HbA1c Test's Accuracy as a Predictor for Diabetes with Complications Diagnosis: Further Analysis of the HbA1c Diabetes Mellitus Test

Author: Liam Cleary

HbA1c Test's Accuracy as a Predictor for Diabetes with Complications Diagnosis: Further Analysis of the HbA1c Diabetes Mellitus Test

by

Liam Cleary

Department of Economics
Senior Honors Thesis
May 2020

Contents

Abstract

HbA1c levels are the most frequently used test for diagnosis and prognosis of diabetes mellitus. Recent studies have shown the biases this test has in particular cohorts, that was not noted when it was originally accepted by the American Diabetes Association in 2008. This study examined how these biases affect HbA1c's ability as a predictor for complications that arise due to diabetes in specific cohorts, those of ethnicity, age, weight, and other patient attributes, compared to other established diabetes prognosis tests. We discovered that both glucose and HbA1c share similar biases as predictors for particular cohorts (the high glucose, high BMI, Asian, African, and Hispanic descent cohorts), HbA1c works better as a predictor when it is combined with the results of a glucose test and more characteristics of the patient compared to a HbA1c test alone with fewer variables, and glucose and HbA1c are better predictors for different diseases, respectively, that may arise due to diabetes mellitus. Multiple patient specific characteristics that include the previously analyzed variables (race, age, BMI) and less analyzed empirically variables (patient's family medical history, patient's susceptibility to specific diseases, and socioeconomic circumstance) all need to be considered when deciding which combination of diabetes prognosis tools are best for that particular patient. The COVID-19 Crisis caused this analysis to switch from risk scores to biomarkers as an indicator for complications, warranting further study be done using risk scores and biomarkers and analyzing the other prognosis tests that were not discussed in this analysis.

Acknowledgements

I would like to thank the Economics Department at Boston College for welcoming me as a part of their department and helping me learn and grow during my four years at Boston College. In particular, I would like to thank Professor Samuel Richardson for advising on my Senior Honors thesis. Our discussions allowed me to narrow down my research and refine my models and regressions in order to properly reach my analysis. I would also like to thank Professor Robert Murphy for guiding me through the Honors program of the Economics Department and leading Omicron Delta Epsilon at Boston College. I would also like to thank the CDC and the Primary Care Network who provided me data necessary to produce this paper. Finally, I would like to thank my parents and family for their support and encouragement during this project.

1. Introduction

Over 30 millions Americans suffer from Type 1 or Type 2 diabetes; this number is expected to increase to 54.9 million by 2030, nearly 23% of those above the age of 65 will have type 1 or type 2 diabetes, with 90% to 95% of those with diabetes having type 2 diabetes (Rowley et al. 2017). Diabetes is a disease that causes a person's body to be unable to produce insulin or have resistance to it. Insulin is the hormone in someone's body that causes the body to absorb excess glucose when its blood glucose level rises above normal. Its counterpart is glucagon which causes certain cells to excrete glucose into the blood stream when its blood glucose level is lower than normal. Type 1 diabetes tends to develop primarily at birth and is characterized by a person being unable to produce insulin, but most people can take insulin to manage their blood sugar levels as their body reacts properly to the insulin hormone. Type 2 diabetes tends to develop later on in a person's life and is characterized by the body developing some intolerance to insulin. The primary symptoms of both diseases are thirst, frequent urinations, hunger, fatigue, and blurred vision due to their cells receiving less sugar compared to if those that have normal insulin levels and insulin tolerance (Nathan 2009).

Both forms of diabetes require similar tracking of blood glucose level to avoid further complications caused by the disease. The higher blood glucose level of diabetics can damage their blood vessels and nerves over time. Someone with uncontrolled diabetes has an increased risk of heart disease, atherosclerosis, and neuropathy (Nathan 2009). It is imperative for someone with diabetes to keep their blood glucose level within the range that their body would have normally kept it if they had normal insulin tolerance. If their blood level is on average too low, they will be fatigued and their cells would receive less glucose than they would need, damaging them over time; if their blood level is too high, they will experience complications due to the high level of blood glucose damaging their blood vessels and nerves (Nathan 2009).

The person's own body cannot maintain optimal blood glucose level so it is up to the person and their physician to consistently test their glucose levels and develop the best lifestyle to maintain optimal glucose levels. There are a range of glucose tests used by physicians and patients to track blood glucose levels. Each test

has inherent costs and benefits; the tests that require the patients to prepare more for it are more accurate, e.g. fasting glucose and oral glucose tolerance tests, and the tests that require minimum preparation by the patient are less accurate, e.g. random blood glucose tests (Picón et al. 2012).

The preferred test for diabetes prognosis today is the hemoglobin A1c or glycated hemoglobin test. The HbA1c test is the most preferred prognosis tool in primary care offices because HCC risk scores use HbA1c levels in their parameters, the American Diabetes Association endorsed it as their preferred diagnosing tool, and is touted as requiring less patient preparation than the fasting glucose test and being more accurate than the random blood glucose test (Bonora et al. 2011).

The HbA1c test does not directly test blood glucose levels, instead it tracks the percentage of red blood cells, erythrocytes, that are covered in glucose. On average, glucose remains on erythrocytes for 60-70 days (Virtue et al. 2004). By tracking how much glucose is on erythrocytes, it is postulated that you can determined how resistant someone is to insulin; a higher average blood glucose level will lead to an increased percentage of erythrocytes coated in glucose. With glucose remaining on blood for an average of 60-70 days, the HbA1c test could be used to determine the average blood glucose level during that time (Virtue et al. 2004).

With the number of patients being diagnosed with diabetes increasing each year and proper blood glucose level monitoring being the most important factors for prognosis, a complete understanding of the limitations of the glycated hemoglobin test is required with it being the preferred prognosis tool of primary care offices.

Limitations of the glycated hemoglobin test arise due to its dependence on the relationship between erythrocytes and glucose. If someone with diabetes suffers from another condition that impairs their erythrocyte production, affects their hormone level, or alters the lifespan of their erythrocyte or the ability of the erythrocyte to be coated in glucose than the glycated hemoglobin test would be a less accurate predictor of their blood glucose level than other tests (Chen et al. 2016). These limitations were known when the American Diabetes Association stated that the HbA1c test was their preferred diabetes diagnostic tool, but the extent of them were

studied in a particular cohort that was not representative of the complete patient population that the diagnostic and prognostic HbA1c test would be used on.

Since HbA1c's acceptance as the preferred test across the world, further tests have been completed from populations outside the United States, particularly in Africa and China, where there is a large population of undiagnosed diabetics. These studies focused on HbA1c's accuracy in diagnosing diabetes (Briker et al. 2019). The results of these studies and others have shown that the HbA1c test is less accurate in diagnosing diabetes for patients outside of the initial tested cohorts. (Briker et al. 2019). Since these results were published, further studies have been done examining the relationship between HbA1c levels and blood glucose levels in diverse populations and how particular patient characteristics (age, pre-existing conditions, race, socioeconomic status) affect HbA1c levels (Villacreses et al. 2019).

This study is focused on how the HbA1c test compares to the fasting glucose test in predicting complications controlling for patient history, BMI, age, socioeconomic factors, and other characteristics over a five year span or predicating future complications using biomarkers. This is the first study that is focused specifically on how HbA1c levels are at predicting complications, using HCC risk score coding indicators or biochemical markers, compared to a more traditional test in a diverse patient population with their near complete medical history and characteristics accounted for.

Question: How is the accuracy of the HbA1c test for predicting diabetes with complications affected by different stated variables (age, socioeconomic status, ethnicity, and weight) compared to the accuracy of a glucose test for predicting diabetes with complications with the same stated variables?

2. Literature Review

In 2008, the American Diabetes Association met and decided to make HbA1c levels the preferred diagnosing tool for diabetes for nonpregnant individuals; this recommendation came from data of David M. Nathan*, International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes*, and others showing its benefits and costs dependent on someone's age and some hormonal conditions. Extensive research was performed to make sure that HbA1c levels correlated to blood glucose levels and glucose tolerance; unfortunately, the patient population used for these experiments were not diverse enough to have external validity hold across the world (Nathan 2009). Recent changes in research parameters and interests have led to healthcare professionals pushing for more diverse populations for testing the accuracies of prognosis and diagnosis tools. This push, in addition to field studies mostly in Asia and Africa, has caused healthcare professionals to reexamine the cost and benefits of the HbA1c test.

In 2011, the HbA1c test was used to test for diabetes in Southeast China and sub-Saharan Africa. Randie R. Little, illustrates in her paper, *Diabetes Blood Tests for People of African, Mediterranean, or Southeast Asian Descent*, the limitations that the HbA1c test has for those of African, Mediterranean, or Southeast Asian Descent. These populations have different forms of hemoglobin that make the HbA1c test less accurate (Little 2011). This study fails to track patients prognosis overtime as it specifically looked at diagnosis. Randie R. Little and Williams Roberts also discussed how different variants of hemoglobin can arise due to age or preexisting condition and how they interfere with HbA1c measurements in *A Review of Variant Hemoglobins Interfering with Hemoglobin A1c Measurement*. This paper specifically examined how HbAS, HbAC, HbAE, HbAD, and HbF affected the immunoassay from different HbA1c tests (Little and Roberts 2009).

Another study that went into further detail about diabetes prognosis and also discussed HbA1c test's correlation with glucose tolerance was, Maria Mercedes Chang Villacreses, Feng Chang, Karnchanasorn Wei, Samoa Raynald Rudruidee, and Ken Chiu's paper discussing the further limitations of the HbA1c test in general and regarding specific races which was titled *SAT-125 Underestimation of the Prevalence of Diabetes and*

Overestimation of the Prevalence of Glucose Tolerance by Using Hemoglobin A1c Criteria. Their results showed that HbA1c test has a tendency for underestimation of DM (diabetes mellitus) and overestimation of NGT (normal glucose tolerance). They focused on diagnosis of diabetes and measurement of blood glucose tolerance (Villacreses et al. 2019) .

In addition to the analysis of HbAlc levels to glucose tolerance and blood glucose levels, the recommended parameters of HbA1c levels are still being defined by the American College of Physician (ACP). In March 2018, the ACP adjusted their recommended optimal A1c levels from between 6.5% to 7% to between 7% to 8%. Hui Shao, Deborah B. Rolka, Edward W. Gregg, and Ping Zhang examined, in their paper *Influence of Diabetes Complications on the Cost-Effectiveness of A1C Treatment Goals in Older U.S. Adults,* the long term healthcare costs and likelihood of complications from groups with HbA1c levels between 7% to 8% and 6.5% to 7% and found maintaining lower HbA1c levels of 6.5% to 7% was just marginally cost effective (Shao et al. 2018). From this research, the ACP changed its recommendation. Shao's paper shows that further examination of how HbA1c levels lead to complications is warranted, as he only examined how specific HbA1c ranges lead to complications and not how accurate HbA1c levels are at predicting calculations compared to fasting glucose levels.

Robert M. Cohen, Robert S. Franco, Paramjit K. Khera, Eric P. Smith , Christopher J. Lindsell, Peter J. Ciraolo, Mary B. Palascak, and Clinton H. Joiner discussed in their paper *Red cell life span heterogeneity in hematologically normal people is sufficient to alter HbA1c* that one large limitation of the HbA1c test is that erythrocytes have different life spans for diabetic patients compared to nondiabetics. Blood cells have been examined to live from 39 to 56 days compared to the normal time of 60-70 days (Cohen et al. 2008). This causes those with diabetes to have a decreased HbA1c level compared to if their blood cells lived a normal period of time.

These papers focused on diabetic diagnosis and correlation to blood glucose levels. The paper that did discuss further complications, focused only on how different HbA1c levels affect complication (Cohen et al. 2008). None of them compared HbA1c to a more established test in predicting diabetes with complications. Analysis of how HbA1c levels correlate to further complications compared to other diabetes prognosis tests in a large patient population is beneficial for the ACP and others as it has not been tested for before and it will help illustrate the accuracy of the HbA1c prognosis test for different patient cohorts.

3. Materials and Methods

4. Primary Care Data

4.1 Data

The data set that will be used first is a proprietary data set provided by a Primary Care Office Network in Upstate New York. The initial data set obtained has 2022 observations for 550 patients. The relevant variables for this data set are HbA1c levels, HCC Risk Scores,, BMI, Age, and Zip code (Table 1). These variables will be a part of the future model, CDC model, and risk scores would have been the left hand side, y, variable. Unfortunately, we could not obtain the full datasets with risk scores but we were still able to use the other variables and include additional variables while using a different variable to replace risk scores function of predicting complications with diabetes ("Primary Care Dataset").

4.2 Assumptions

There are some assumptions that are being made for this model to hold internal and external validity. We are assuming with the dataset that: the patient population will be diverse enough that all of them living in Upstate New York will not have an effect on the data; we have a diverse enough patient population that the race interactive terms will be valid and internal validity will hold, there are enough patients in each cohort that the

data will be significant in holding external validity, zip code will be a strong enough indicator for socioeconomic status, HCC Codes were coded properly so they can be an accurate indicator for complications, and complications are close enough that they can be made into one group ("Primary Care Dataset"). Unfortunately, we were unable to obtain the larger dataset so not all of these assumptions will hold.

4.3 Model

The model that was used to analyzed the Primary Care Network data is a multiple linear regression model that analyzes HbA1c's ability to predict future complications in different cohorts. The complications coefficient is the y variable and was constructed based on reported Risk scores of the most common complications caused by diabetes grouped into one dummy variable (Fig. 1 and Fig. 2). The Model includes HbA1c, BMI, age, socioeconomic status, an interactive term between HbA1c and age, an interactive term between HbA1c and BMI, and an interactive term between the socioeconomic status and HBA1c levels (Fig. 3). The inclusion of BMI, age, and socioeconomic status shows the average percentage of complications that an increase of age or BMI causes. The interactive terms illustrates how HbA1c's ability as a predictor changes based on the cohort that is being analyzed. The socioeconomic status was broken down to examine those who are closer to the poverty line compared to those who are not. This was accomplished by using patient's zip code as a weak replacement for socioeconomic status; those from the zip codes with the lowest average household income were recorded as being in the "poverty" group and have an interactive term between their identifier variable and HbA1c scores. This model was supposed to include more variables, but we were unable to do so given the current situation. The Primary Care Model is listed below (Fig. 3).

Figure 1. Risk Score Variable list

Set E11.65= type 2 diabetes mellitus with hyperglycemia = 3
Set E11.21= Type 2 diabetes mellitus with intercapillary glomerulosclerosis (scarring of kidney blood vessels)
Set E11.22 = Type 2 diabetes mellitus with diabetic chronic kidney disease
Set E11.40 = Type 2 diabetes mellitus with diabetic neuropathy,
Set E11.329= Type 2 diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema,
Set E11.51= Type 2 diabetes mellitus with diabetic peripheral angiopathy without gangrene
Set E11.9 = Type 2 diabetes mellitus without complications

Source: Primary Care Dataset

The Risk Score Code variable (dmdx) was relabeled to show the diagnosis of the patient from their risk score. This relabeling shows that this patient population has a range of further complications caused by diabetes. Some of these complications show high comorbidity with other complications in this patient population, but the comorbidity was not analyzed further as it is out of the scope of this study. The most pertinent diagnosis is Type 2 diabetes mellitus without complications, E11.9, as that will be or standard or control in the model.

Figure 2. Creation of Bernoulli Distribution for Diabetes with and without complications

For Analysis set Ell.9 (Type 2 without complications)=0

Any Score that includes complications=1
For 2016: gen score2016=0 if dmdx2016=="E11.9
replace score2016=1 if missing(score2016)
" complications=1, no complications=0"
Regressed a1c2015 on score 2016

Source: Primary Care Dataset

In the initial regressions, all diagnoses that were not Type 2 without complications were treated as the same diagnoses of Type 2 with complications. This allows the correlation between HbA1c levels and diabetes with complications to be analyzed in a simplified model. Upon obtaining the complete data set with more observations, further analysis is warranted on the correlation between HbA1c levels and the different diabetes with complication diagnoses.

Figure 3. Primary Care HbA1c Alone Model

$$Complications_{i,t} = \beta_0 + \beta_1(HbA1c_{i.t-1}) + \beta_2(bmi) * (HbA1c_{i.t-1}) + \beta_3(age) * (HbA1c_{i.t-1}) + \beta_4(\ socioeconomic\ status) * (HbA1c_{i.t-1}) + \beta_5(BMI) + \beta_6(age) + \beta_7(socioeconomic\ status) + \ \beta_8 * X_{i,t} + \varepsilon$$

5. Primary Care Results

Due to the small number of variables, the model produced interesting results that are difficult to determine any significance from. The only variables that were statistically significant were the age variable and age HBA1c interacting variable (Fig. 4). This was expected as an increase in age tends to lead to an increase of HbA1c and complications. The low statistical significance may be due to the low number of observations, 789, compared to what the larger dataset would have, the low number of variables not accounting for all the variance in complication predictions that each patient may have, and HbA1c having a non-linear relationships that cannot

be tracked using this multiple linear regression (Fig. 4). This small regression that ended up being a failure and the only use to come from the Primary Data, illustrates the importance of using larger datasets with more variables and finding a new dataset to perform a more complete analysis compared to this model.

Figure 4. Results of Primary Care HbA1c Alone Model

	(1)
VARIABLES	HbA1c Predictors
a1c2015	-0.0606
	(0.121)
bmia1c	0.00161
	(0.00222)
BMI	-0.00927
	(0.0153)
agea1c	0.00328***
	(0.00114)
age	-0.0156*
	(0.00802)
poora1c	-0.0169
	(0.0316)
poor	0.182
	(0.227)
Constant	0.177
	(0.842)
Observations	789
R-squared	0.254

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Source: Primary Care Dataset

5.1 Observed Errors for Primary Care Data

With the initial regression, the largest observed error is that patients were broken down into two distinct sets: those with diabetes without complications and those with diabetes with complications (Fig 1 and Fig. 2). The severity of complications cannot be observed in this model as all complications are a part of the same group. This regression's purpose was to support the idea that further analysis of HbA1c to diabetes with complications is needed (Fig. 4). Unfortunately, the wide definitions of complications will affect the next

models, CDC models, as well. The other observed errors are that not all patients complete the recommended number of HbA1c tests a year, all the patients are from the Upstate New York area, and we do not have a complete medical history for patients. Due to the small size of this model and regression, it is best to not focus on the errors, as this model is understood to be incomplete and an example of why the larger datasets were needed. Fortunately, the CDC was able to supplement this dataset during this difficult time.

6. CDC DATA

6.1 Data

With the COVID-19 crisis, we could not obtain the whole dataset from the Primary Care Network as all non-COVID related clinical research was put on hold. We decided to use the publicly available patient data from the CDC National Center for Health Statistics. We combined their General Demographic Survey, Fasting time Questionnaire, Standard Biochemistry Profile, and Glycohemoglobin Profile from 2015-2016. We are able to combine these datasets due to the CDC tracking each subject with a unique identification number that identifies what surveys and laboratory profiles they took part in (CDC National Center for Health Statistics).

The major CDC dataset that will be used is the Standard Biochemistry Profile dataset as it contains biochemistry markers that can be used to diagnose patients for the diseases or at risk for the diseases that the risk Scores for diabetes tracked. This dataset pertains to average glucose levels, fasting glucose levels, and other biochemistry markers for 6401 observations. The observations have been recorded across America and age group. The Standard Biochemistry Profile from 2015-2016 is being used to analyze the relationship between a glucose level that could be considered uncontrolled or diabetic, greater than 170 mg/dL, and biochemical markers that are indicative of diseases (complications) that arise due to diabetes (“National Health and Nutrition Examination Survey Standard Biochemistry 2015-16 Data”).

The five area of complications being examined in the Standard Biochemistry Profile data set are heart diseases, nerve damage, Kidney damage, liver damage, and cardiac damage. The biochemistry marker level that is indicative of these damages or diseases was obtained through the data from the CDC, Mayo Clinic, and peer reviewed medical articles and are the levels that a doctor would diagnose the patient who has these levels as having complications or be at risk for complications if they had diabetes ("National Health and Nutrition Examination Survey Standard Biochemistry 2015-16 Data").

The biochemical markers for heart disease and liver damage is alanine aminotransferase (ALT) and aspartate aminotransferase (AST) which are indicative of liver damage and heart disease from **hepatitis,** infection, **cirrhosis,** liver cancer, **heart damage and disease**, or other liver diseases ("Liver Function Tests"). Though ALT and AST are indicators for similar diseases, ALT is an indicator more for liver diseases and AST is an indicator more for heart disease so they are not perfectly collinear. The diseases pertaining to diabetes with complications that ALT and AST levels are indicators for are hepatitis (more susceptible with diabetes), cirrhosis, and other heart and liver diseases. An ALT or AST level of 100 IU/L or greater is an indicator that the subject has one of these disease or is at risk for them ("Liver Function Tests").

The biochemical marker for nerve damage that diabetes can cause, **neuropathy**, is creatine phosphokinase. A creatine phosphokinase level of 2000 IU/L or greater is an indicator for moderate to high nerve damage ("Test ID: CK").

The biochemical markers for kidney damage are blood urea nitrogen to creatinine ratio and uric acid. Uric acid is used as an indicator for **gout, renal failure,** leukemia, and psoriasis. The diseases pertaining to diabetes with complications that uric acid levels are an indicator for are gout and renal failure. An uric acid level of 7mg/dL is an indicator that the subject has one of these disease or is at risk for them ("High Uric Acid Level Causes"). A high level of blood urea nitrogen to creatinine ratio (BUN), greater than 20:1 or close to 20:1 with a high level of creatinine is indicative of heart failure, **chronic dehydration, kidney damage** . The

diseases pertaining to diabetes with complications that the BUN ratio is an indicator for are chronic dehydration and kidney damage ("Blood Urea Nitrogen (BUN) Test").

Lastly, the biochemical marker for ketoacidosis, when an acidic substances called ketones build up to dangerous levels in your body, which is an often complication with diabetes is bicarbonate levels. A bicarbonate level of less than 18mEq/L with a glucose concentration above 150mg/dL is an indicator that a subject has ketoacidosis or is at risk for ketoacidosis ("Test ID: HCO3").

These biochemical markers were chosen as indicators for complications of diabetes as they indicate the primary disease that arise due to diabetes and the other diseases that they are indicators for are rare. The other biochemical markers that can be indicators for complications due to diabetes which were considered but ultimately not included were cholesterol, gamma-glutamyl transaminase (GGT), lactate dehydrogenase (LDH), albumin, and hyperkalemia ("Liver Function Tests," "National Health and Nutrition Examination Survey Standard Biochemistry 2015-16 Data"). Cholesterol was not included as it has a high correlation to glucose which would cause near perfect collinearity in the model. GGT was not included as it can be an indicator for alcoholism or frequent alcohol use which would skew the results. LDH was not used as an indicator as it is a similar indicator for cirrhosis as ALT. Albumin was not used as it had a relation to HbA1c, their levels affect each other, which is too complicated for the proposed models. Lastly, hyperkalemia (high potassium levels) was not used due to glucose affecting potassium level and potassium levels also being affected by diet and fasting time ("Liver Function Tests") .

The last biochemical marker used from the Standard Biochemistry profile was glucose, in mg/dL. Glucose levels can be used for an indicator for complications with diabetes (those with high glucose are more likely to develop diabetes with complications), and or an indicator for diabetes or prediabetes, or uncontrolled glucose levels ("National Health and Nutrition Examination Survey Fasting Questionnaire 2015-16 Data"). The Fasting Questionnaire was combined with the Biochemistry dataset so that we could see whether a subject was

in a fasting, ate eight or more hours ago, or non-fasting state, ate 3 or less hours ago. A glucose level of 140 to 199mg/dL is indicative of prediabetes and a glucose level of 200mg/dL or greater is indicative of diabetes for non-fasting individual. A glucose level of 100 to 125mg/dl is indicative of prediabetes and a glucose level of 125mg/dL or greater is indicative if diabetes for fasting individuals ("Diabetes"). From this data, we were able to determine whether a subject had diabetes or prediabetes level of glucose based on their glucose level and fasting status, recorded as PHAFSTHR in the dataset.

The Glycohemoglobin Survey was combined with the Biochemistry Profile to determine a subject's HbA1c levels. The HbA1c levels are the most common used indicator for diabetes and complications for diabetes and their ability as a predictor will be compared to glucose ("National Health and Nutrition Examination Survey Glycohemoglobin 2015-16 Data"). Glucose and HbA1c will also be incorporated into the same model to determine if using both of them as predictors together for complications is better than using HbA1c or glucose alone.

The General Demographic Survey was used to determine a subject's age, ethnicity, socioeconomic status, pregnancy status, weight, and height. Based upon previous studies, it was shown that HbA1c may be a worse predictor for subjects of Asian, Hispanic, or African descent ("National Health and Nutrition Examination Survey Demographic Variables"). The ethnicity, recorded as RIDRETH3 in the dataset, of a subject was obtained to see how accurate glucose or HbA1c levels are as predictors for complications compared to those of European descent. We removed anyone who indicated that they were pregnant, recorded as RIDEXPRG in the dataset, in the Demographic Survey as both glucose and HbA1c level do not work as indicator for those that are pregnant. The height and weight of a subject was obtained to determine the subjects BMI and, therefore, their overall fitness level to determine if BMI affects glucose or HbA1c's ability as a predictor ("National Health and Nutrition Examination Survey Demographic Variables"). The socioeconomic status of the subject was obtained to determine how having a lower economic status, or being closer to the

poverty line, affects glucose's and HbA1c's ability as a predictor and was determined using a variable the CDC created, recorded as INDFMPIR in the dataset, that calculated the household's income to the poverty line for the subject's state. If the subject's household income to poverty line ratio was 2 or less, the subject was recorded as being close to the poverty line ("National Health and Nutrition Examination Survey Demographic Variables"). The age, recorded as RIDAGEYR in the dataset, of the subject was obtained to determine if there is a correlation between age and HbA1c's or glucose's ability as a predictor ("National Health and Nutrition Examination Survey Demographic Variables").

6.2 Models

Upon combining the data sets and analyzing the data, it was determined that three multiple linear regressions that examine glucose's, HbA1c's, and HbA1c's with glucose's ability to linearly predict any complications or future complications that arose or will arise in a subject, on average, due to diabetes based upon the biochemical markers in specific cohorts would be used to analyze HbA1c versus glucose's effectiveness as a prognosis tool for diabetes with complications. The y variable of all these models will be a dummy variable for having the required biochemical marker level for any of the previous discussed diseases to indicate a complication that can arise due to diabetes (Fig. 5). If a subject meets one or many of the discussed required biochemical levels chosen as indicators for complications, then complications will equal one. We chose to focus on all possible complications in a regression instead of focusing on one biochemical marker in multiple regressions because HbA1c and Glucose may be better indicators for some specific diseases respectively, but we are concerned with focusing on the predicating diabetes with complications in general and replicating the Risk Score Analysis using biochemical markers as replacements and using a more robust dataset. The Glucose Alone and HbA1c Alone Regressions are concern with seeing how glucose and HbA1c are as predictors for complications in the specific cohorts. HbA1c's or Glucose's change in predictor ability for specific cohorts were recorded using interactive variables; the HbA1c level or Glucose level was multiplied by the dummy variable for the cohort (i.e. HbA1c*dummy variable, where dummy variable==1 if the subject is in

that cohort) and compared to the base group to determine if there was significant difference between the coefficients. These dummy variables for the cohorts were included in the regression as standalone variables as well to determine the variance in the average level of complications for each cohort and to make sure that there is no bias in the interacting variables (Fig. 5). Also, the p-values and standard deviations were calculated for the pertinent variables with nearly all of them, besides the Asian descendent cohort, being statistically significant (Appendix Fig. 16). The HbA1c Complete Model does this as well, but additionally includes glucose levels and an interactive term between glucose and HbA1c level (i.e. HbA1c*glucose) to determine if using glucose levels with HbA1c is better than using HbA1c by itself as a predictor and vice versa.

Figure 5. Variables and Complications for CDC Combined Dataset

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
glucose	5,985	102.3	40.35	19	610
HbA1c level	6,050	5.743	1.098	3.800	17
Fasting hours	2,988	10.92	3.575	0	37
complications	9,654	0.549	0.498	0	1
High glucose group	9,654	0.402	0.490	0	1
Fasting high glucose	9,654	0.00673	0.0818	0	1
Pre-high glucose	9,654	0.0267	0.161	0	1
Fasting pre-high glucose	9,654	0.00663	0.0812	0	1
Hispanic descent	9,654	0.315	0.464	0	1
African descent	9,654	0.215	0.411	0	1
Asian descent	9,654	0.105	0.307	0	1
Old (>60 yrs)	9,654	0.189	0.391	0	1
Poor	9,654	0.480	0.500	0	1
Medically Obese	9,654	0.579	0.494	0	1

Source: CDC National Center for Health Statistics

6.3 Variables

There are 23 variables that we were interested in all of the models. The Glucose Alone Model has 18 variables, which are Complications, Glucose level, High Glucose Indicator for Diabetes* Glucose, Prediabetes Glucose Indicator* Glucose, Obese, Hispanic Descent, African Descent, Asian Descent, Poverty, Glucose*

Obese, Glucose* Poverty, Glucose* Hispanic Descent, Glucose* African Descent, and Glucose* Asian Descent (Fig. 6). The Complications variable is a dummy variable that indicates whether a patient is likely to have or develop a disease or condition that which diabetes can help cause or be a complication on top of diabetes based on the disease's biochemical markers. The list of possible complications, the biochemical marker levels indicative of that complication, and why they were used as opposed to other were discussed previously. Whether a patient's biochemical market meets or exceeds the minimum level for a complication or expectation of a future complication have been calculated using data from the CDC and Mayo Clinic that indicates the minimum level of biochemical marker that a doctor would diagnose a patient of having that disease or be at risk for that disease if their patient had that level of biochemical marker. The Complications variable is the y variable for all of the CDC data models (Fig 6, Fig. 7, and Fig 8.).

Figure 6. CDC Glucose Alone Model

Complications= β_1 + β_2 (glucose level)+ β_3 (glucose level)* (high glucose indicator)+ β_4 (glucose level)*(fasting indicator for diabetes)+ β_5 (glucose levels)* (fasting indicator for prediabetes) + $\beta_6(Asian\ descent) *$ (glucose level) + $\beta_7(African\ descent) *$ (glucose level) + $\beta_8(Hispanic\ descent) *$ (glucose $level$) + $\beta_9(obese) * (glucose\ levels) + \beta_{10}(poverty) * (glucose) + \beta_{11}(old) *$ $(glucose)$+ β_{12}(glucose level)* (high glucose indicator for prediabetes) + $\beta_{13}(old) + \beta_{14}(hispanic) + \beta_{15}(asian) + \beta_{16}(black) + \beta_{17}(poverty) + \beta_{18}(overweight) + \beta_{19} * X_i + \varepsilon$

The Glucose level variable is the recorded levels of glucose in the subject's blood in mg/dL. The glucose level variable indicates how an increase of 1mg/dL of a subject's glucose increases the likelihood that a subject will have complications of further complications on average, or the likelihood of being considered to have high biochemical markers for diseases that are common for diabetes with complication. The glucose levels varies on a normal person depending upon when they last ate and therefore fasting time needs to be accounted for. The High Glucose Indicator variable accounts for the fasting time and is a combined dummy variable of a Fasting High Glucose variable, which tracks which subjects who ate 8 hours or later and have glucose levels indicative of diabetes given their fasting status, and a Non-Fasting High Glucose Variable, which tracks subjects who ate 3

hours or less and have a glucose level indicative of diabetes given their non-fasting status. The glucose levels indicative of diabetes for fasting and non-fasting subjects were discussed previously. The High Glucose Indicator for Diabetes*Glucose interactive variable indicates how having glucose levels indicative of diabetes affects glucose's ability to predict complications or future complications for a subject on average. The Prediabetes Indicator variable is similar to the High Glucose Indicator variable as it accounts for the fasting time and is a combined dummy variable of a Fasting Prediabetes variable, which tracks which subjects who ate 8 hours or later and have glucose levels indicative of prediabetes or uncontrolled given their fasting status, and a Non-Fasting Prediabetes Variable, which tracks subjects who ate 3 hours or less and have a glucose level indicative of prediabetes or uncontrolled given their non-fasting status (Fig. 5). The glucose levels indicative of prediabetes for fasting and non-fasting subjects were discussed previously. The Prediabetes Indicator*Glucose interactive variable indicates how having glucose levels indicative of prediabetes or uncontrolled affects glucose's ability to predict complications or future complications for a subject on average (Fig. 6). The obese variable is a dummy variable that indicates whether a subject is considered medically obese, BMI greater or equal to 30 given height and weight. The Glucose*Obese interactive variable indicates how having a BMI greater than 30, or being considered medical obese, affects glucose's ability to predict complications or future complications for a subject on average (Fig. 6).

The Asian Descent, Hispanic Descent, and African Descent dummy variables indicates whether a subject is from Asian descent, Hispanic descent, or African descent respectively. The Glucose* Hispanic Descent, Glucose* African Descent, and Glucose* Asian Descent interactive variables indicate how being of Hispanic, African, or Asian descent affects glucose's ability to predict complications or future complications for a subject on average respectively (Fig. 5). The Poverty variable is a dummy variable that indicates whether a subject is close to the poverty line, household income 2x poverty level or less. The Glucose* Poverty interactive variable indicates how being closer to the poverty line, or to be in a lower socioeconomic class, affects

glucose's ability to predict complications or future complications for a subject on average. The Old variable is for any subject 60 years or older and shows how Glucose works as a predictor for an older cohort (Fig. 6).

Figure 7. CDC HbAlc Alone Model

Complications = β_1 + β_2 (HBA1c level) + β_3 (HBA1c level)* (high glucose indicator)+ β_4 (HBA1c level) *(glucose indicator for prediabetes) + β_5 (HBA1c level) *(fasting indicator for diabetes)+ β_6 (HBA1c level) *(fasting indicator for prediabetes) +$\beta_7(Asian\ descent) *$ (HBA1c level) + $\beta_8(African\ descent) *$ (HBA1c level) + $\beta_9(Hispanic\ descent) *$ (HBA1c level) + $\beta_{10}(obese)(HBa1c) + \beta_{11}(poverty)(HBa1c) + \beta_{12}(old)(HBa1c) + \beta_{13}(old) + \beta_{14}(hispanic) + \beta_{15}(asian) + \beta_{16}(black) + \beta_{17}(poverty) + \beta_{18}(overweight) + \beta_{19} * X_i + \varepsilon$

The HbA1c Alone Model has 18 variables, which are Complications, HbA1c levels, High Glucose Indicator for Diabetes* HbA1c, Prediabetes Glucose Indicator* HbA1c, Obese, Hispanic Descent, African Descent, Asian Descent, Poverty, HbA1c* Obese, HbA1c* Poverty, HbA1c* Hispanic Descent, HbA1c* African Descent, and HbA1c* Asian Descent (Fig. 7). The HbA1c Alone model is a very similar model to that of the Glucose Alone model with HbA1c levels, recorded in percentage, replacing Glucose levels and in the interactive variables. In this Model, the HbA1c level variable indicates how an increase of 1% of HbA1c in the subject's blood increases the likelihood that a subject will have complications of further complications on average. As well, the interactive terms, High Glucose Indicator for Diabetes* HbA1c, Prediabetes Glucose Indicator* HbA1c, HbA1c* Obese, HbA1c* Poverty, HbA1c* Hispanic Descent, HbA1c* African Descent, and HbA1c* Asian Descent are examining the same thing as the Glucose Model variables, except they are examining how being part of a specific cohort or having a medical condition affects HbA1c's ability to predict complications or future complications for a subject on average (Fig. 7).

Figure 8. CDC Complete HbAlc Model

Complications = β_1 + β_2 (HBA1c level) + β_3 (HBA1c level)* high glucose indicator+ β_4 (HBA1c level) *(glucose indicator for prediabetes) + β_5 (HBA1c level) *(fasting indicator for diabetes)+ β_6 (HBA1c level) *(fasting indicator for prediabetes) +$\beta_7(Asian\ descent) *$ (HBA1c level) + $\beta_8(African\ descent) *$ (HBA1c level) + $\beta_9(Hispanic\ descent) *$ (HBA1c level) + $\beta_{10}(obese) * (HBa1c) + \beta_{11}(poverty) * (HBa1c) + \beta_{12}(old) * (HBa1c)$+$\beta_{13}$(glucose levels)+$\beta_{14}$ (HBA1c level)*(glucose levels) + $\beta_{14}(old) + \beta_{15}(hispanic) + \beta_{16}(asian) + \beta_{17}(black) + \beta_{18}(poverty) + \beta_{19}(overweight) + \beta_{20} * X_i + \varepsilon$

The Complete HbA1c Model has 19 variables and is the HbA1c Alone Model with glucose incorporated into it. The variables that are added to the HbA1c model in the Complete HbA1c Model are the Glucose levels variable and the HbA1c*Glucose levels variable (Fig. 8). The Glucose level has the same function as it did in the Glucose Alone Model. The HbA1c*Glucose levels interactive variable shows how using both Glucose and HbA1c to track complications or future complications affects the Glucose levels and HbA1c levels variables ability of predicting complications or future complications themselves. The HbA1c*Glucose levels interactive variable shows how an increase of 1 percentage of HbA1c and 1mg/dL increase of glucose together affects glucose's and HbA1c's ability to predict complications or future complications for a subject on average respectively (Fig 8).

6.4 Assumptions

In these models, we are assuming certain things about the nature of the dataset and variable to maintain external and internal validity. The CDC states that the population for the demographic survey and Biochemistry Profile are diverse enough to be a good representation for the US populations. We are also assuming that the reported levels by the CDC and indicated levels of a biomarker to predict a disease by the Mayo Clinic and the Medical Research papers are accurate and good indicators for diabetic complication diseases to maintain internal validity. We are also assuming that there are enough subjects in each specific cohort to maintain external validity. We are also assuming that subjects reported their demographic data honesty and there was no selection bias for those who chose to partake in the Biochemistry Profile or other laboratory tests. Lastly, we are assuming the relationship between glucose or HbA1c and predicting complications is linear in parameters and can be analyzed through a multiple linear regression. With these assumptions, we can use the results from the three models to see

how glucose and HbA1c are as predictors for complications in general and specific cohorts. We can also see how combining glucose levels with the HbA1 model affects HbA1c's ability as a predictor.

6.5 Expected Results

From the Literature Review and initial regression using the Primary Care network, we are assuming that HbA1c combined with glucose levels is a better predictor of complications than HbA1c alone and HbA1c's ability of a predictor is dependent upon the particular cohort being analyzed (e.g HbA1c is a worse predictor for those of African descent , Asian descent, Hispanic descent, prediabetic glucose levels, High Glucose levels, or older in age). We cannot assume whether glucose or HbA1c is a better glucose in this dataset as the way glucose levels were taken in this population is closest to a random glucose test which is less accurate than an Oral Glucose test or purely fasting glucose test.

7. CDC Results

7.1 Glucose Model

The results of the Glucose Alone Multiple linear regression shows that glucose is a predictor for complications associated with diabetes, ceteris paribus. The coefficient for the glucose variable was 0.00129 and is statistically significant, meaning that an increase of 1mg/dL of glucose for a subject increases their likely hood of having or being at risks for complications associated with diabetes by 0.129%, ceteris paribus. Glucose has a much wider range than that of HbA1c so the coefficient for glucose needs to be compared, possibly multiplied by 20 or 15 given the range of glucose can be 200 and the range of HBA1c is around 10 (Fig. 9). For the different cohorts, there was no statistically significant interactive term between glucose and the cohort identifier in the original regression. However, a linear combination of estimators (lincom) regression of the glucose variable plus the cohort identifier does show a statistically significant difference (Fig. 10). The cohorts where glucose had a different coefficient as a predictor were those of Hispanic descent, African descent, close

to the poverty line, having prediabetic levels of glucose, having diabetic levels of glucose, and considered medically obese (with a p values less than 0.1 but greater than 0.05). For the Hispanic descent cohort, glucose is less accurate of a predictor for complications, ceteris paribus, their coefficient being 0.00103 compared to the base glucose coefficient of 0.00129. This is the same for the African descent cohort to a lesser extent, their coefficient being 0.00110 compared to the base glucose coefficient of 0.00129. The glucose coefficient for the high glucose group was the most lowest compared to the base glucose coefficient, 0.000915 compared to 0.00129, ceteris paribus, showing that glucose's effect at predicting diabetes can begin to flatten out after a certain level for those who are diabetic with a high glucose level, ceteris paribus, (Fig 10.). However, the glucose predictor coefficient is higher for those in the prediabetic range, 0.00142, ceteris paribus. This may be because this is the range where the change in glucose levels have the most effect on the subject's likelihood of developing or having complications. All the subjects in the high glucose cohort have such high glucose that it produces diminishing returns on predicting complications for each 1mg/dL increase. The prediabetes cohort is in the range where a small increase in glucose can damage the body greatly. This may be why the medically obese cohort has a lower coefficient as well, 0.000905, ceteris paribus (Fig. 10). They are more likely to have a higher glucose score given their BMI and face diminishing return as their glucose is already above the range at which 1mg/dL has the greatest effect on their body. For those closer to the poverty line, the coefficient for glucose as a predictor was higher than the base glucose as well, 0.00137, ceteris paribus (Fig. 10). This may be because the dummy variable in the first model alone was not statistically significant and was greatly negative which may have skewed these results. It may also be because there are more outliers of glucose levels and those with complications in this cohort making glucose a better predictor. However, it cannot be certain and one increased coefficient is not enough to assume anything concrete. The old interactive term is positive in this regressions as well. This may be because the biochemical markers we used as indicators for complications and glucose increase with age naturally making this cohort have skewed results.

Also, the cohort identifier variables alone show that those of Hispanic descent, old, closer to the poverty line, and of Asian descent are at higher risk of complications on average in the glucose model, these cohorts have higher indicator constants compared to the base group in the glucose model (Fig. 9). It also shows that those of African descent are at a lower risk of complications on average when comparing glucose. These constants can affect glucose's ability as a predictor with glucose being less of a predictor in a higher risk group due to the already high average level of complications. However, the effect would not be that great and it is best to compare the interactive term to determine the ability of glucose as a predictor for each cohort. The constant for all the analyses (β_1) is not pertinent for our analysis as it pertains to the average percentage of subjects who have complications or are at risk in the base group (Fig. 9). We are more concerned with comparing glucose to HbA1c and seeing the biasness of them in different cohorts.

Figure 9. Results of CDC Glucose Alone Model

	(1)
VARIABLES	Percentage Increase in Predicting Complications
Glucose	0.00129***
	(0.000478)
Glucose*high glucose	-0.000379
	(0.000263)
Glucose*prediabetes	0.000125
	(0.000221)
Hispanic*glucose	-0.000264
	(0.000360)
Hispanic	0.125***
	(0.0396)
African*glucose	-0.000196
	(0.000428)
African	-0.0797*
	(0.0454)
Asian*glucose	-0.000751
	(0.000585)
Asian	0.196***
	(0.0613)
Old*glucose	-0.000728**
	(0.000296)
old	0.163***
	(0.0347)
Poor*glucose	7.89e-05
	(0.000290)
poor	-0.0214
	(0.0318)
Overweight* glucose	-0.000388
	(0.000292)
overweight	0.0756**
	(0.0323)
Constant	0.0911**
	(0.0464)
Observations	5,985
R-squared	0.043
	Standard errors in parentheses
	*** p<0.01, ** p<0.05, * p<0.1

Source: CDC National Center for Health Statistics

Figure 10. Combined Comparisons for CDC Glucose Alone Model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variables	Hispanic Descent	African Descent	Asian Descent	Obese	Poor	High Glucose	Pre-Diabetes
Glucose Predictor Coefficients	0.00103**	0.00110**	0.000543	0.000905*	0.00137***	0.000915**	0.00142***
	(0.000463)	(0.000493)	(0.000610)	(0.00048)	(0.000489)	(0.000393)	(0.000441)
Observations	5,985	5,985	5,985	5,985	5,985	5,985	5,985

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: CDC National Center for Health Statistics

7.2 HbA1c Alone Model

The HbA1c Alone model is more difficult to analyze as the variable that we are most interested in, HbA1c levels, is not statistically significant. This does not necessarily mean that HbA1c cannot be used as a predictor for complications. With the high glucose, prediabetes, Hispanic descent, and old cohorts being statistically significant in the initial regression, it may be indicative that HbA1c levels can be a predictor for complications but in a non-linear way (Fig. 11). For those with normal glucose levels, HbA1c may not work as a good predictor for complications. When a subject's glucose level reaches diabetic or prediabetic levels, then HbA1c can be used as a predictor for complications. For the prediabetic cohort, an increase of 1 percentage of HbA1c levels leads to on average 1.13% increase in complications, ceteris paribus (Fig. 11). The High Glucose group has HbA1c as an even better predictor with their coefficient being 0.0187, ceteris paribus. This increase between the prediabetes and high glucose group is in contrast to the Glucose Alone model having the opposite relationship between the groups (Fig. 12). This may occur due to the A1c coefficient being statistically insignificant so the higher groups do not illustrate a slowing down due to not having the base linear coefficient to begin with. This idea is also reinforced in the lincom regressions with Hugh Glucose having a higher

coefficient than the Prediabetes group in those regression as well. These coefficient increases are similar to an increase of glucose by 20mg/dL.

In the base model, the Hispanic descent and HbA1c interacting variable is statistically significant (Fig. 11). However, in the lincom regressions none of the ethnicity cohort interacting terms are statistically significant (Fig. 12). This may be due to the base variable being non-statistically significant so multiplying it by a dummy terms keeps it still as statistically insignificant. It may also be because the High Glucose and Prediabetes groups are the cohorts with statistically significant coefficients so the interactive term for those cohorts would only pertain to those in the high glucose or prediabetes group as well and those not in it would skew the results and make them statistically insignificant. Even with them being statistically insignificant, the negative or near negative variable recorded in all of the lincom regressions does show that there is a trend for HbA1c to be a less powerful predictor in different ethnicity cohorts (Fig. 12). This trend cannot be completely confirmed with the coefficients not being significant.

The Old interactive term is also negative which is different from the Glucose Alone model and the expected results. This may again be due to the statistically insignificant coefficient and the trend for Older people to have higher glucose and A1c. This trend for higher glucose and A1c would put them into the prediabetes and high glucose group with high A1c levels (Fig. 11). Due to this, HbA1c may show diminishing returns as well eventually when a cohort reaches such a high glucose level or enough of them in general are on the upper end of glucose scores. HbA1c loses some of its predictability, old term is negative 0.0443,ceteris paribus, for a cohort that has a higher trend of glucose in general due to their age (Fig. 11). As well, those closer to the poverty line has a noted slight increase that is statistically significant, 0.0320, ceteris paribus. This may be because of the variance discussed in the glucose model that makes it be more impactful added on to the statistically insignificant HbA1c variable. However, it is hard to determine why it is increased in this Model as well. In the HbA1c model, the cohorts for those of Asian descent, Hispanic descent, and older again had a high chance of complications on average.

It is hard to compare the Glucose Alone to this HbA1c Alone model due to the HbA1c variable being statistically insignificant. It cannot be determined which one has more of a bias for a group or cohort based on this regression (Fig. 11). It can be noted that Glucose has a more linear relationship that begins to decrease it slope in the higher glucose group compared to HbA1c. Both Glucose and HbA1c show the same biases in specific cohorts, but it is again hard to determine with HbA1c being statistically insignificant. A larger model that compares both of them together and their ability to predict complications when combined is the best way to be sure which one is more accurate or less biased. Glucose may be seen as a preliminary better predictor, comparing glucose coefficient *20 to a1cprediabetes or high glucose coefficient, slightly but with their relationships being different it is again hard to determine.

7.3 Comparing CDC HbA1c Alone to Primary Care Network Alone Model

This Model is very similar to the Model that we were able to perform with the Primary Care data that we have. Upon comparing the two, it is clear that the Primary Care data does not have enough variables for it to be a proper predictor for HbA1c's ability to predict complications in different cohorts. In that model, there may be other non-accounted for variables that skew the results (Fig. 3). This is not saying that the CDC HbA1c Alone Model is perfect either; the CDC HbA1c has its own limitation due to the nature of the dataset, using biomarkers as replacements for Risk Scores, and not including glucose levels as compared to the complete model (Fig. 7). However, the CDC HbA1c Model is better at showing HbA1c's ability to glucose's ability to predict complications in specific cohorts and HbA1c's relationship to predicting complications in general.

Figure 11. Results of CDC HbA1c Alone Model

	(1)
VARIABLES	Percentage Increase in Predicting Complications
A1c	0.0211
	(0.0150)
a1c*prediabetes	0.0113**
	(0.00447)
a1c* high glucose	0.0187***
	(0.00442)
Hispanic*a1c	-0.0265**
	(0.0134)
Hispanic	0.247***
	(0.0775)
African*a1c	-0.0205
	(0.0147)
African	0.0234
	(0.0854)
Asian*a1c	-0.0351
	(0.0224)
Asian	0.323**
	(0.128)
Poor*a1c	0.0110
	(0.0105)
poor	-0.0761
	(0.0613)
Old*a1c	-0.0443***
	(0.0113)
old	0.354***
	(0.0685)
Overweight*a1c	-0.00767
	(0.0107)
Medically overweight	0.0807
	(0.0629)
Constant	0.0975
	(0.0831)
Observations	6,050
R-squared	0.045

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: CDC National Center for Health Statistics

Figure 12. Combined Comparisons of CDC HbA1c Alone Model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Hispanic Descent	African Descent	Asian Descent	Obese	Poor	High Glucose	Pre-Diabetes
HbA1c Predictor Coefficients	-0.00545	0.000533	-0.0141	0.0134	0.0320**	0.0398***	0.0324**
	(0.0139)	(0.0145)	(0.0215)	(0.0145)	(0.0152)	(0.0143)	(0.0148)
Observations	6,050	6,050	6,050	6,050	6,050	6,050	6,050

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: CDC National Center for Health Statistics

7.4 Complete HbA1c Model

The Complete HbA1c Model that incorporates glucose shows that HbA1c works as a better predictor when glucose is used along with it. HbA1c's coefficient is statistically significant in the Complete Model and has a coefficient of 0.0391, indicating that increasing HbA1c by 1 percentage point increases the likely hood of complications by 3.91%, which is quite larger than any of the HbA1c Alone model (Fig. 13). As well, the glucose coefficient is not statistically significant, indicating that HbA1c should be used as the predictor in this model. This may be because the cohort interactive terms for this model are used for HbA1c, but the addition of Glucose still makes the HbA1c variable a better predictor than when it was in its own Model. As well, the high glucose interactive variable with HbA1c is significant and negative, -0.0157, ceteris paribus, which is in accord with the previous reports showing that HbA1c works less as a predictor in subjects with high glucose, considered diabetic, on average (Fig. 13). Unfortunately, the lincom regression for the high glucose cohort variable is statistically insignificant which put these results in question. The prediabetic cohort interacting variable is statistically significant in the lincom regression, 0.0373 compared to 0.0391, ceteris paribus, suggesting that HbA1c loses some of its ability as a predictor as well; this is in accordance with the previous studies as well (Fig. 14). The only descendent cohort that had a statistically different HbA1c predictor variable compared to the base was the Hispanic descent cohort, 0.0287 compared to 0.0391, ceteris paribus. All of the

descendent cohorts show a trend for having HbA1c work less as a predictor than the base, but the Hispanic descent cohort was the only one that can be asserted with statistical significance. In the lincom regression the obese cohort has a lower HbA1c coefficient, 0.0388, ceteris paribus, than the base group. This could be due to the obese cohort having higher glucose scores and higher rates of complications on average and may explain why the High Glucose cohort is not statistically significant (Fig. 14).

As well, the closer to poverty line cohort has a higher HbA1c coefficient, 0.0460, ceteris paribus, which may be due to similar reasons discussed in the previous models. Many of the cohort alone identifier variables lost statistical significance as well (Fig. 13 and Fig. 14). This may have arose due to glucose levels, HbA1c levels, and their interaction with HbA1c levels explaining more of the variance between the cohorts compared to the models with less variables.

Figure 13. Results of CDC Complete HbA1c Model

	(1)
VARIABLES	Percentage Increase in Predicting Complications
HbA1c level	0.0391**
	(0.0167)
a1c*prediabetes	-0.00178
	(0.00492)
a1c*high glucose	-0.0157**
	(0.00682)
Hispanic*a1c	-0.0104
	(0.0137)
Hispanic	0.154*
	(0.0794)
African*a1c	-0.0147
	(0.0149)
African	-0.0232
	(0.0867)
Asian*a1c	-0.0191
	(0.0223)
Asian	0.222*
	(0.127)
Poor*a1c	0.00689
	(0.0106)
poor	-0.0533
	(0.0620)
Old*a1c	0.0103***
	(0.00229)
Overweight*a1c	-0.000282
	(0.0107)
Medically overweight	0.0319
	(0.0628)
Glucose	0.000656
	(0.000585)
a1c*glucose	-4.90e-05
	(6.23e-05)
Constant	-0.0248
	(0.0978)
Observations	5,980
R-squared	0.042

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: CDC National Center for Health Statistics

Figure 14. Combined Comparisons of CDC Complete HbA1c Model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Asian Descent	African Descent	Hispanic Descent	Obese	Poor	High Glucose	Pre-Diabetes
HbA1c Predictor Coefficient	0.0200	0.0245	0.0287*	0.0388**	0.0460***	0.0234	0.0373**
	(0.0228)	(0.0170)	(0.0171)	(0.0167)	(0.0175)	(0.0179)	(0.0163)
Observations	5,980	5,980	5,980	5,980	5,980	5,980	5,980

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: CDC National Center for Health Statistics

7.5 Errors in CDC Dataset

There are errors that arise in the CDC dataset due to it being a combination of surveys, laboratory data, and biochemical markers not being a perfect substitute for risk scores. The first error that arises due to the nature of the dataset is that it is not longitudinal; we are trying to predict subject having complications or developing future complications from biochemical markers that are just a snapshot of a subject's health. They may not develop future complications even with a high level of a biochemical marker. Using biochemical markers for Risk Scores is also difficult because we do not have an actual diagnosis report; a subject may have a high level of a biochemical marker that is indicative of complications but it actually arose due to some other preexisting condition. The decided upon levels of a biochemical marker to warrant a subject being considered to have complications or future complications was also chosen upon averages and reported data on that biochemical marker; biomarkers are not exact indicators of complications or the specific diseases ("National Health and Nutrition Examination Survey Standard Biochemistry 2015-16 Data"). A full analysis of a subject's past medical history and characteristics would be needed to be certain that these biochemical levels are indicative for complications in that individual. Also, even though we selected for the biochemical markers that indicate the most common forms of complications associated with diabetes, there are other diseases that were not included and the diseases that were included may affect specific cohorts differently.

With the height, weight, ethnicity, and age being based upon a survey, subjects may have not reported accurately which may have skewed the results. The additional laboratory data, HbA1c test and Biochemical Survey, was voluntary and may have produced a selection bias for those who have known conditions to agree to participate in the laboratory tests because they are used to the test or to see their biochemical levels; this can be seen with the high level of BUN ratio in participants (though many of them could just be very dehydrated) and the high level of fasting in participants (those who have done laboratory tests before know not to eat beforehand) (Fig. 5). As well, due to the high BUN ratio in the population, another regression was performed excluding the BUN ratio for complications (Appendix Fig. 15). The definition of fasting can also skew the data as a non-fasting state tends to be 3 or less hours after a meal and non-fasting tends to be 8 or more hours after a meal. It is hard to determine a normal glucose range between these hours as they tend to change per individual and depends on diet. This makes it hard to classify those who ate between 4 to 7 hours ago in a high glucose or prediabetic cohort.

Lastly, the regression models we have used have been Linear Probability Regression Models. Linear Probability Models always violate homoscedasticity, but can be unbiased. It remains unbiased as homoscedasticity is not needed for determining biasness. However, it is needed for the ordinary least squares model to be the best linear unbiased estimator. Without homoscedasticity, the expected error for this group may not be the actual error for the population. This may cause the regressions to be less powerful than in the actual population or a decreased R^2 value, but nothing that completely invalidates our analysis.

8. Conclusion

From the literature review and preliminary analysis from the Primary Care Network data, further analysis is warranted. The initial research of David Nathan and the new research of Randie Little illustrates some of the possible risks of using HbA1c test exclusively to diagnose and prognose a diverse patient population. This research is incomplete, unfortunately, which my research has illustrated, in some ways, in the

areas that relate to diabetes mellitus with complications diagnosis tracked via HbA1c tests and Glucose tests in a diverse patient population. Hitherto, diabetes research has been focusing on how accurate HbA1c is for diabetes diagnosis and how accurate it predicts normal glucose tolerance. The analyses done for this study were different from previous research as it allowed us to see what the best way to predict future complications from diabetes using different diabetes' prognosis tools and their biases. In addition, the analysis of the biomarkers dataset showed us further strengths and weaknesses in the tools' ability to diagnose specific diseases, depending on the prognosis tool, that has not been seen before due to not focusing on biomarkers or future complications in general and not having to construct a diagnosis variable based upon a biomarker database. The new HCC risk coding system allowed us to perform a simple analysis initially but due to the lack of the full dataset, we had to transition to the CDC data. This initial analysis showed the importance of using as many variables as possible as the small variable size made the results mostly statistically insignificant; it also illustrated the simplicity of using risk scores for future complications compared to biomarkers. It is suggested to use a large dataset like the CDC's but use risk scores for indicators in the future to avoid unneeded work and errors due to the complex relationship many biomarkers have with predicting diseases.

From the CDC data, further and more complete analysis between HbA1c levels in predicting complications or future complications associated with diabetes mellitus is warranted as well. Though the analysis for future complications is not as definitive with risk scores, the biomarker analysis allowed us to see general trends and compare HbA1c's ability as a predictor as compared to a random glucose test and compare HbA1c's ability with and without glucose levels incorporated into its analysis.

The analysis between the Glucose Alone and HbA1c Alone model shows that Glucose and HbA1c share similar trends of biases for predictability in the cohorts of Asian, African, and Hispanic descendants. The analysis also illustrated that glucose has a more linear relationship of predictability than HbA1c when they used as the sole prognosis tool. However, glucose does lose some of its ability as a predictor as a prognosis tool for those of higher glucose values and the comparisons of these two models were complicated due to many of the

variables not being statistically significant and them ignoring the relationship between HbA1c and Glucose. These complications, combined with a comparison of the HbA1c Alone to the HbA1c Complete coefficient for HbA1c levels, illustrate the importance of using multiple diabetes prognosis test to account for the biomarkers' interactions and short comings each test has. The analysis of the CDC data also illustrated another reason to use multiple tests for prognosis of complications; besides the already discussed costs and benefits of diabetes prognosis tests (e.g. time, cost, requirements for patients, etc,), the construction of the complications variable illustrated that some complications that may arise due to diabetes may be predicted better with a glucose test or HbA1c test, respectively, depending on the disease's relationship with the biomarker. For example, albumin is a biomarker for many liver diseases and has a stronger correlation to HbA1c levels than glucose levels (note: albumin had to be removed from the complications variable due to its complex nature that could not be tracked via a dummy variable).

In addition to using multiple prognosis tests for a patient, our analysis also illustrated the importance of considering all attributes of a patient with the prognosis tools. Besides the biases in the cohorts discussed, the smaller model from the Primary Care Data illustrated why a larger dataset and considering more attributes of the patients was needed for a more complete analysis; the smaller model had fewer statistically significant variables on average and did not account for ethnicity or glucose level which has been shown to affect HbA1c's as a predictor in previous studies.

The comparison of the Primary Care to the CDC data showed that a more robust model with more variables from a larger dataset is better for analyzing HbA1c's ability as a prognosis tool and biases in different cohorts. Not having even more variables was a problem in the Complete HbA1c Model as well. Though it had many more variables than the Primary Care Model, its lack of a complete patient medical history is why the results may be skewed and the biomarkers are not a perfect replacement for risk score diagnoses. However, the Complete HbA1c Model still illustrates the importance of using multiple prognosis tools in its comparison to the HbA1c Alone Model, HbA1c loses some of its ability as a prognosis tool for those of prediabetic to diabetic

levels of glucose even with glucose considered, and that every test has its biases for specific cohorts that need to be considered in the future.

The results of the three CDC models does warrant further analysis of HbA1c's and other diabetes prognosis tools' ability as a predictor for complications in a larger data set, comparing more cohorts, and analyzing the relationship between different prognosis tools. Though biomarkers can be used as indicator for complications and future complications, it is suggested to utilize risk scores, as was initially intended, as they are less complicated to analyze and are based upon doctor diagnoses for that patient, i.e. the doctor does the analysis for the patient so the researcher does not have to. Due to COVID-19 this paper had to change greatly, but as a results we learned additional information about HbA1c and other diabetes prognosis tools that would not have been learned using the initial Primary Care Data and tentatively finished the analysis that we originally sought to do. With this knowledge, it is suggested to expand upon it and further analyze the cost and benefits of all prognosis tools for diabetes to better treat and help the patients that suffer from it. We are getting closer to understanding the complete limitations and benefits of the HbA1c test, but we are still not there. Future analysis of the accepted tests of oral glucose tolerance test and random glucose test and new tests of continuous blood glucose test and Fructosamine test is recommended based off of the initial data.

References

"Blood Urea Nitrogen (BUN) Test." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 2 July 2019, www.mayoclinic.org/tests-procedures/blood-urea-nitrogen/about/pac-20384821.

Bonora, Enzo MD, PHD, Tuomilehto, Jakko MD, MA, PHD. "Diabetes The Pros and Cons of Diagnosing Diabetes With A1C." *American Diabetes Association Diabetes Care* vol. 34,2 S184-S190. May. 2011.

Briker, Sara M. , Aduwo1 Jessica Y., Mugeni, Regine, Horlyck-Romanovsky, Margrethe F., W. DuBose, Christopher W., Mabundo, Lilian S., Hormenu, Thomas, Chung, Stephanie T., Ha, Joon, Sherman., Arthur, Sumner, Anne E. "A1C Underperforms as a Diagnostic Test in Africans Even in the Absence of Nutritional Deficiencies, Anemia and Hemoglobinopathies: Insight From the Africans in America Study." Endocrinology, August. 2019.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Standard Biochemistry 2015-16 Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, September 2017, https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/BIOPRO_I.htm#Analytic_Notes.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Glycohemoglobin 2015-16 Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, September 2017, https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/GHB_I.htm.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Demographic Variables and Sample Weights 2015-16 Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, September 2017, https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DEMO_I.htm.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Fasting Questionnaire 2015-16 Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, September 2017, https://wwwn.cdc.gov/nchs/nhanes/2015-2016/fastqx_i.htm.

Chen, Jean MD, Diesburg-Stanwood, Amy, Bodor, Geza, MD, Rasouli, Neda, MD. "Led Astray by Hemoglobin A1c: A Case of Misdiagnosis of Diabetes by Falsely Elevated Hemoglobin A1c." *Journal of Investigative Medicine High Impact Case Reports* vol. 4,1. January. 2016.

Cohen, R. M. Franco, Robert S. , Khera, Paramjit K., Smith, Eric P., Lindsell, Christopher J., \Ciraolo, Peter J., Palascak, Mary B., Joiner, Clinton H. "Red cell life span heterogeneity in hematologically normal people is sufficient to alter HbA1c." *Blood* vol. 112,10. 4284–4291, https://doi.org/10.1182/blood-2008-04-154112 (2008).

"Diabetes." Mayo Clinic, Mayo Foundation for Medical Education and Research, 8 Aug. 2018, www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451.

"High Uric Acid Level Causes." Mayo Clinic, Mayo Foundation for Medical Education and Research, 30 Nov. 2018, www.mayoclinic.org/symptoms/high-uric-acid-level/basics/causes/sym-20050607.

Little, Randie. "Diabetes Blood Tests for People of African, Mediterranean, or Southeast Asian Descent." *National Institute of Diabetes and Digestive and Kidney Diseases*, U.S. Department of Health and Human Services, 1 Oct. 2011, www.niddk.nih.gov/health-information/diagnostic-tests/diabetes-blood-tests-african-mediterranean-southeast-asian.

Little, Randie R, Roberts, Williams "A review of variant hemoglobins interfering with hemoglobin A1c measurement." *Journal of diabetes science and technology* vol. 3,3 446- 51. 1 May. 2009, doi:10.1177/193229680900300307

"Liver Function Tests." Mayo Clinic, Mayo Foundation for Medical Education and Research, 13 June 2019, www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595.

Iskandar, S., Migahid, A., Kamal, D. et al. "Glycated hemoglobin versus oral glucose tolerance test in the identification of subjects with prediabetes in Qatari population." BMC Endocr Disord (2019) 19: 87. https://doi.org/10.1186/s12902-019-0412-1

Nathan, David M, "International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes." *American Diabetes Association Diabetes Care* vol. 32, 7 1327- 1334. July. 2009.

Park, Paul H., Pastakia, Sonak D."Access to Hemoglobin A1c in Rural Africa: A Difficult Reality with Severe Consequences," Journal of Diabetes Research, vol. 2018, Article ID 6093595, 5 pages, 2018. https://doi.org/10.1155/2018/6093595.

"Primary Care Dataset." General Physicians PC, October 2019.

Picón, María José, Murri, Mora, Muñoz, Araceli, Fernández-García, José Carlos, Gomez- Huelgas, Ricardo, Tinahones, Francisco J. "Hemoglobin A1c Versus Oral Glucose Tolerance Test in Postpartum Diabetes Screening." *American Diabetes Association Diabetes Care* vol. 35,8 1648-1653. August. 2012.

Rowley, William R et al. "Diabetes 2030: Insights from Yesterday, Today, and Future Trends." *Population health management* vol. 20,1 (2017): 6-12. doi:10.1089/pop.2015.0181

Shao, Hui , Rolka, Deborah B., Gregg, Edward W., Zhang, Ping. "Influence of Diabetes Complications on the Cost-Effectiveness of A1C Treatment Goals in Older U.S. Adults." *American Diabetes Association* vol. 67,s1. July. 2018.

Tello, Monique. "Rethinking A1c goals for type 2 diabetes." *Harvard Medical School, Harvard Health Publishing.* March. 2018.

"Test ID: CK Creatine Kinase (CK), Serum." *CK - Overview: Creatine Kinase (CK), Serum, Mayo Clinic*, Mayo Foundation for Medical Education and Research, www.mayocliniclabs.com/test-catalog/Overview/8336.

"Test ID: HCO3 Bicarbonate, Serum." *HCO3 - Clinical: Bicarbonate, Serum*, Mayo Foundation for Medical Education and Research, www.mayocliniclabs.com/test-catalog/Clinical and Interpretive/876.

Villacreses, Maria Mercedes Chang, Feng, Wei, Karnchanasorn, Rudruidee, Samoa, Raynald, Chiu, Ken, SAT-125 Underestimation of the Prevalence of Diabetes and Overestimation of the Prevalence of Glucose Tolerance by Using Hemoglobin A1c Criteria, *Journal of the Endocrine Society*, Volume 3, Issue Supplement_1, April-May 2019, SAT–125, https://doi.org/10.1210/js.2019-SAT-125.

Virtue, Mark A., MD, Furne, Julie K., BS, Nuttall, Frank Q., MD, PHD, Levitt, Michael D. MD. "Relationship Between GHb Concentration and Erythrocyte Survival Determined From Breath Carbon Monoxide Concentration." *American Diabetes Association Diabetes Care* vol. 27,4 931-935. April. 2004.

Yi, Whitley M., PharmD, Jones, Emily M., PharmD, Hansen, B. Kyle, PharmD, Vora, Jay PharmD. "The Impact of Self-Monitoring Blood Glucose Adherence On Glycemic Goal Attainment in an Indigent Population, With Pharmacy Assistance." *P&T Community* vol. 44,9 554- 559. September. 2019.

Appendix

Appendix Table 1 : CDC Results without BUN Ratio

This Figure Presents the Results from the CDC regression excluding the BUN Ratio for the Complete Regression Model. The decreased coefficients show the effect the ratio had. The coefficients have the same sign as the original regression, but fewer of them are statistically significant and even less economically significant.

Figure 15. Appendix: Complications Regression Excluding BUN Ratio

VARIABLES	Complications (no BUN Ratio)
HbA1c level	0.0121*
	(0.00730)
a1c*prediabetes group	-0.00171
	(0.00215)
a1c*high glucose group	-0.00454
	(0.00298)
Hispanic*a1c	-0.00790
	(0.00599)
Hispanic descent	0.0529
	(0.0347)
African*a1c	-0.00582
	(0.00652)
African descent	0.0415
	(0.0379)
Asian*a1c	-0.000906
	(0.00976)
Asian descent	0.00257
	(0.0557)
Poverty*a1c	-0.000390
	(0.00464)
poor	0.0126
	(0.0271)
Old group* a1c	0.000270
	(0.00100)
Overweight group* a1c	0.00162
	(0.00469)
overweight	0.00929
	(0.0275)
Glucose level	0.00102***
	(0.000256)
a1c*glucose	-6.96e-05**
	(2.73e-05)
Constant	-0.106**
	(0.0427)
Observations	5,980
R-squared	0.010

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: CDC National Center for Health Statistics

Appendix Table 2: Analyzing correlation between endogenous variables (chi squared table)

This Figure Presents the analysis for the comparison of the relationship between endogenous variables using t-tests and χ^2 tests. It shows the difference in means for the relevant variables and their p-values based off of the t-tests and χ^2 tests.

Figure 16. Appendix: Variables for the CDC Dataset Analyzed with P-values.

	NO COMPLICATIONS	COMPLICATIONS	P-VALUE
N	4354	5300	
Glucose, Refrigerated Serum (Mg/dl), Mean (SD)	101.0 (39.0)	105.6 (43.7)	<0.001
Glycohemoglobin (%), Mean (SD)	5.7 (1.1)	5.9 (1.2)	<0.001
BMI, Mean (SD)	28.3 (6.6)	27.9 (6.9)	0.024
Total Length Of 'Food Fast', Hours, Median (IQR)	11 (10, 13)	11 (10, 13)	0.12
Non-Fasting High Glucose Group	147 (3.4%)	3733 (70.4%)	<0.001
Fasting High Glucose Group	21 (0.5%)	44 (0.8%)	0.038
Non-Fasting Pre-Diabetic Levels Of Glucose Group	137 (3.1%)	121 (2.3%)	0.009
Fasting Pre-Diabetic Levels Of Glucose Group)	20 (0.5%)	44 (0.8%)	0.025
Race(Hispanic)	1189 (27.3%)	1850 (34.9%)	<0.001
Race(Black)	1068 (24.5%)	1010 (19.1%)	<0.001
Race(Asian)	445 (10.2%)	572 (10.8%)	0.36
Old (>60)	1093 (25.1%)	731 (13.8%)	<0.001
Poor (Household Earnings >2 * Poverty Level)	1994 (45.8%)	2641 (49.8%)	<0.001
Medically Obese	1692 (38.9%)	3901 (73.6%)	<0.001

Source: CDC National Center for Health Statistics