by the KOLs as 'high', 'medium' or 'low' priority and compared against the data variables due to be collected.

For the first phase of the cohort study, four high-priority research questions were identified. These were the research questions that, on average, ranked the highest among the KOLs, and mapped to the most feasible data variables (Table 2). All research questions were mapped similarly. These were then presented back to the KOLs for re-discussion and confirmation of their priority.

Research Questions:	Core	Core+	Core++	Core+++
What is the current incidence and prevalence of Type 1 Diabetes within the specified population?				
What factors are associated with achieving and maintaining favorable control of Type 1 Diabetes within the study population?				
What are the mortality rates and causes of death among individuals with T1D over long-term follow-up periods?				
How do different demographic groups or treatment approaches influence disease progression?				

Table 2: Research questions for phase one of GC-CDiC with corresponding core module requirement. Dark green represents fully able to answer the question at this level of data collection, light green is mostly able to answer the research question, yellow is able to answer with some flaws and red is unable to answer.

Finally, the modules were reassessed alongside the research question mapping. Select variables that were deemed to be essential for the prioritized research questions were shifted towards the ‘Core’ module while those of lower impact were shifted towards the ‘Core+++’ module, in part to reduce the number of variables required to be collected within the first two modules (final core modules are available in Appendix II, p34).

c. Data gaps.

A principle behind this virtual cohort is to cause minimal change to current clinical practice during the study. This includes the minimization of variation from normal data collection such as laboratory testing, genetic testing, mental health screening, quality of life screening, or any other patient-level investigation which may not have otherwise taken place. As a result, data gaps are expected to be present. Some of these specific gaps are anticipated – such as scores from quality-of-life assessments or other detailed questionnaires. As such, these are not built into the core modules for collection. Of the original 35 health challenges proposed by the KOLs, nine of them are not possible with the current structure of the virtual cohort. These are largely related to behavior, diet, quality of life and maternal outcomes (which are not followed up in many of the diabetes-specific clinics).

d. Data collection system to enable international research.

Harvard collaborated with a technology company, Dure Technologies, who would implement a co-designed data collection system for use in all CDiC countries, consisting of:

- i. Electronic Health Records (EHRs) at the clinic level**
- ii. Registries at the national level**
- iii. A mechanism for data transfer for cohort assembly and research**

These key features will work together, pooling data to form the global cohort. Clinics in each country will input data for the core modules into their EHRs. This data will then be transferred within each country to their national registry. KOLs will have a functionality within the country registry to automatically modify the data such that personal identifiable information is removed. The data can then be exported and shared by the KOLs in each country with the research team at Harvard to form the global cohort. The collated data will then be analysed by the Harvard research team with input by the KOLs for publication. These steps are summarized in Figure 2.

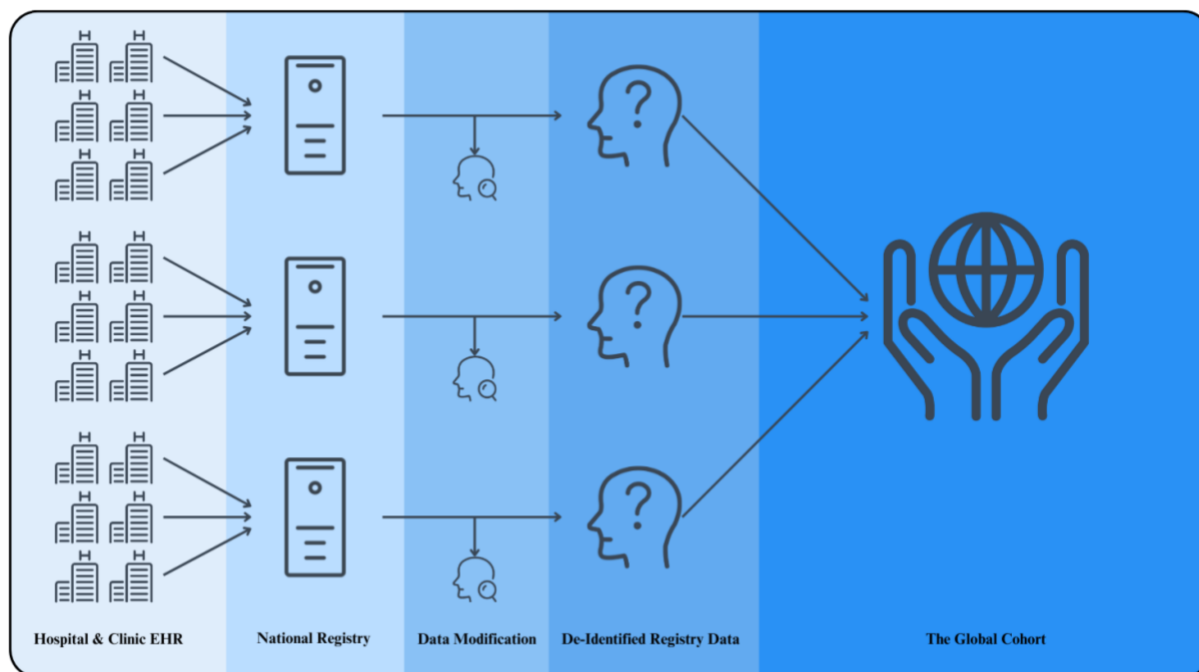


Figure 2: Data collection and transfer processes. Data is collected in hospitals and clinics via EHRs and transferred to each country's national registry. Before transfer to form the global cohort, data is modified to remove or alter personal identifiable data. Data is finally pooled at Harvard, forming the global cohort.

3. Data Acquisition and Management for T1D Research

a. Data governance and stewardship.

The collection, storage, sharing and use of data is influenced by each country's laws and regulations, and policies and procedures at the facility level. Compliance with these policies and regulations is crucial to data acquisition and management for T1D cohort study research.

1) Laws, regulations, and policies governing the use of data at a global and national level.

While laws on data protection and usage vary from country to country, the principles are similar. Generally, laws and policies cover the need for fair and lawful usage, transparency, minimization of risk via limited collection and storage time, accuracy of data, confidentiality and accountability. Additionally, these laws mainly only apply to *personal* data; this is data which identifies or could identify an individual. An indicative list of some international data protection laws for 'wave 1 countries' is summarized in Table 3.

Data Law	Description
Health Insurance Portability and Accountability Act of 1996: HIPAA (USA)	Healthcare specific and only applies to 'covered entities' (healthcare providers, health plans and healthcare clearinghouses). Outlines their definition of protected personal information and the steps to de-identify data. ⁹
General Data Protection Regulation: GDPR (European Union)	Only applies personal data* of EU citizens and residents. Data usage must be fair and transparent, used for only legitimate purposes, only the minimum possible data may be collected that is needed for a specific purpose, it must be kept for the shortest time possible, and it must be stored securely. ¹⁰
The Information Technology Act, 2000 (India)	Recognizes that informational privacy is a facet of privacy. Personally sensitive data or information can only be collected with consent. Data may only be stored in servers in India. Does not apply to anonymized or de-identified data. ¹¹
Personal Data Protection Act, 2010: PDPA (Malaysia)	Personal data* may only be processed if the participant has given their explicit consent. Personal data can only be transferred outside of Malaysia if specified by the Minister in charge of data protection. Does not apply if data cannot identify the individual it is about. ¹²

African Union Convention on Cyber Security and Personal Data Protection (27 African countries including Cameroon and Guinea ¹³)	Personal data* may only be processed after authorization by the national protection authority. Data subject must have given their consent. Largely aligned with GDPR principles. Rules do not apply fully if data is de-identified or anonymized. ¹⁴
Constitution of the Federal Democratic Republic of Ethiopia, 1995 (Ethiopia)	The country's constitution outlines the right of all people to privacy. It also describes that privacy may be restricted for the protection of health. ¹⁵ In April 2024, Ethiopia passed the Personal Data Protection Proclamation (PDPP), which will bring data laws in line with other international standards. Personal data* will be required to be stored securely, lawfully processed and may not be transferred to countries without adequate data protection standards. Additionally, there will be limitations on personal data* collection and use. ¹⁶

Table 3: International data laws and key details from each. *Refers only to information relating to an identified or identifiable person

2) Data use and data security agreements at the institutional level.

Data use agreements (DUAs) outline the terms and limitations on data usage by research partners and are fundamental in the establishment of a virtual cohort. The DUA is a contract which allows for the transfer of data from a data 'provider' to a 'recipient'. The DUA established for this study outlines the rights of 3rd parties, such as Harvard, involved in the GC-CDiC to engage country research teams to use the data for the prioritized research purposes as well as outlining the limits on data usage. An overview of the key principles used in the DUA between Harvard and the data providers is provided in Table 4.

(1) Research publication	States that Harvard and Harvard researchers are free to publish results from the use of the data as they see fit, provided the data provider is recognized for their contribution as the source of data in any public disclosures of research outputs resulting from the data.
(2) Data security	States that appropriate safeguards must be put in place by Harvard to prevent unauthorized use or distribution of the dataset, such as limiting access to only Harvard researchers who require access to the data for the purpose of conducting research for the GC-CDiC and by ensuring the data is not shared with any person outside of Harvard.

(3) Data Contents	Describes the data which should be shared in the data set. This includes the broad categories of patient history, clinical evaluation, disease management and disease outcome.
(4) Multiple Transmission	States that data may be shared in multiple instances (by facsimile or, more likely, electronic transmission) as is necessary.
(5) Use limits of data	Clarifies that only Harvard researchers may access the data, it may only be kept for five years (unless a new DUA is agreed upon) and it may not be attempted to be used to diagnose or treat individual human subjects.
(6) Other	Outlines steps to take in the case of a data breach, including informing the provider. Clarifies the right to termination of the contract by the data provider (30-day's notice) and Harvard (notification and data return/destruction).

Table 4: Overview of key features in the GC-CDiC DUA between data providers and Harvard

Data security agreements outline the required protections of confidential information and may also include stipulations around data access, storage, processing, transfer and confidentiality – based on the data security level. These are also legally binding contracts, with breaches potentially leading to fines or criminal sanctions for some types of data or agreements. These are generally only required when restricted data is shared. In the case of GC-CDiC, stipulations around data security are outlined in the DUA and will not require a separate document.

Additionally, throughout this process, data ownership, responsibility and rights must be transparent and clearly outlined. In this study, data will be ownedⁱⁱⁱ entirely by the countries (or a party within the country) in which the data was originally collected. Research outputs from Harvard will be presented as a collaborative output between the involved GC-CDiC collaborators and the Harvard research team.

b. Data storage

Data security is of the utmost importance, particularly at the country level where data will be identifiable and in its most complete form. Any data breach could have potentially significant repercussions for patients and researchers and must be prevented via appropriate data storage and maintenance.

ⁱⁱⁱ A data owner (a person or institution) has legal control and responsibility over a dataset. They ultimately determine who can use their data and for what purposes. They are also responsible for the safe keeping of their data.

1) Country level storage.

There are two main methods for data storage which may be utilized: localized physical storage or cloud storage. Local physical storage requires the installation of a data server to which data would be uploaded. This option requires a dedicated and safe location for the server, basic utilities to ensure continuous uptime and an individual and staff responsible for the troubleshooting, maintenance and security of the server. A comprehensive comparison of local vs. cloud storage is provided in Table 5.

Cloud storage utilizes existing storage options for purchase/subscription (for example from major cloud service providers such as Microsoft, Amazon, among others) and may cost from \$0.001 to \$0.15 per GB monthly depending on user location, volume and retrieval/input frequency.¹⁷⁻¹⁸ For a dataset that is regularly updated, this cost will likely be in the \$0.018-0.03 per GB monthly range. At later stages of the cohort study, it may be possible to have intermittent uploads of data or store a portion of data in ‘deep archive access’ which could lower the price range towards <\$0.01 per GB monthly. Anticipating these prices is crucial to promote sustainability of the cohort study.

	Local Server	Cloud Storage
Storage	Storage limited by server size – modular upgrades may be possible	Storage limited by cost but can otherwise be increased as needed
Access	Local offline access and online access when server is connected to the internet	Offline access available if data is downloaded, online access available globally with internet connection
Safety	Data safety run by local team – varies by expertise	Data safety run by expert host – may be more secure
Maintenance	Maintenance and upkeep required by local team	Maintained by Cloud service provider
Data loss	Greater risk of data loss or destruction	Data backed up across multiple servers in different world regions
Costs	High up-front set up costs	Low up-front set up costs, longer operational costs?
Upload	Data can be uploaded locally without internet access	Requires internet access for upload – but can be saved to a local device and uploaded when internet available

Table 5: Comparison of local serve and cloud storage for data storage.

2) International level storage.

Data will be sent from each participating country (the data provider) to Harvard (the data recipient) as de-identified data. Data will not be shared with any non-Harvard party. This includes any industry, sponsors, the technology partner and other participating country research teams. As such, data outside of the country level will only be stored by the Harvard team as stipulated by the DUA.

All data used at Harvard is classified into the University's data security levels (DSL), from DSL1 (lowest risk, public information) to DSL5 (extremely sensitive data) (Figure 3). Given the de-identified nature of the GC-CDiC data, it is classified as level 2, low-risk.¹⁹

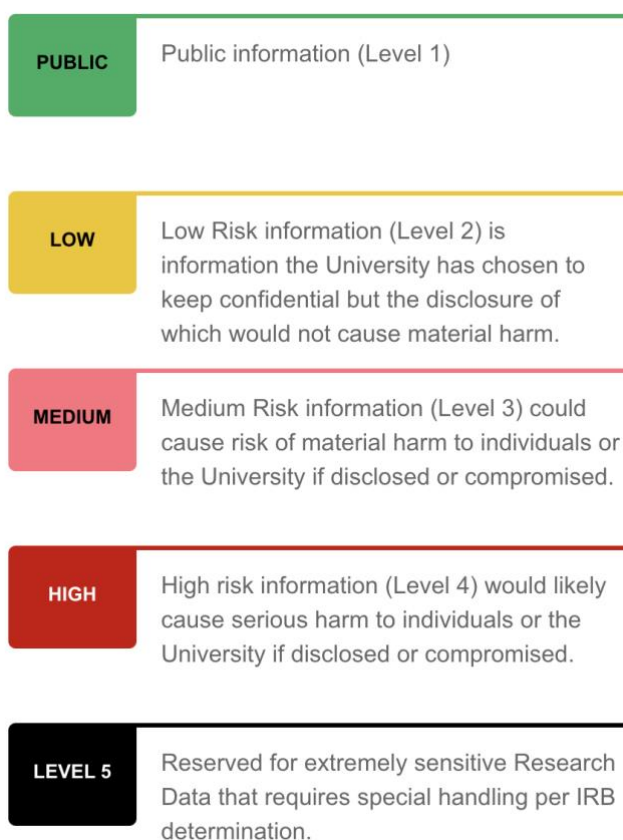


Figure 3: Harvard data risk classifications²⁰

At this level of data security, data will be stored on a Harvard-managed Dropbox, provided there are no public repositories.²¹ Other requirements for this level of data security include the need for accessing computers to meet Harvard security requirements and to only share data with authorized users on the research team included in the DUA.

c. Mechanisms to pool data at national and global levels using Cloud technology.

There are two methods to pool de-identified data from a national registry to then share data on Harvard's secured File Transfer Protocol (FTP – software that allows secure exchange of files over a network) server and domain: 1) manual data upload, 2) real-time data transfer. This server would host all data from the wave 1 countries and future countries and can be accessed for research purposes. The two methods to upload/share de-identified data with Harvard are as follows:

1) Manual data upload.

i. Registry data download:

- The data provider logs into the country registry and selects 'Patient Record List', where the complete list of T1D patients and their visits is available.
- The data provider selects 'Download De-Identified Data' to save in the data on their personal system.

ii. Upload process:

- The data provider logs into the secure Harvard FTP server portal and uploads the previously downloaded data files to the Harvard portal.

iii. Notification and confirmation:

- Upon successful upload, the system will confirm receipt of the data.
- E-mail notifications will be sent to Harvard and the data provider to confirm that the data has been successfully uploaded.

2) Real-time data transfer from national registry.

This outlines the architecture and workflow for exporting de-identified data from a PostgreSQL database (an open-source database management system) via a Representational State Transfer Application Programming Interface (REST API – a way to create web services that exchange data between client and server applications), formatting the data into an appropriate file format, and uploading it to an Amazon S3 bucket (a secure online data storage service). The process aims to ensure secure, efficient, and automated handling of the data export and upload process.

i. Data architecture workflow (Figure 4):

1. Data Retrieval: The application fetches de-identified data from a PostgreSQL database using a REST API endpoint.
2. Data Transformation: The retrieved data is formatted into a file format suitable for storage and further processing (e.g., JSON, Excel, CSV).
3. Data Upload: The formatted file is uploaded to an Amazon S3 bucket for storage and accessibility.

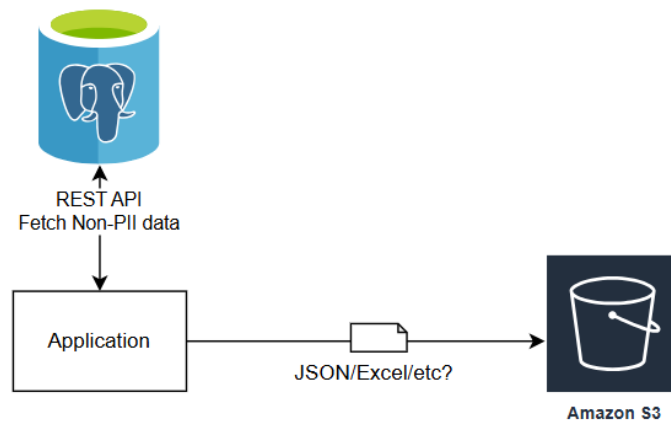


Figure 4: Data architecture workflow.

ii. Functional flow of data transfer:

- The data provider logs into the country registry and clicks on ‘Patient Record List’, where the complete list of T1D patients and their visits is available.
- The data provider selects the time-period from the menu and clicks on “Share de-identified data with Harvard”.
- E-mail notifications will be sent to Harvard and the data provider to confirm that the data has been successfully uploaded.

3) General considerations.

- i. File formats to use: JSON, Excel, or CSV.
- ii. Data Security: Although the data is de-identified, encrypting the file before uploading to the S3 bucket is recommended to add a further layer of security.

- iii. Frequency of uploads - defining the upload schedule based on data requirements:
 - Daily uploads: for time-sensitive or frequently updated data.
 - Monthly uploads: for less dynamic or periodic data updates.
- iv. S3 Bucket Configuration:
 - Ensure the bucket has the correct settings in place to allow uploads while maintaining security.
 - Use Identity and Access Management (IAM) roles to protect and limit access to the S3 bucket.
 - Set up lifecycle policies to archive or delete older files automatically.

4. Using Data to Conduct T1D Research

a. Identifier definitions

Direct identifiers are ones which reference a person, such as a full name or ID number. Given their nature, these are easier to identify for removal from a dataset. In recent years, attention has been drawn to indirect identifiers, in particular, after a researcher proved that research efforts were not sufficient enough to protect people's privacy.^{22,iv} Indirect identifiers are ones which can be used in combination with publicly available data or other indirect identifiers, such as eye color, gender and age, to identify a person. These can be any type of data point, with some less obvious than others, and can be more difficult to pick out for removal or modification. Their status as an indirect identifier depends on the context of the data set, including who it describes (i.e. in a dataset of only US Senators, age would be considered an indirect identifier) and what data variables are collected.

b. De-identification and anonymization.

There are numerous ways to protect the identification of individuals participating in research. De-identified data is data that has had all direct and indirect identifiers removed or manipulated such that they cannot be linked to the real world. For example, this would include removing a patient's name or changing a patient's address to only a city, region or state. Anonymized data is a higher

^{iv} For example, in 1977 when Dr. Latanya Sweeney, Harvard Professor of the Practice of Government and Technology, reidentified medical records which had been de-identified in line with the HIPAA regulations at the time

level of security than de-identified data. For data to be anonymized, it must not be able to be linked back to the person for which it pertains by anyone, including the researchers, and there must not be a way that any participant can be identified to have even taken part in the research.

c. De-linking.

Research data is often linked or ‘coded’ to a research identification number or code and can be a highly useful technique to allow longitudinal analyses of data from a set of patients over time. Linking of data is conducted by assigning each participant a code, which can be produced randomly or with a particular coding method (e.g. a combination of a participant’s social security number, 0123-45-678, month of birth, February, and year of birth, 1979, forms the code ‘0123feb79’). The formed codes are included alongside participants’ data in the main dataset. In a separate document, the ‘master list’ or ‘code list’, codes are linked with identifying information (e.g. 0123feb79 = John Smith). This code list is then stored separately and securely from the dataset (in some cases in a physical format), ideally accessible by only one or two individuals.

While a research ID/code is not a direct identifier, as it is only held by the research team, it can lead to significant data breaches if the method of coding or code list is leaked. Additionally, linked data often falls under different data regulations from de-identified data.

De-linking (or de-coding) of data is a straightforward process. It can be done by either deleting (or destroying if a physical copy) the code list or by removing the research ID/code variable for each participant. If a specific method of code formation was used, removal of the code variable alone is not sufficient, as those with knowledge of the code makeup could recreate each code, and the code list will also need to be destroyed.

Within GC-CDiC, data will not be coded at any level, so will not require additional de-linking measures.

d. De-identification techniques.

In order to conduct research that is not classified as human subject research in the US, all data must be fully de-identified, however it is not required to be anonymized.²³ The process of de-

identifying data can be conducted in two manners according to the US Department of Health and Human Services (DHHS) to comply with the Health Insurance Portability and Accountability Act (HIPAA). The first is the ‘safe harbor’ method, by which all HIPAA personal identifiers are fully removed from the data (Table 6). This is the safest and most simple method for data anonymization. However, this method may result in the loss of data required for analysis within the planned research. The alternate method is via ‘expert determination’. In this, a person with “appropriate knowledge of and experience with generally accepted statistical and scientific principles” applies their knowledge and experience to determine that the risk of identification is “very small” if the information is used either alone, or in conjunction with other data, by the intended data recipient to identify the individual that the data regards.²⁴ This ‘very small risk’ will not necessarily be consistent between the same variables in different datasets or from different sources, so this should be reassessed in each instance.

Additionally, these guidelines are targeted at the transferring entity, defined as a healthcare provider (including doctors, clinics, pharmacies and others), a health plan or a health care clearinghouse.²⁵ It is important to note that these guidelines were formed in, and only legally apply to, the US. However, these principles stand up as good ethical practice and may also align with country-specific laws and regulations of international collaborators.

Another important regulation is the general data protection regulation (GDPR), which offers protections to the privacy of EU citizens and was designed as one of the world’s toughest and most comprehensive privacy laws. However, even within this stringent framework, the proposed level of data protection would not constitute a breach of GDPR, regardless of consent status for the data’s use.²⁶⁻²⁷

Identifier Domain	Collected?
1. Names	Yes
2. Geographic location up to state level	Yes
3. Dates, except year (all dates >89 years old)	Yes
4. Telephone (and fax) numbers	Yes
5. Vehicle ID’s and serial numbers	No

6. Fax numbers	No
7. Device ID's	No
8. Email addresses	No
9. URLs	No
10. Social security or national ID numbers	Yes
11. Medical record numbers	Yes
12. Biometric identifiers	No
13. Health plan numbers	No
14. Full face photographs	No
15. Account numbers	No
16. Any other unique ID numbers	No
17. Certificate or license numbers	No

Table 6: HIPAA Individually Identifiable Health Information (IIHI) and if they are collected in the national registries which will provide the data for the GC-CDiC research

e. Data transformation required to ensure data de-identification.

These processes must be conducted at the country level, before being transferred outside the control of the responsible officer. Any external transfer of data from the country of origin without prior de-identification would be considered a serious breach of data protection in terms of the DUA, HIPAA and GDPR. Additionally, any transfer of data within or external to the country to non-authorized recipients, regardless of de-identification status, would constitute the same serious breach.

GC-CDiC will employ the expert determination technique across the HIPAA IIHI identifier domains. This allows for the inclusion of variables deemed to be very low risk (e.g. month of clinic visit) that otherwise would be removed in the blanket measure of the safe harbor method, while still maintaining participant data safety. As outlined in Table 6, the cohort study will collect data across six of the 17 domains. Of the six domains collected from, data from four will be fully removed before data transfer, data from one will be edited to remove the most identifying portion and one will have both editing and data conversion (Table 7).

Identifier Domain	Collected Variable	Risk Mitigation Strategy
1. Names	Patient first and last name	Removed before transfer.
2. Geographic location up to state level	Patient address, clinic address	Clinic address retained if ≥ 50 patients treated at site, otherwise removed before transfer. Clinic address converted to clinic ID for all patients. Patient address converted to urban/rural designation, and region or state retained.
3. Dates, except year (all dates >89 years old)	Date of birth, clinical visit date, date of diagnosis, date of death	Month of birth and death, day of birth and death, day of clinic visit, and day of diagnosis removed before transfer.
4. Medical record numbers	Medical record number	Removed before transfer.
5. Telephone (and fax) numbers	Phone number for patient and/or guardians	Removed before transfer.
6. Social security or national ID numbers	National ID number	Removed before transfer.

Table 7: IIHI data removal or conversion plan

Data within the ‘geographic location up to state level’ domain will require the most modification. All clinics will be given an ID number which will be linked to their metadata. All patients will have this clinic number included in their modified data. Clinics with ≥ 50 T1D patients included in the study will also retain the specific information regarding clinic site. It would not be responsible to collect this detailed data from all clinics, as it is estimated that some clinics may have as few as one patient who will be included in the study (Table 8). Additionally, address of all patients will be edited to remove all but the country and state/region the patient lives in, as well as converting the patient’s address to a ‘rural’ or ‘urban’ designation.

Country	Patients Per Clinic:			
	Mean	Median	Low	High
Cameroon (n=3)	28	30	15	40
Ethiopia (n=10)	66	50	10	235
Guinea (n=9)	61	52	12	160
India (n=35)	56	30	5	270
Malaysia (n=26)	16	10	1	100

Table 8: Estimated patients per clinic in participating CDiC countries from early survey data

f. Data analyses.

Once data is pooled at the country level, de-identified and then shared with Harvard it will be analyzed to answer the research questions. This will be done through a range of statistical modeling; each modality will be best suited to the individual research questions.

This analysis will be conducted on the data across two time points: retrospectively collected data and new data. The retrospective data will be formed of any data added to the national registries that was collected before the initiation of the data collection solution. New data encompasses any data collected after the implementation of the data collection solution. Retrospective data will be shared as soon as it is available, whereas new data will be collected over the course of approximately two years before it is shared for analysis.

5. Creating an Enabling Environment for a CDiC Research Data Pipeline**a. Principles.**

In a large international study such as this one, global cooperation and collaboration is key to success. Effective and regular communication is fundamental in fostering this. Finally, a collective understanding and trust in the longevity and long-term commitment required by all partners is necessary if the cohort study is to be valuable beyond the initial two-year phase. This joint belief in a greater overall benefit through longitudinal research and policy development brings the most potential future benefit.

b. Components.

In the GC-CDiC, we have kept communication channels open via regular email updates, monthly online international alignment meetings and weekly or fortnightly progress calls between local country leads and the Harvard team (described in more detail in Table 9). Additionally, two in-person meetings (at the World Health Assembly and the International Society of Pediatric and Adolescent Diabetes' annual conference) have been arranged during the first half of phase 1 where progress is shared, and collaborative learning can occur. Beyond this, we aim to nurture the longevity of this project through in person executive education courses aimed towards relevant


stakeholders from each country and through online research capacity strengthening mini courses, facilitating future independent research by in-country research teams.

	Participants	Frequency	Purpose
Internal	1) Harvard Research Team	Bi-weekly and as needed	- Project updates - Discuss next steps of study implementation
Country-Level	1) Harvard Research Team 2) Individual KOLs and their research teams	Weekly or Fortnightly and as needed	- To gather regular feedback from KOLs on study design and implementation - To provide logistical support to KOLs for IRB submission and research preparations
Global	1) Harvard Research Team 2) CDiC Country KOLs 3) Dure Technologies	Monthly	- Study-wide alignment - Share international progress - Build

Table 9: Meeting participants, cadence and purposes.

c. Governance and Management

The three main stakeholders in establishing the GC-CDiC are team members of the Harvard Health Systems Innovation team who are responsible for managing and executing the four interrelated streams of work comprising GC-CDiC (I) research, (II) data systems, (III) innovation, and (IV) translation; Dure Technology, the technology provider and implementer; and the country teams and key opinion leaders of each of the six ‘wave 1 countries’. As the entity responsible to oversee the entire study, the Harvard team has established the roles and responsibilities of each set of stakeholders in the first phase of the study (see Table 10).

Partner	Key Individuals	Primary Focus
CDiC Country  <p>30 partner countries as of Q1 2024</p>	1) Co-Principal Investigators: lead, supervise and guide within the country 2) Country Researchers: form the local research team under the Co-PI to assist in study implementation, data management and clinic engagement	Design and operation of data systems, data management and research activities

	1) Principal Investigator: holds overall responsibility for the study and its design 2) Senior Project Manager: manages and coordinates the study at Harvard and across countries 3) HSIL Researchers and Country Liaisons: research activities, country engagement, assisting in design of study	Research study design, coordination and management of the GC-CDiC cohort and activities
	1) Dure team representatives: build and assist in the implementation of the data system, troubleshoot technical issues and provide ongoing assistance with the EHR, registries and cohort	Implementation of the data system

Table 10: Stakeholders, roles within each group and responsibilities of each stakeholder.

d. Roadmap, 2024-2025 and country plans.

Phase one of GC-CDiC involves the onboarding of six CDiC countries. This first phase will reach its completion at the end of 2025, with final data translation being completed in early 2026 (Figure 5). As the collaborative progresses to phase two and onwards, additional CDiC countries will be included in data collection until all 30 CDiC partnered countries are part of the virtual cohort. This onboarding will occur alongside phase one, with the next group of countries beginning onboarding as soon as Q4 of 2024.

	2024												2025												2026		
	March	April	May	June	July	August	September	October	November	December	January	February	March	April	May	June	July	August	September	October	November	December	January	February	March		
1 Research Preparation																											
1.1 Development of a research protocol		x	x																								
1.2 Support for countries IRB submission		x	x	x																							
1.3 Development of a mechanism to create a T1D virtual 'cohort' for research		x	x	x	x	x																					
2 Implementation of a data collection system at the clinic level																											
2.1 Development of a clinic readiness assessment to identify clinics in need of research capacity building		x	x	x	x	x																					
2.2 Assessment of clinic data collection systems in relation to CDiC modules and SWEET		x	x	x	x	x	x																				
3 Capacity building / Education and Training																											
3.1 Development of research training sessions to conduct proposed research		x	x	x	x	x	x																				
3.2 Provision of research training and capacity building sessions						x	x	x	x																		
4 Knowledge sharing																											
4.1 Organization of 2 day event at the World Health Assembly in May 2024		x	x	x																							
4.2 Organization of a 2 day workshop at the ISPAD Conference in Lisbon						x	x	x																			
5 Virtual cohort formation (retrospective data) + data analysis																											
5.1 Data acquisition							x		x	x																	
5.2 Cleaning of data									x	x	x																
5.3 Analysis of data											x	x															
6 Virtual cohort formation (new data) + data analysis																											
6.1 Data acquisition																		x	x	x							
6.2 Cleaning of data																			x	x	x						
6.3 Analysis of data																				x	x	x					
7 Data translation																											
7.1 Development of manuscripts for publication												x	x	x									x	x			
7.2 Development of policy reports and recommendations													x	x	x									x	x		
7.3 Development of overall project report																								x			

Figure 5: GC-CDiC Phase 1 timeline

6. References

1. Ward, Z.J., Yeh, J.M., Reddy, C.L., Gomber, A., Ross, C., Rittiphairoj, T., Manne-Goehler, J., Abdalla, A.T., Abdullah, M.A., Ahmed, A. and Ankotche, A., (2022). Estimating the total incidence of type 1 diabetes in children and adolescents aged 0–19 years from 1990 to 2050: a global simulation-based analysis. *The lancet Diabetes & endocrinology*, 10(12), pp.848-858.
2. Chow, C.K., Ramasundarahettige, C., Hu, W., AlHabib, K.F., Avezum, A., Cheng, X., Chifamba, J., Dagenais, G., Dans, A., Egbujie, B.A. and Gupta, R., (2018). Availability and affordability of essential medicines for diabetes across high-income, middle-income, and low-income countries: a prospective epidemiological study. *The lancet Diabetes & endocrinology*, 6(10), pp.798-808.
3. Ward, Z.J., Yeh, J.M., Reddy, C.L., Gomber, A., Ross, C., Rittiphairoj, T., Manne-Goehler, J., Abdalla, A.T., Abdullah, M.A., Ahmed, A. and Ankotche, A., (2022). Estimating the total incidence of type 1 diabetes in children and adolescents aged 0–19 years from 1990 to 2050: a global simulation-based analysis. *The lancet Diabetes & endocrinology*, 10(12), pp.848-858.
4. Ward, Z.J., Yeh, J.M., Reddy, C.L., Gomber, A., Ross, C., Rittiphairoj, T., Manne-Goehler, J., Abdalla, A.T., Abdullah, M.A., Ahmed, A. and Ankotche, A., (2022). Estimating the total incidence of type 1 diabetes in children and adolescents aged 0–19 years from 1990 to 2050: a global simulation-based analysis. *The lancet Diabetes & endocrinology*, 10(12), pp.848-858.
5. Ward, Z.J., Yeh, J.M., Reddy, C.L., Gomber, A., Ross, C., Rittiphairoj, T., Manne-Goehler, J., Abdalla, A.T., Abdullah, M.A., Ahmed, A. and Ankotche, A., (2022). Estimating the total incidence of type 1 diabetes in children and adolescents aged 0–19 years from 1990 to 2050: a global simulation-based analysis. *The lancet Diabetes & endocrinology*, 10(12), pp.848-858.
6. Changing Diabetes in Children programme for children with type 1 diabetes. Novo Nordisk. Available at:
<https://www.novonordisk.com/content/nncorp/global/en/sustainable-business/access-andaffordability/changing-diabetes-in-children.html> (Accessed April 12, 2023.)

7. Novo Nordisk. (2024) 'Changing Diabetes in Children', Novo Nordisk. Available at: <https://www.novonordisk.com/sustainable-business/access-and-affordability/changing-diabetes-in-children.html#> (Accessed: 26 June 2024).
8. Moreire A., Larty K.F., Chhabria S., Reilly A., Carpenter K., Mita C., Shakir Z., Azhar B., Figi J.T., Reddy C.L., Atun R., (2024, pending publication). *Type 1 Diabetes Registries for Children and Adolescents: A Scoping Review*
9. U.S. Department of Health & Human Services. (2024) 'Health Insurance Portability and Accountability Act (HIPAA)', U.S. Department of Health & Human Services. Available at: <https://www.hhs.gov/hipaa/index.html> (Accessed: 10 June 2024).
10. GDPR.eu. (2024) 'General Data Protection Regulation (GDPR)', GDPR.eu. Available at: <https://gdpr-info.eu/> (Accessed: 10 June 2024).
11. *The Information Technology Act, 2000*, India. Available at: <https://eprocure.gov.in/cppp/rulesandprocs/kbadqkdlcswfjdelrquehwuxcfmijmuixngudufgbuubgubfugbububjxcgfvbsdihbgfGhdfgFHtyhRtMjk4NzY=#:~:text=%5B9th%20June%2C%202000%5D%20An,communication%20and%20storage%20of%20information%2C>
12. *Act 709, Personal Data Protection Act 2010*, Malaysia. Available at: <https://www.pdp.gov.my/jdpdpv2/assets/2019/09/Personal-Data-Protection-Act-2010.pdf>
13. African Union, (2023). List of countries which have signed, ratified/acceded to the African Union Convention on Cyber Security and Personal Data Protection.
14. African Union, (2014). African Union convention on cyber security and personal data protection. *African Union*, 27.
15. FDRE, H., (1996). Constitution of the federal democratic Republic of Ethiopia. *Federal democratic republic of Ethiopia*.
16. (2024) *Ethiopia's New Data Protection Law: Enhancing Privacy and Security in the Digital Age*. rep. Addis Ababa, Ethiopia: Mesfin Tafesse & Associates, pp. 1–6.
17. Amazon Web Services (AWS). (2024) 'Amazon S3 Pricing', Amazon Web Services. Available at: <https://aws.amazon.com/s3/pricing/> (Accessed: 26 June 2024).

18. 'Azure Blob Storage Pricing', Microsoft Azure. Available at:
<https://azure.microsoft.com/en-us/pricing/details/storage/blobs/> (Accessed: 26 June 2024).
19. Harvard University. (2024) 'Data Security Levels and Research Data Examples', Harvard University. Available at: <https://privsec.harvard.edu/data-security-levels-research-data-examples> (Accessed: 26 June 2024).
20. Harvard University. (2024) 'Harvard University Information Security Policy', Harvard University. Available at: <https://policy.security.harvard.edu/> (Accessed: 26 June 2024).
21. Harvard University. (2024). 'Collaboration Tools Matrix', Harvard University. Available at: <https://privsec.harvard.edu/collaboration-tools-matrix> (Accessed: 26 June 2024).
22. Sweeney, L., (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05), pp.557-570.
23. U.S. Department of Health & Human Services. (2024). 'OHRP Decision Charts', U.S. Department of Health & Human Services. Available at:
<https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts-2018/index.html> (Accessed: 14 June 2024).
24. U.S. Department of Health & Human Services. (2024). 'De-identification and its Rationale', U.S. Department of Health & Human Services. Available at:
<https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#rationale> (Accessed: 14 June 2024).
25. U.S. Department of Health & Human Services. (2024). 'Covered Entities', U.S. Department of Health & Human Services. Available at:
<https://www.hhs.gov/hipaa/for-professionals/covered-entities/index.html> (Accessed: 14 June 2024).
26. Basin, D., Debois, S. and Hildebrandt, T., (2018). On purpose and by necessity: compliance under the GDPR. In *Financial Cryptography and Data Security: 22nd International Conference, FC 2018, Nieuwpoort, Curaçao, February 26–March 2, 2018, Revised Selected Papers 22* (pp. 20-37). Springer Berlin Heidelberg.

27. Comandè, G. and Schneider, G., (2022). Differential data protection regimes in data-driven research: Why the GDPR is more research-friendly than you think. *German Law Journal*, 23(4), pp.559-596.

7. Appendix

Appendix I – Full set of 35 research questions originally set out by the KOLs

Area of Interest	Research Question
Epidemiology of T1D	What is the current incidence and prevalence of Type 1 Diabetes within the specified population?
Determinants of T1D Development	What are the key risk factors influencing the development and progression of Type 1 Diabetes?
	How does T1D manifest and progress in pediatric populations?
	What factors influence the long-term health outcomes and quality of life for children diagnosed with T1D?
	What biomarkers or genetic factors predict the progression or response to specific therapies in T1D?
	What behavioral interventions are most effective in promoting healthy habits among individuals with T1D?
T1D Treatment and Control	What factors are associated with achieving and maintaining favorable control of Type 1 Diabetes within the study population?
Economic Burden of T1D	What are the direct and indirect costs associated with the management of Type 1 Diabetes, and how do these costs impact individuals and the healthcare system?
	What is the economic burden of T1D on individuals, families, and healthcare systems?
	How do socioeconomic factors influence disease management and outcomes?
Healthcare Utilization and Access	At what stage do Type 1 Diabetes patients enter the healthcare system for the given population?
	What are the patterns of healthcare utilization among individuals with T1D?
	How does access to healthcare services and resources impact disease management and outcomes?
Long-term Complications and Outcomes	What are the complications and long-term outcomes experienced by individuals with Type 1 Diabetes, and how do these factors evolve over-time?
Treatment and Management	What are the most effective treatments or interventions for managing T1D?
	How do various treatment modalities (insulin therapies, diet, exercise, etc.) impact disease outcomes and quality of life?
	What role do immune system changes and autoimmunity play in the development and progression of T1D?
	How can understanding immune responses lead to novel therapeutic strategies or prevention methods?

Risk Factors and Etiology	What are the environmental, genetic, or lifestyle factors associated with the development of T1D?
	How do these risk factors vary across different age groups or geographical regions?
	How do dietary patterns, physical activity, and lifestyle choices affect T1D management and disease progression?
	What is the natural course of T1D in terms of disease progression, complications, and comorbidities over time?
	How do different demographic groups or treatment approaches influence disease progression?
Health Outcomes and Complications	What are the long-term health outcomes associated with T1D, such as cardiovascular complications, neuropathy, retinopathy, etc.?
	How do these outcomes differ based on treatment adherence, duration of the disease, or specific patient demographics?
	What are the mortality rates and causes of death among individuals with T1D over long-term follow-up periods?
	How does disease duration impact mortality risk and causes in T1D patients?
Technology and Innovations	What is the impact of technological advancements like continuous glucose monitoring (CGM) or insulin pumps on disease management and patient outcomes?
	How do novel innovations in T1D care influence treatment adherence and quality of life?
Pregnancy and Maternal Health	How does T1D affect pregnancy outcomes and maternal health?
	What are the best practices for managing T1D during pregnancy to ensure both maternal and fetal well-being?
Psychosocial and Quality of Life aspects	How does T1D impact the quality of life, mental health, and psychosocial well-being of individuals living with the condition?
	What interventions or support systems improve the psychological aspects and overall well-being of T1D patients?
	How do social support networks and psychological interventions influence the mental health and coping mechanisms of individuals with T1D?
	What strategies can improve social support and mental well-being in this population?

Appendix II – Recommended Variables to be Collected for the Core to Core+++ Modules

	Core		Core+		Core++		Core+++	
	Item	Frequency	Item	Frequency	Item	Frequency	Item	Frequency
Demographic Data	Sex at birth	Once	Religion	Once				
	Date of Birth	Once	Migrant Status	Once				
	Ethnicity	Once	Gender	Annually				
	Address (City)	Once (and when changes)						
	Address (Town)	Once (and when changes)						
Socioeconomic Data			Household Income	Once	# of people in household	Annually	Mortality of parent	Annually
			Parent Education Level	Once	Patient education level	Annually		
			Insurance Status	Annually				
Clinical Data	Height	Every visit, Until age 21	ED or inpatient stay	Annually	Mental health disorders	Annually	Other comorbidities (coded as ICD-11 codes)	Annually
	Weight	Every visit	- If yes, reason:	Annually	Celiac disease screen	Year 0, 2, 5	Pubertal stage	Until age 18
	Body Mass Index (BMI)	Calculated (kg/m ²)	Smoking	Annually	- Celiac antibodies result	If tested		
	Blood Pressure (Systolic/Diastolic)	Every visit			Telemedicine Consultation	Every visit		
	BMI Z-score	Calculated						
	Family History of T1D	Once						
	Date of diagnosis	Once						
	Symptoms at Diagnosis	Once						
	Patient death	If occurs						
	Date of death	If occurs						
	Cause of death	If occurs						
	Type of Diabetes	Once						
Medical Therapy	Type of Insulin	Every visit	Oral diabetic medication	Every visit	Pump vs. injection	Annually		
	Insulin Regimen		Current medications	Every visit	Frequency of missed doses	Every visit		
	- Insulin injections per day	Every visit	Injectable diabetic medication	Every visit	Glucose sensor use	Annually		
	- Daily insulin dose	Every visit			Closed loop	Annually		
	- Daily basal dose	Every visit						
	- Daily prandial dose	Every visit						
	- Type of basal insulin	Every visit						
	- Type of prandial insulin	Every visit						
	SMGB Frequency	Every visit						
Lab Values	HbA1c	Every 6 months	Fasting lipids:	Annually	Urine Creatinine	Annually	OGTT	Once
	Patient reported glucose range	Every visit	- Total Cholesterol		CBC	Annually	Pancreatic autoantibodies	
			- LDL		- WBC		ICA	Once
			- HDL		- RBC		IAA	Once
			- Triglycerides		- Hemoglobin		GADA	Once
			Serum creatinine	Annually	- Hematocrit		IA-2A	Once
			Thyroid function tests	Annually to 2 yearly	- MCV		C-peptide	Once
			- TSH		- MCH		Genetic testing	Once
			- Free T4		- Platelets		- Number of subtype according to ISPAD classification	Once
					- Neutrophils		- Specification of diabetes subtype	Once
					- MCHC		Time in range during last 2 weeks	Every visit
					- RDW		Time below range during last 2 weeks	Every visit
					- Lymphocytes (monocytes)			
					- Lymphocytes (eosinophil)			
					- Lymphocytes (basophil)			
					Thyroid peroxidase Ab	If TFT abnormal		
					Antithyroglobulin Ab	If TFT abnormal		
Complications	DKA	Every visit	Hospitalizations for hypoglycemia	Every visit	CVA	Annually		
			Eye exam for retinopathy	Annually	MI	Annually		
			- If yes, retinopathy?	Annually				
			Skin filament test for neuropathy	Annually				
			- If yes, neuropathy?	Annually				
			Microalbuminuria for nephropathy	Annually				
			- If yes, nephropathy?	Annually				
90 items total	27 items total		23 items total		27 items total		13 items total	