INST 327
Section 0203
Team Final Project
05/15/2025
Names: Selam Fesseha, Genevieve Koduol, Sarah Bamba, Alicia Debra, Ana Gabriela Salazar

Introduction

Diabetes affects millions of people worldwide, leading to serious health complications like heart disease, kidney failure, and vision loss. Managing diabetes requires continuous monitoring, lifestyle adjustments, and medical support, making it a major public health challenge. Despite the amount of diabetes research, accessing well-organized data that connects lifestyle factors, genetics, and medical indicators remains difficult. Our goal is to create a structured database that improves how diabetes-related information is stored.

For this project, we are working with a diabetes dataset compiled by Ankit Batra. This dataset includes 70,000 records from hospitals, public health databases, and research studies, covering multiple types of diabetes, including Type 1, Prediabetes, and Steroid-Induced Diabetes. The dataset contains a range of medical, genetic, and lifestyle attributes, allowing us to explore key factors that contribute to diabetes risk and progression. However, in its raw form, the dataset is vast and unstructured, making it difficult to extract meaningful insights efficiently.

To address this, we will design a well-organized database that makes it easier to search, analyze, and interpret diabetes-related data. Our database will focus on 7 key factors: alcohol consumption, cholesterol level, history of PCOS, blood pressure, insulin levels, genetic markers, and digestive enzyme levels. These attributes are critical in understanding diabetes risks and patterns. By cleaning the dataset for analysis, our database will provide a more user-friendly and efficient tool for answering important research questions, such as the role of genetics in diabetes and how lifestyle choices impact disease progression. Additionally, we will ensure that the database is handled responsibly, given the sensitive nature of health-related data. By improving access to structured diabetes data, we hope to contribute to better research, prevention, and management strategies for diabetes.

Database Description

Our database is designed to investigate diabetes risk factors in adults. The schema is centered around a table titled “Patients,” where each patient is assigned a patient_id. The

“Patients” table is connected to six other tables, AutoimmuneProfiles, LabResults, MedicalHistory, PregnancyHistory, VitalSigns, and LifestyleFactors. LabResults include clinical measurements, including insulin, cholesterol, blood pressure, glucose levels, digestive enzymes, liver function, and neurological assessments. MedicalHistory tracks genetic markers, history of PCOS, and steroid use, LifestyleFactors currently focuses on alcohol consumption with room to expand to other behavioral variables, AutoimmuneProfiles tracks potential autoimmune disease for patients, PregnancyHistory tracks if the patient has been pregnant and the symptoms surrounding this, and VitalSigns tracks basic vitals taken at appointments such as glucose, blood pressure, and insulin level.

Our database is normalized so that there are minimal data redundancies and more query efficiency. Separating the patient data into tables allows us to complete complex joins and aggregate queries. This normalized structure makes it easier in terms of efficient querying and comprehensive analysis of the multiple factors that may contribute to diabetes risk in older adults. Overall, our schema supports both exploratory and targeted analysis of the diabetes risk in adults while also remaining adaptable for the future in terms of expanding data as well as refining it.

Sample Data

For our sample data, we first filtered out the ages from 30-40, then we used the first 15 columns of the diabetes database. Below are all of our tables, including the sample data.

Patients Table

This table allows patients to store data pertaining to their birth weight and type of diabetes to make a potential connection between the two.

patient_id	birth_weight	target
1	2629.00	Steroid-Induced Diabetes
2	1881.00	Neonatal Diabetes Mellitus (NDM)
3	3622.00	Prediabetic
4	3452.00	Type 1 Diabetes
5	1770.00	Wolfram Syndrome
6	3835.00	LADA
7	4426.00	Type 2 Diabetes
8	1644.00	Wolcott-Rallison Syndrome
9	3721.00	Secondary Diabetes
10	4206.00	Secondary Diabetes
11	3965.00	Secondary Diabetes
12	2947.00	Type 1 Diabetes
13	3588.00	Prediabetic
14	2202.00	Neonatal Diabetes Mellitus (NDM)
15	2984.00	LADA
NULL	NULL	NULL

Lifestyle Factors Table

This table contains data relating to the type of lifestyle that the given patient lives. This table will allow users to store data of lifestyle habits that affect diabetes.

lifestyle_id	patient_id	smoking_stat...	dietary_habi...	alcohol_consumpti...
1	1	Smoker	Healthy	High
2	2	Non-Smoker	Healthy	Moderate
3	3	Smoker	Unhealthy	High
4	4	Smoker	Unhealthy	Moderate
5	5	Smoker	Healthy	Moderate
6	6	Non-Smoker	Healthy	Low
7	7	Non-Smoker	Healthy	Low
8	8	Smoker	Unhealthy	Low
9	9	Smoker	Healthy	Low
10	10	Non-Smoker	Unhealthy	Low
11	11	Non-Smoker	Healthy	Low
12	12	Smoker	Unhealthy	Moderate
13	13	Smoker	Healthy	Moderate
14	14	Non-Smoker	Unhealthy	Moderate
15	15	Non-Smoker	Healthy	High
16	1	Smoker	Healthy	High
17	2	Non-Smoker	Healthy	Moderate
18	3	Smoker	Unhealthy	High
19	4	Smoker	Unhealthy	Moderate
20	5	Smoker	Healthy	Moderate
21	6	Non-Smoker	Healthy	Low
22	7	Non-Smoker	Healthy	Low
23	8	Smoker	Unhealthy	Low
24	9	Smoker	Healthy	Low
25	10	Non-Smoker	Unhealthy	Low
26	11	Non-Smoker	Healthy	Low
27	12	Smoker	Unhealthy	Moderate
28	13	Smoker	Healthy	Moderate
29	14	Non-Smoker	Unhealthy	Moderate
30	15	Non-Smoker	Healthy	High

LabResults Table

This table stores data relating to different types of lab results that the given patient received. We included lab results from blood pressure, cholesterol level, digestive enzyme level, liver function treatment, neurological assessments, and blood glucose levelsv

lab_result_id	patient_id	blood_pressu...	cholesterol_lev...	digestive_enzyme_lev...	liver_function_te...	neurological_assessme...	blood_glucose_lev...
1	1	124	201	56	Normal	3	168
2	2	73	121	28	Normal	1	178
3	3	121	185	55	Abnormal	1	10
4	4	100	151	60	Abnormal	2	121
5	5	103	146	24	Normal	1	289
6	6	127	208	52	Normal	2	142
7	7	115	237	96	Abnormal	3	186
8	8	80	157	29	Normal	1	206
9	9	138	185	74	Normal	3	160
10	10	136	259	42	Abnormal	2	192
11	11	134	193	59	Normal	3	192
12	12	91	195	60	Abnormal	1	114
13	13	128	191	76	Abnormal	1	113
14	14	71	126	29	Normal	1	175
15	15	116	163	43	Normal	2	136

Medical History Table

This table allows the user to store any data related to their medical history. We included medical history such as genetic markers, genetic testing, history of PCOS, and steroid use.

history_id	patient_id	genetic_marke...	genetic_testi...	history_of_PCOS	steroid_use_hist...
1	1	Positive	Positive	No	No
2	2	Positive	Negative	Yes	No
3	3	Positive	Negative	Yes	No
4	4	Negative	Positive	No	No
5	5	Positive	Positive	No	No
6	6	Negative	Negative	No	No
7	7	Positive	Negative	No	Yes
8	8	Negative	Negative	Yes	No
9	9	Positive	Positive	No	Yes
10	10	Negative	Positive	No	Yes
11	11	Positive	Negative	No	No
12	12	Positive	Negative	Yes	No
13	13	Positive	Positive	No	Yes
14	14	Negative	Negative	Yes	Yes
15	15	Positive	Positive	No	No

Vital Signs Table

This table allows patients to store basic vital sign data. We included vital signs such as glucose level, blood pressure, and insulin level.

	vital_id	patient_id	glucose_level	blood_pressu...	insulin_level
	1	1	168.00	124.00	40.00
	2	2	178.00	73.00	13.00
	3	3	105.00	121.00	27.00
	4	4	121.00	100.00	8.00
	5	5	289.00	103.00	17.00
	6	6	142.00	127.00	17.00
	7	7	186.00	115.00	29.00
	8	8	206.00	80.00	10.00
	9	9	160.00	138.00	47.00
	10	10	192.00	136.00	21.00
	11	11	192.00	134.00	16.00
	12	12	114.00	91.00	8.00
	13	13	113.00	128.00	22.00
	14	14	175.00	71.00	9.00
	15	15	168.00	124.00	40.00

Pregnancy History Table

The table demonstrates whether any patient has been pregnant in the past and the type of symptoms and complications they had. We included previous gestational diabetes, weight gain during pregnancy, and pregnancy history in order to see if these relate to a certain type of diabetes.

patient_id	previous_gestational_diabe...	weight_gain_during_pregna...	pregnancy_hist...
1	No	15	Normal
2	Yes	15	Normal
3	Yes	19	Complications
4	Yes	26	Complications
5	No	11	Normal
6	Yes	14	Normal
7	Yes	28	Complications
8	Yes	15	Complications
9	No	26	Complications
10	No	17	Complications
11	No	20	Complications
12	Yes	12	Normal
13	Yes	14	Normal
14	Yes	25	Complications
15	Yes	23	Complications

AutoimmuneProfiles Table

This data allows the patient to store their data based on if they have autoimmune diseases and how it relates to the type of diabetes they have. We decided to include the columns of autoantibodies, early onset symptoms, pancreatic health, and cystic fibrosis diagnosis.

	profile_id	patient_id	autoantibodi...	early_onset_sympto...	pancreatic_hea...	cystic_fibrosis_diagno...
	1	1	Negative	No	36	No
	2	2	Negative	Yes	26	Yes
	3	3	Positive	Yes	56	Yes
	4	4	Positive	No	49	Yes
	5	5	Negative	No	10	No
	6	6	Negative	Yes	40	Yes
	7	7	Negative	No	62	Yes
	8	8	Negative	Yes	13	Yes
	9	9	Positive	No	91	No
	10	10	Negative	No	86	Yes
	11	11	Positive	No	64	No
	12	12	Negative	No	67	No
	13	13	Negative	No	63	Yes
	14	14	Positive	No	16	No
	15	15	Negative	No	76	No

Changes from the original design

In the old design, all test results and medical history lived in just four tables: Patients, DiagnosticTests, MedicalHistory, and LifestyleFactors with one big table holding every metric and foreign keys stuck on Patients. In the new design, we split DiagnosticTests into two tables: VitalSigns for measures like glucose and blood pressure, and LabResults for lab assays like cholesterol and enzyme levels. We added AutoimmuneProfiles with its own ID to store immune data and a separate PregnancyHistory table for pregnancy details. MedicalHistory now has its own ID and only holds genetic markers, testing info, PCOS, and steroid-use history. LifestyleFactors still tracks smoking, diet, and alcohol but now requires smoking status and dietary habits to have values. Every table has a clear primary key (either its own ID or, in one-row tables, patient_id), a single patient_id foreign key back to Patients, and at least one required column. After feedback from the TAs we added those tables to make our data more complex. These changes organize the data better, avoid repeating information, keep it accurate, and make queries easier.

Database Ethics Considerations

Demographic Representation

Our database limits patients to ages 30–40 to focus our analysis, but this choice means we're missing insights about younger or older adults. We also rely on "high" or "low" risk labels for ethnicity rather than actual race or background. Those labels can hide real differences between groups, like unequal access to hospitals or healthy food, and might overstate or understate someone's true risk. In future work, we should look for more detailed demographic data or at least acknowledge that our risk labels are rough and may not capture everyone's reality.

Lifestyle and Environmental Factors

We kept diet, BMI, and waist measurements, but those numbers don't tell us why someone has an unhealthy diet or high BMI. Wealth and neighborhood play a big role: people in food deserts or without safe places to exercise will often score lower on lifestyle measures. We need to be clear that these columns reflect opportunity as much as personal choice. Wherever possible, we'll note in our reports that lifestyle scores are only rough indicators unless we can gather more detail (for example, types of foods eaten or gym access).

Medical and Genetic Biases

Adding genetic markers or autoantibodies can help explain some diabetes cases, but it can also make genetics seem like the only cause and downplay how much environment, stress, or income matter. We must avoid implying that patients "get" diabetes just because of their genes. When we write up our findings, we'll highlight both genetic and non-genetic factors and remind readers that health outcomes come from a mix of biology, behavior, and environment.

Data Privacy and Fair Use

We protect patient privacy by never storing names or personal details; each person is known only by a random patient ID. Our use of this data is strictly for our class project, which falls under fair use, but we still treat it carefully. We won't share any raw data outside our team, and we'll present only aggregated results (for example, counts or averages) so individual patients can't be identified.

By keeping these points in mind, representation, context for lifestyle data, balanced view of genetics, and strong privacy safeguards, we aim to build a database that is inclusive, fair, and responsible.

Lessons Learned (Selam)

Working with the Diabetes dataset has taught our team about the implications of using medical information and creating a database based on healthcare data. Unlike other datasets, using healthcare data requires an understanding of how the body works to interpret and create meaningful tables. Our team members spent a lot of time researching what kind of factors impact people who have diabetes. We had to learn about medical terminology and understand how the information was recorded and shared. Additionally, a lot of the diabetes data had a lot of inconsistencies, such as no depth in what kind of symptoms patients with diabetes were facing, which made it hard for us to draw some conclusions from our data. Another lesson we learned was realizing that the data does not automatically mean you know what to do with it. We had to critically think about what kinds of queries would reveal something useful from our data. It made us more thoughtful about our research questions and got us thinking like actual analysts.

Potential Future Work (Alicia)

For the future, we plan to increase the number of tables and symptoms we include but also the complexity within our queries. To go into detail, we plan to include more samples. We only included 15 entries for sample data;; however, we should include up to 1000 data entries to get a more conclusive understanding of the given diabetes database. One of the main focuses of this course was data ethics, and one key component of having ethical data is having data that represents the entire population. The database has over 70,000 entries, so it is clear that 15 entries is not an accurate representation of the data.

We also plan to include more detailed column categories such as pulmonary function, cystic fibrosis diagnosis, and autoantibodies. We initially didn't include these categories because they seem to require a heavier understanding of complex health terms, which would require further research on how these symptoms intersect with diabetes risk.

Finally, as we add more columns and entries, we plan to improve our database normalization and develop advanced queries/views using subqueries and CTEs.

Citations

Batra, Ankit. *Diabetes Dataset*. Kaggle, 2022, https://www.kaggle.com/datasets/ankitbatra1210/diabetes-dataset.

Kanica Yashi, & Daley, S. F. (2023, June 19). *Obesity and Type 2 Diabetes*. Nih.gov; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK592412/