2023

Association of Diabetes Prevalence with Labor Force Participation Rates in Kentucky Counties

Ryan Montgomery
University of Kentucky, ryanmontgomery379@gmail.com

Association of Diabetes Prevalence with Labor Force Participation Rates in Kentucky Counties

Ryan Montgomery

University of Kentucky

Martin School of Public Policy and Administration

Spring 2023

Faculty Advisor:

Dr. Karen Blumenschein

Table of Contents

Executive Summary

Background:

Compared to other states in the USA, the Commonwealth of Kentucky has one of the highest rates of diagnosed diabetes and one of the lowest levels of labor force participation. While many disparate factors likely contribute to each of these problems, evidence suggests that diabetes contributes to poor economic outcomes, including reduced participation in the workforce. Although previous studies have sought to measure various economic impacts of diabetes, both in Kentucky and the wider world, there does not yet exist research that estimates the correlation between diabetes prevalence and labor force participation in Kentucky communities.

Purpose:

This project seeks to estimate the strength of the relationship between county-level diabetes prevalence and labor force participation rates across Kentucky's 120 counties. While it does not establish causality or the direction of any causal relationship, it does offer a benchmark by which the diabetes epidemic can be understood both as an issue afflicting the health of Kentucky's people and as a pressing concern for the Commonwealth's economic well-being.

Research Design:

This project employs a multivariable linear regression model to estimate the correlation between county-level diabetes prevalence and county-level labor force participation rates in Kentucky counties from 2015-2019, using publicly available data from the Centers for Disease Control's Behavioral Risk Factor Surveillance System (BRFSS) and the U.S. Census Bureau's American Community Survey (ACS). Labor force participation rates serve as the dependent variable and diabetes prevalence rates serve as the primary independent variable, alongside nine other county-level variables meant to control for likely confounding factors: urban/rural categorization, mean age, sex ratio, proportion of the population aged 65 years or older, race, mean educational attainment, mean household income, poverty rate, and disability rate. Various secondary analyses and sensitivity analyses were performed and are discussed.

Findings:

The correlation between diabetes prevalence and labor force participation rates in Kentucky counties was estimated at

-0.425 (95% CI: -0.779, -0.0717). In other words, for every 1% increase in its diabetes prevalence, an average Kentucky county would expect to see its labor force participation rate decrease by 0.425%. Secondary analyses consistently found a negative relationship between diabetes prevalence and labor force participation rates, although some analyses returned diminished and non-significant results. Sensitivity analyses demonstrated that the relationship could lose statistical significance if the ACS labor force participation rate estimates used to build the model were significantly overestimated.

Future Directions:

More research is needed to further elucidate the relationship between labor force participation and diabetes and to investigate any specific causal mechanisms that may connect the two measures. The approach taken in this project could be applied to other geographical areas or to counties in the United States overall. Quasi-experimental research designs could estimate the causal directionality of the relationship between diabetes and labor force participation. Studies could explore the relative impacts of diabetes-associated disability, caregiver drain, unemployment-driven health outcomes, and even perverse healthcare system incentives in explaining the correlation observed in this project.

INTRODUCTION

With a high prevalence of diabetes and a low labor force participation rate[1], the Commonwealth of Kentucky faces interconnected healthcare and economic challenges. The state-wide prevalence of diabetes has risen precipitously in the last two decades, nearly doubling from 6.5% in 2000 to 13.3% in 2019 (KY CHFS et al., 2021). Meanwhile, Kentucky's labor force participation rate was estimated in 2019 to be a mere 58.2%, tied for the seventh lowest rate in the country (FRED, 2019). Prior research has linked diabetes to worsened economic outcomes, including decreased employment rates, lower earned incomes, and reduced time spent working (Pedron et al., 2019; Seuring et al., 2015). The progression of the disease can lead to complications that severely impact individuals' ability to work, including blindness and amputation (American Diabetes Association, 2023). It is therefore reasonable to suppose that Kentucky's high prevalence of diabetes and its low level of workforce participation could be related. However, there currently exists no research estimating the correlation of these two factors in Kentucky communities. The present study seeks to fill that void by making use of county-level estimates published by the Centers for Disease Control and the U.S. Census Bureau. A multivariable linear regression model is employed to estimate the correlation between diabetes prevalence rates and labor force participation rates in Kentucky counties from 2015-2019 while controlling for nine potential confounding variables.

[1] *The labor force participation rate is defined as the proportion of the non-institutionalized civilian population that is currently working or looking for work. It is usually calculated for the portion of population aged 16 years and older.*

LITERATURE REVIEW

Section 1: Diabetes as a Disease

Diabetes Type 1 and Type 2

Diabetes mellitus is a progressive disease characterized by impaired insulin signaling. Type I Diabetes Mellitus, the less common form of the disease, occurs when the pancreas loses the ability to make insulin. It is usually diagnosed early in the lifespan and is thought to be caused by autoimmune reactions linked to genetic factors and viral infections. Type II Diabetes Mellitus (T2DM) arises when the body loses the ability to respond appropriately to the insulin that it does produce, and more commonly occurs later in the lifespan as damage to the endocrine system accumulates (Ozougwu et al., 2013). In the United States, Type II Diabetes accounts for around 90% of diabetes diagnoses (U.S. Centers for Disease Control and Prevention, 2022).

Causes and Consequences of Type 2 Diabetes

Excessive caloric intake and sedentary lifestyle are key risk factors for developing both T2DM and obesity. As consumed calories exceed the body's energy output, chronically high levels of blood sugar can lead to insulin resistance, leading to dysregulation of glucose metabolism (Maggio & Pi-Sunyer, 2003). Once significant damage to the insulin signaling system has accumulated, reversing or curing diabetes is often impossible. Treatment instead focuses on managing blood sugar levels in order to slow the progression of the disease and delay the development of complications. Diabetes-related health complications include severe and life-threatening conditions, such as stroke, heart disease, kidney failure, blindness, opportunistic infections, and death (American Diabetes Association, 2023).

Recent decades have seen the emergence of a dual epidemic of obesity and diabetes. According to one estimate, the total prevalence of diabetes in the USA increased from 7.7% in 1999-2000 to 13.3% in 2015-2016, with an accompanying rise in obesity cited as a principal cause (Fang, 2018). The CDC estimates that approximately a quarter of adults with diabetes remain undiagnosed, meaning that the true prevalence of the disease is even higher than diagnosis rates suggest.

Kentucky has been affected particularly badly by the diabetes epidemic. The proportion of adults diagnosed with diabetes rose from 6.5% in 2000 to 13.7% in 2018, one of the highest rates in the nation. The worst-struck regions in the Commonwealth have diagnosed diabetes rates over 20%, including in the eastern Cumberland Valley, Kentucky River, and

Big Sandy regions (Centers for Disease Control and Prevention, 2000-2018). Over 1,500 Kentuckians died from diabetes in 2020, the 13th highest rate of diabetes-caused mortality in the country (Centers for Disease Control and Prevention, 2022).

Section 2: Diabetes as an Economic Determinant

The diabetes epidemic incurs a heavy economic burden, both for the USA overall and in Kentucky specifically. The American Diabetes Association estimates that diabetes resulted in $237 billion in healthcare costs and $90 billion in lost earnings nationwide in 2017 alone (American Diabetes Association, 2018). Research has shown that diabetes is negatively associated with economic factors such as employment, earnings, and worker productivity (Pedron et al., 2019; Seuring et al., 2015). This makes Kentucky's epidemic diabetes rate an attractive explanation for the Commonwealth's low labor force participation rate, which at 57.5% ranks at the 7th lowest in the nation as of 2023 (United States Bureau of Labor Statistics, 2023).

However, different studies have found varying and even contradictory results as far as the magnitude of diabetes-related economic effects and the degree to which different populations experience them. Additionally, it is inadvisable to draw direct, unambiguous causal relationships between diabetes and economic factors, since much of the available evidence is correlational (Seuring et al. 2015; Pedron, et al. 2019).

Seuring et al. (2015) – The Economic Costs of Type 2 Diabetes

A 2015 systematic review by Seuring et al. compiled 109 studies assessing the economic burden of diabetes published globally from January 2001 to October 2014. Of the 109 included studies, 86 were characterized as cost-of-illness studies and 23 were classified as labor market studies. The direct and indirect costs of diabetes varied significantly between nations, and different studies often showed ambiguous or even contradictory economic effects of diabetes in different groups, particularly in men and women. For example, one study found no significant relationship between diabetes and employment outcomes for Australian women (Zhang et al., 2009) while another found a nearly 45% decrease in employment for women with diabetes in the United States (Minor, 2011). These disparities suggest that the economic effects of diabetes vary based on the context in which they are measured, limiting the generalizability of any one study to the overall population. However, common patterns and areas of concordance are revealed by the review. Most studies included in the review estimated that individuals with diabetes are between 3% to 10% less likely to be employed. The studies were also consistent in showing a negative association between diabetes and earned incomes, although again with highly heterogeneous conclusions. Similarly, diabetes was consistently linked to reduced time spent working among individuals who were employed (Seuring et al., 2015).

In the same way that the negative health complications of diabetes worsen as the disease progresses, so do its economic complications appear to worsen with time. A study by Tunceli et al., (2009) shows how economic disparities widen between people diagnosed with diabetes and the non-diabetic population with advancing age. In the 20-44 year age range, individuals with diabetes had a 1.2% higher disability rate, were 3.1% more likely to report having health-related work limitations, and had a 3.4% lower chance of being employed, compared to those without diabetes. These disparities were significantly greater for the 45-64 age range, where diabetes diagnosis predicted a 3.4% higher disability rate, 5.7% more health-related work limitations, and 8.1% lower chances of employment. The results of this study are corroborated by the work of Minor (2013), which found that men with diabetes began experiencing reduced incomes around 6 to 10 years following their diagnosis. These data suggest that interventions that delay the progression of diabetes-related health complications might also slow the economic damage it incurs.

Clark et al. (2019) – The Economic Impact of Diabetes in Kentucky

A 2019 study by Clark et al. offers the most comprehensive available estimates of the economic impact of diabetes specific to the Commonwealth of Kentucky, including decreased employment. Drawing on data from the Kentucky Behavioral Risk Factors Surveillance System, the study noted that while 54% of all Kentucky adults were employed in 2016, only 26% of adults with diabetes were working during the same time frame. However, diabetes prevalence is higher among groups who also experience reduced employment, regardless of diabetes diagnosis. This complicates efforts to estimate the effects of diabetes itself. Accordingly, the researchers developed a logistic regression model to estimate the impact of diabetes on employment while controlling for gender, race, education, state of residence, and other health conditions that could also potentially affect employment. This model found that diabetes reduced adult employment by 3.1% for men and 4.2% for women, with the greatest effect size found for the middle-aged population. This equated to a loss of 15,700 workers in Kentucky, which Clark et al. estimated resulted in lost worker earnings of $551.3 million and reduced tax revenue of $33.1 million in 2016 alone.

Clark, et al. note that their conclusions are limited by the likely presence of complicated causal relationships between diabetes and labor market outcomes, a problem echoed across much of the available literature on the subject. For example, it is probable that loss of employment may contribute to the development and progression of diabetes at the same time that diabetes contributes to loss of employment and reduced hours worked. The cost of drug therapies such as insulin is frequently identified as a key barrier to effective diabetes self-management (Adu et al., 2019; Mogre et al., 2019); accordingly, diabetic individuals

with reduced employment or income may be at risk of faster disease progression because they cannot afford treatment. A 2019 Dutch study by Herber, et al. showed that loss of employment is correlated with poor mental and physical health, including obesity, with health effects appearing particularly strong for individuals experiencing chronic unemployment of 5 years or longer. If this pattern holds true in Kentucky, it could help explain the especially high incidence of diabetes in economically disadvantaged regions of the Commonwealth. However, the likely bi-directional relationship between diabetes and reduced employment complicates any attempt to estimate the magnitude of one variable's effect upon the other.

Pedron et al. (2019) - The Impact of Diabetes on Labor Market Participation

The problem of "reverse causality" between diabetes and labor market outcomes is given special attention in a systematic review by Pedron et al., (2019). The authors selected 30 studies and identified four main labor participation outcomes potentially affected by diabetes: absence of employment, unemployment, early retirement, and disability pension. They note that the reliance of previous cost-of-illness studies on focusing on productivity losses caused by morbidity and mortality likely leads to underestimating the real effects of diabetes on the labor market, since that methodology fails to account for potentially lower workforce participation rates among individuals with diabetes compared to the general population. Included studies showed a significant negative association between diabetes and employment, but as in the 2015 review by Seuring et al., the magnitude of this effect varied greatly between studies and sample populations.

Pedron et al. (2019) found that the overall evidence supported statistically significant reductions in employment for both men and women due to diabetes, with men generally being more greatly affected. Seven cross-sectional studies were identified that employed statistical methods to test for endogeneity of diabetes while estimating the disease's effect on employment. In other words, these seven studies employed various statistical techniques to attempt to pick apart the effects of diabetes itself from confounding variables associated with both diabetes and employment. Such methods included instrumental variables (IVs) and multivariate models. However, there was no clear consensus between studies regarding the presence of endogeneity or its magnitude if it does exist. One standout study found that people suffering both from diabetes and other comorbid health conditions had a 12% lower chance of participating in the labor force compared to those with diabetes but no other comorbidities (Ng et al., 2001).

Literature Review Summary

The overall picture painted by global systematic reviews is that diabetes seems to reduce participation in the labor force, as well as time spent working and income earned (Seuring et al., 2015; Pedron, et al., 2019). The work of Clark, et al. (2019) offers valuable estimates of the yearly diabetes-associated reduction in employment in Kentucky. However, there does not appear to exist any previous study which attempts to correlate diabetes prevalence and labor force participation rates on a county-by-county level. Despite the inability of a correlational model to test causal inferences, such a project would expand the existing knowledge of how diabetes prevalence and labor market outcomes are intertwined.

RESEARCH DESIGN

Summary and Variables

This project comprises a quantitative, retrospective, correlational analysis of the relationship between diabetes prevalence and workforce participation rates in Kentucky counties using estimated 2015-2019 5-year averages for both primary variables (i.e. county-level diabetes prevalence and county-level labor force participation rates). Labor force participation rate is considered the dependent variable, while diabetes is considered the primary independent variable. A multivariable linear regression model was constructed to measure the correlation between diabetes and labor force participation while controlling for nine potentially confounding variables: mean age, proportion of the population 65 years or older, sex ratio, race, mean educational attainment, mean household income, poverty rate, disability rate, and the county's urban/rural classification. A summary and description of each variable used in the regression can be found in **Table 4** in Appendix II**.** The equation representing the multivariable regression model used for the primary statistical analysis is expressed in **Figure 1.**

Hypothesis

The project's null hypothesis predicted that its model would find no significant correlation between county-level diabetes rates and county labor force participation rates after controlling for confounding variables. The alternative hypothesis predicted that counties with higher diabetes rates would exhibit significantly lower labor force participation rates, even after accounting for confounding factors.

Data Sources

The project employs data compiled and made available for public use by the U.S. Census Bureau and the Centers for Disease Control. The 2019 American Community Survey (ACS) 5-year estimates provide county-level labor force participation rates, as well as the economic, social, and demographic information used for eight of the nine control variables. While newer 5-year estimates have recently become available, the decision was made to use the 2015-2019 estimates to avoid potential distortions in the data arising from the COVID-19 Pandemic. The Behavioral Risk Factor Surveillance System (BRFSS), administered by the CDC, provides estimates for county-level diabetes prevalence rates in Kentucky. These data are provided in annual reports. For this study, the point estimates of diabetes prevalence from each year from 2015-2019 were averaged for each county in Kentucky, creating a 5-year average matching the time frame of the 2015-2019 ACS 5-year estimates. The National

Center for Healthcare Statistics provided the final data source in the form of its 2013 Urban-Rural Classification Scheme for Counties, from which the variable of urban/rural classification was created.

Statistical Assumptions

Constructing a multivariable regression model involves making several assumptions about the data used to create the model, as well as the residuals and predicted values that the model generates. These assumptions include *independence of observations*, *linearity*, *normality, homoscedasticity*, *non-multicollinearity*, and the *absence of outlier effects* (Tranmer et al., 2020). The failure to satisfy one or more of these assumptions undermines the statistical validity of the model and thus the interpretation of its results. A complete description of the methods used to test each assumption is found in **Table 7** in Appendix II, the findings of those tests are revealed in the Results section, and their bearing on the project's conclusions is explicated in the Discussion section. Observed or likely violations of statistical assumptions were addressed through several secondary analyses which attempted to estimate the bearing of the violations on the project's findings.

Primary Analysis

The primary analysis evaluates the correlation between diabetes and labor force participation rates, including all nine of the control variables and using unmodified point estimates for each variable. The equation representing the multivariable regression model used in the primary analysis is expressed in **Figure 1**, below.

Figure 1: Multivariable Linear Regression Equation

$$LaborForce = ß0 + ß(Diabetes) + ß(UrbanRural) + ß(Age) + ß(Education) + ß(Male) + ß(White) + ß(SixtyFivePlus) + ß(Disability) + ß(Poverty) + ß(Income) + ε$$

In **Figure 1**, ß0 represents the Y-intercept, ε represents the error term, and ß(variable) represents the correlation coefficient for the respective independent variable as it relates to the dependent variable, labor force participation rate. The greater the absolute value of a coefficient, the more strongly the variable predicts counties' labor force participation rates. Positive coefficients mean that a variable varies directly with workforce participation, whereas a negative coefficient means that as a variable increases, workforce participation falls.

Secondary Analyses

In addition to the primary analysis described above, the regression was repeated under several alternative conditions as part of six secondary analyses. First, the regression was repeated with the variable of county-level disability rates removed, in order to estimate the contribution of disability to the relationship between diabetes and labor force participation. Secondly, average household income was removed from the regression as it was observed to have the highest level of multicollinearity with the other variables in the model, particularly educational attainment (see **Table 7**). Thirdly, the regression was repeated in the absence of the mean age variable, due to the concern that mean age and the proportion of county residents aged 65 or olde could be redundant measures. Fourthly, the variables representing mean age, male sex ratio, and non-Hispanic mono-racial white ethnicity were excluded, as these variables did not exhibit linear relationships with labor force participation when examined outside the multivariable model. Fifthly, the regression was repeated while excluding both mean age and household income, in order to test the model while excluding the most problematic variables (mean age being highly correlated with the proportion of residents aged 65 or older, and household income exhibiting the highest level of multicollinearity). Lastly, counties ranking among the 5.5% highest or lowest rates of labor force participation and/or diabetes prevalence were excluded from the data set[2], along with three counties with particularly extreme levels of education, income, and labor force participation: Clay, Oldham, and Wolfe counties. In this final analysis, the regression was again repeated with potential outliers removed.

Sensitivity Analyses

The estimates yielded by the primary and secondary analyses in this project rely on previous ACS and BRFSS estimates. Rudimentary sensitivity analyses were conducted to examine the robustness of the model's findings assuming conditions of significant error in the estimated county-level labor force participation rates and diabetes prevalence rates. The 2015-2019 ACS data provide point-estimates of labor force participation rates with margins of error corresponding to the 90% CI of each estimate. Likewise, annual BRFSS data provides point estimates as well as upper and lower limits of county-level diabetes prevalence, again demarcating the 90% confidence interval.

[2] *A cutoff of 5.5% was chosen because it corresponded to the six counties with the most extreme estimates for diabetes prevalence and/or labor force participation. This was the closest possible option to excluding the most extreme 10% of the overall data distribution.*

Four new variables were created, corresponding to the average upper and lower limits of each county-level estimate for diabetes prevalence and labor force participation rate. Alongside the point-estimates used in the primary analysis, this process yielded nine total combinations of the six variables, which were tested with a series of eight corresponding sensitivity regressions[3]. In each sensitivity regression, the labor force variable (whether the upper limit, lower limit, or point estimate) served as the dependent variable, while the diabetes variable (whether the upper limit, lower limit, or point estimate) served as the independent variable. In each instance, all nine control variables from the primary analysis were included as independent variables. **Table 1** below summarizes the variables utilized in the eight sensitivity analyses.

Table 1: Variables used in sensitivity analyses

Dependent Variables	*Independent Variables (Primary)*	*Independent Variables (controls)*
Labor Force Participation Rate (Upper Limit)	Diabetes Prevalence (Upper Limit)	• UrbanRural • Age • Male • SixtyFivePlus • White • Education • Income • Poverty • Disability
Labor Force Participation Rate (Point Estimate)	Diabetes Prevalence (Point Estimate)	
Labor Force Participation Rate (Lower Limit)	Diabetes Prevalence (Lower Limit)	

[3] *One combination (diabetes point estimate vs/ labor force point estimate) was identical to the primary analysis and therefore not repeated as a sensitivity analysis.*

RESULTS

Statistical Assumptions

Multiple statistical assumptions of a multivariable linear regression model were found to have been potentially violated or deemed impossible to assess. The nature of using secondary data from the Census Bureau and CDC made testing for *independence of observations* unfeasible, as an exhaustive analysis of the data collection methods and statistical weighting used by those entities was beyond the scope of this project. When testing *linearity*, graph matrix visualization suggested that the variables representing age and sex ratio (and to a lesser extent race) did not vary linearly with labor force participation when viewed as individual variables. Furthermore, a significant p value obtained from a RESET (Ramsey Regression Equation Specification Error Test) indicated a possible non-linear relationship between labor force participation and the independent variables as a group.

While visualization of a residual-versus-fitted plot did not reveal any violation of the *normality* assumption, a Shapiro-Wilk test for normality revealed that the residuals of the model may not be normally distributed. A Breusch-Pagan test for *homoscedasticity* was successful, indicating that the residuals of the regression did not have significantly unequal variance for different values of fitted values and of predictors. This means that the regression should return accurate standard errors[4]. While no formal test was conducted for the *independence* assumption, it can be reasonably assumed that this assumption is violated. In order to satisfy it, there should be no remaining association between the observed data points that is not accounted for by the predictors in the regression model. This project does not presume to perfectly model county-level labor force participation, which is almost certainly a highly endogenous variable influenced by factors not captured in the regression model presented here.

A VIF (Variance Inflation Factor) test was employed to test the assumption of *no multicollinearity,* which states that each independent variable should constitute a unique predictor, rather than overlapping with information contained by another independent variable The VIF test revealed that the variable representing mean household income had a high degree of multicollinearity with the other independent variables (VIF >10). Four other variables exhibited a moderate degree of

[4] *The standard errors and 95% confidence intervals calculated in this project are inaccurate for a different reason: its lack of error propagation for the uncertainties present in the original ACS and BRFSS estimates. This problem receives further consideration in the Discussion section.*

multicollinearity: mean age, poverty rate, proportion aged 65 years or older, and education (10 > VIF > 5). Finally, visualization of the data revealed the potential presence of extreme values that could violate the assumption of *no outlier effects*. These included counties with dramatically higher or lower labor force participation rates, diabetes rates, and other economic measures compared to the rest of the data set.

To address the likely violations of the *linearity, no multicollinearity, and no outlier effects* assumptions, secondary analyses were conducted to test whether adjusting the project's regression model to account for these violations would significantly alter its findings.

Primary Analysis

The correlation between county-level diabetes prevalence and labor force participation rates was estimated at -0.425 (95% CI: -0.779, -0.072) while controlling for all nine potential confounders. In other words, the model predicts that for an average Kentucky county, each 1% increase in the prevalence of diagnosed diabetes is accompanied by a 0.425% *decrease* in the county's labor force participation rate. The 95% confidence interval does not cross zero, indicating that the negative relationship between diabetes and labor force participation is statistically significant. The R^2 value of the regression is 0.907, which (at face value) implies that over 90% of the observed variance is accounted for by the model. **Figure 2** contains the full output of the regression as run in the statistical software package Stata.

Figure 2: Primary regression Stata output

```
. regress LaborForce Diabetes UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability
```

Source	SS	df	MS	Number of obs	=	120
				F(10, 109)	=	106.45
Model	8005.09388	10	800.509388	Prob > F	=	0.0000
Residual	819.70937	109	7.52026945	R-squared	=	0.9071
				Adj R-squared	=	0.8986
Total	8824.80325	119	74.1580105	Root MSE	=	2.7423

LaborForce	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Diabetes	-.4253408	.1784327	-2.38	0.019	-.7789885	-.0716931
UrbanRural	.5868778	.2591184	2.26	0.025	.0733135	1.100442
Age	-.034805	.2227117	-0.16	0.876	-.4762123	.4066024
Male	-.3523765	.0409735	-8.60	0.000	-.4335846	-.2711684
SixtyFivePlus	-.9879941	.2696092	-3.66	0.000	-1.522351	-.4536375
White	-.0672284	.0573761	-1.17	0.244	-.1809461	.0464892
Education	-.0971857	2.443786	-0.04	0.968	-4.94069	4.746318
Income	-.0000336	.0000799	-0.42	0.675	-.000192	.0001249
Poverty	-.7319414	.0929914	-7.87	0.000	-.9162473	-.5476354
Disability	-.1832703	.0824768	-2.22	0.028	-.3467367	-.0198039
_cons	136.6695	11.50718	11.88	0.000	113.8626	159.4763

Secondary Analyses

Six secondary analyses were conducted to test the regression model under several different circumstances, as summarized in **Table 2**. Five of the six analyses were performed in response to observed violations of statistical assumptions inherent in the regression model. The remaining analysis, in which disability rates were excluded from the regression, is of greater practical interest. When excluding disability rates as a control variable, the observed coefficient of correlation of diabetes prevalence on labor force participation rate grew in magnitude to -0.525 (-0.873, -0.176), compared to the estimated coefficient of -0.425 (-0.779, -0.077) obtained in the primary analysis. This result suggests that disability mediates the relationship between diabetes and labor force participation but does not fully explain it.

Four of the five secondary analyses addressing violated statistical assumptions yielded broadly similar findings compared to the primary analysis. However, the iteration of the regression which excluded mean age, male ratio, and race generated a coefficient of -0.282 (-0.720, +0.156) for the correlation of diabetes prevalence and labor force participation rate. These variables were each observed to not exhibit an independent linear relationship with labor force participation, but when they were removed from the model the results of the regression lost statistical significance.

Table 2: Summary of Secondary Analyses

Secondary Regression	*Purpose of Secondary Regression*	*Included Control Independent Variables*	*Coefficient of diabetes prevalence on labor force participation rate (95% CI)*	*R^2 value*
1: Excluding disability rates	Estimating mediating effect of disability on correlation between diabetes and labor force participation	UrbanRural Age Male SixtyFivePlus White Education Income Poverty	-0.525 (-0.873, -0.176)	0.903
2: Excluding household income	Income variable highly multicollinear with other independent variables (VIF > 10)	UrbanRural Age Male SixtyFivePlus White Education Poverty Disability	-0.409 (-0.753, -0.065)	0.907
3: Excluding mean age	Age variable highly correlated with SixtyFivePlus variable	UrbanRural Male SixtyFivePlus White Education Income Poverty Disability	-0.429 (-0.778, -0.080)	0.907
4: Excluding mean age, male ratio, and proportion white race	Age, Male, and White variables are not linearly correlated to dependent variable (LaborForce).	UrbanRural SixtyFivePlus Education Income Poverty Disability	-0.282 (-0.720, +0.156)	0.843
5: Excluding mean age and household income	Assessing regression output in the absence of the two most potentially problematic control variables.	UrbanRural Male SixtyFivePlus White Education Poverty Disability	-0.410 (-0.752, -0.067)	0.907
6: Excluding counties in the top/bottom 5% for labor force participation or diabetes prevalence.	Assessing the strength of the primary analysis after excluding potential outliers.	UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.556 (-0.992, -0.120)	0.877

Sensitivity Analyses

Eight sensitivity regressions were performed to test the sensitivity of the study's conclusions to large hypothetical errors in the ACS labor force participation rate estimates and BRFSS diabetes prevalence estimates. The results of these analyses is summarized in **Table 3** below. In the three conditions where the lower limit of the labor force participation rate estimates was used, the correlation coefficient between the two primary variables was significantly reduced and failed to reach statistical significance. These findings are interpreted in greater detail in the Discussion section below.

Table 3: Summary of Sensitivity Analysis Regressions

Sensitivity Analysis Regression #	*Dependent Variable*	*Primary Independent Variable*	*Coefficient of Diabetes on Labor Force Participation (95% CI)*	*R^2 Value*
1	Labor Force (Upper Limit)	Diabetes Prevalence (Upper Limit)	-0.623 (-0.997, -0.249)	0.894
2	Labor Force (Point Estimate)	Diabetes Prevalence (Upper Limit)	-0.433 (-0.808, -0.058)	0.907
3	Labor Force (Lower Limit)	Diabetes Prevalence (Upper Limit)	-0.243 (-0.637, +0.150)	0.911
4	Labor Force (Upper Limit)	Diabetes Prevalence (Point Estimate)	-0.658 (-1.00, -0.307)	0.897
N/A *(identical to primary analysis)*	Labor Force (Point Estimate)	Diabetes Prevalence (Point Estimate)	-0.425 (-0.779,-0.072)	0.907
5	Labor Force (Lower Limit)	Diabetes Prevalence (Point Estimate)	-0.195 (-0.568, +0.178)	0.910
6	Labor Force (Upper Limit)	Diabetes Prevalence (Lower Limit)	-0.666 (-1.00, -0.331)	0.898
7	Labor Force (Point Estimate)	Diabetes Prevalence (Lower Limit)	-0.424 (-0.765, -0.082)	0.907
8	Labor Force (Lower Limit)	Diabetes Prevalence (Lower Limit)	-0.181 (-0.542, +0.179)	0.910

DISCUSSION

Interpretation of Results

This project's primary and secondary analyses consistently demonstrate a negative relationship between diabetes prevalence and labor force participation rates in Kentucky counties. In both the primary analysis and the majority of the secondary analyses, this correlation was both statistically significant and practically meaningful. These findings reinforce previous research that connects diabetes to negative economic outcomes, and provide novel estimates specific to Kentucky communities. If a 1% increase in a Kentucky county's diabetes prevalence is associated with a 0.425% decrease in its labor force participation rate, then interventions aimed at preventing or managing diabetes might plausibly yield indirect economic benefits through increased employment. Likewise, efforts to increase a community's employment opportunities could also indirectly improve the health of its residents. The results of the secondary analysis excluding disability rates as a control variable suggest that the correlation between diabetes and labor force participation is at least partially independent of the effects of disability. This could mean that efforts to prevent the development or early progression of diabetes could generate economic benefits, rather than only interventions that prevent advanced disease or disability.

However, it is important to note that the model is designed to provide a *correlational* estimate, not a *causal* inference. While the model conceives of labor force participation rates as the dependent variable and diabetes prevalence as the independent variable that predicts it, the two measures likely exhibit bi-directional causality, wherein decreased workforce participation rate could cause diabetes prevalence at the same time that increases in diabetes rates depress labor force participation. Likewise, both diabetes and labor force participation rate are each likely influenced by a wide array of different factors, including many that are not accounted for in the present model.

Limitations

Aside from the causal uncertainty inherent in its correlational design, this project has important limitations that should inform how its findings are interpreted. Several assumptions inherent in its statistical design were likely violated. Secondary analyses intended to address violated assumptions did consistently reinforce the negative correlation between diabetes prevalence and labor force participation. However, the removal of multiple potentially objectionable variables in one secondary analysis decreased the strength of the correlation, and the coefficient of diabetes prevalence varied considerably between the various analyses (see Table 4).

The model may also exhibit some degree of overfitting, a problem that arises when a statistical model exactly accounts for observed data at the expense of failing to accurately capture the true underlying relationship being studied (Ying, 2019). The apparent R^2 value of the primary analysis was remarkably high (>0.9). However, secondary analyses showed that multiple variables could be removed from the model without a significant decrease in the R^2 value, implying that their addition in the primary analysis may not have truly increased the model's predictive power. While the secondary analysis that removed age and household income still found a significant negative correlation between the primary variables, overfitting is nonetheless a likely weakness of the primary analysis.

Even if the findings of the project are taken at face value, it should be noted that there was considerable uncertainty surrounding the point estimate of the correlation between diabetes prevalence and labor force participation. In the primary analysis, the point estimate of the correlation coefficient was -0.425. However, the associated 95% confidence interval included values from -0.779 to -0.072; the upper and lower limits lie a full order of magnitude apart from one another. It would therefore be imprudent to interpret the -0.425 correlation coefficient as a precise estimate of the relationship between diabetes and labor force participation in Kentucky counties.

An additional limitation of the project is its reliance on prior estimates from ACS and BRFSS data. While these data sets are used ubiquitously by researchers in the social sciences, the project's reliance on these secondary data is not without risks. The response rate to telephone surveys has dropped considerably in recent years, prompting the advent of new strategies for weighting data to produce accurate estimates in the face of likely sampling bias (Keeter, et al. 2017). BRFSS data has shown concordance with other national surveys that rely on self-reporting, but the level of consistency varies among different health domains and BRFSS estimates diverge from surveys that include objective physical data in addition to self-reported metrics (Pierannuzi, Hu, and Balluz, 2013). Likewise, ACS estimates are often accompanied by wide margins of error, especially in small geographic regions such as those studied in this project. The large uncertainties arise both from the methodologies employed by the U.S. Census Bureau and exogenous error beyond the control of government researchers (Spielman, Folch, and Nagle, 2014).

The uncertainty in the original estimates used in the projects highlights one of its most serious limitations: the absence of robust error propagation. The project's sensitivity analyses are a crude substitute, only taking into account improbably large,

unidirectional errors in the two primary variables while ignoring the uncertainties present in the secondary variable estimates. While they did find that the model would fail to generate statistically significant results in the presence of a large overestimation of county-level labor force rates, the sensitivity analyses do not adequately assess the vulnerability of the model's findings to the uncertainty in its constituent data. The standard errors, 95% confidence intervals, and R^2 values reported by the project's models must be interpreted as inappropriately precise because they do not represent the error present in the original estimates. Better results could be obtained through Monte Carlo simulation methods, whereby the primary and secondary analyses could each be repeated thousands of times with variables that vary randomly according to the error published in the original estimates (James, 1980). However, such an simulationist method for error propagation was beyond the means of the current project.

Directions for Future Research

Various mechanisms may explain the negative correlation between diabetes and labor force participation in Kentucky counties. Future research could elucidate these mechanisms, which would guide policymakers and community members in tailoring solutions to both relieve the burden of diabetes and foster improved labor market outcomes. Some possible mechanisms are offered here, both to highlight the uncertainty in interpreting the results of the study and to identify directions for further study.

Perhaps the most obvious potential mechanism is diabetes-associated disability. The progression of the disease, if left unchecked, can lead to debilitating symptoms that make employment extremely difficult. The results of this project imply that this mechanism is present to a meaningful degree in Kentucky counties: the correlation between diabetes and labor force participation was *stronger* when disability rates were not controlled for in the regression model, growing from -0.425 to -0.525. However, this project's correlational findings are unable to differentiate between the effects of disability arising directly from diabetes versus other possible relationships between diabetes, disability, and labor force participation. Future research could compare the employment characteristics of diabetic individuals with and without diabetes-associated disabilities to explore this question.

Another possible mechanism relates to an increased demand for caregiving when diabetes prevalence rises. Friends and family members of individuals with diabetes may find themselves spending significant time and effort in the role of caregiver,

whether through providing direct health care, managing medications, providing transport to medical appointments, or helping with impaired activities of daily living. For individuals with poorly-controlled diabetes, having an informal caregiver is associated with greater medication adherence and better self-care metrics (Bouldin, et al., 2017). However, the demands of being a diabetes caregiver could make it harder to maintain traditional employment. Increasing rates of diabetes could therefore decrease labor force participation by indirectly removing caregivers from the formal workforce. Future studies could compare the relationship between diabetes caregiver status (or hours spend providing diabetes care) and employment outcomes.

A third potential explanation for the correlation between diabetes and lower labor force participation is a perverse incentive created by the American healthcare system. Diabetes is an extremely expensive disease; even with commercial prescription insurance, patients often face high copays for insulin and other medications (Adu et al., 2019; Mogre et al., 2019). Many Kentuckians with diabetes rely on means-tested Medicaid insurance plans to obtain their medications (KY CHFS et al., 2021). For some individuals, exceeding Medicaid income limits could mean a disproportionate increase in healthcare costs that income from additional employment could not accommodate. Some Kentuckians with diabetes might therefore choose to work fewer hours, seek income through the informal economy, or not seek work at all for fear of losing their health insurance benefits. A regression discontinuity design study could compare employment outcomes for Kentuckians with diabetes with incomes close to the cutoff for Medicaid eligibility to investigate whether this phenomenon is significant. If so, expanding access to high-quality, affordable health insurance and prescription coverage could bolster the Commonwealth's economy.

The three mechanisms proposed above all conceive of diabetes as the driver for reduced labor force participation. However, it is also plausible that decreased labor force participation rates cause higher rates of diabetes. For example, a longitudinal Danish study showed that the loss of employment led to more unhealthy dietary consumption, including a long-term substitution of fresh food sources for increased carbohydrates and added sugars (Smed, et al., 2018). Adverse economic conditions such as reductions in employment opportunities could put a county's residents at heightened risk of new or worsening diabetes and other health problems, potentially contributing to a destructive feedback loop of damaged health and economic vitality. Additionally, reductions in county-level labor force participation rates might also be accompanied by decreases in residents' economic resources and ability to access healthcare. Both financial cost and transportation have been identified as important barriers to treating diabetes (Adu et al., 2019; Mogre et al., 2019). Loss of employment could therefore

reduce individuals' ability to afford medications or be able to travel to their health providers, leading to worsening progression of diabetes or new-onset diabetes in the prediabetic population. Longitudinal cohort studies could track the effects of unemployment and labor force drop-out on subjects' likelihood to develop diabetes.

Finally, future research designs could improve and expand upon this project without investigating the specific causal mechanisms relating diabetes to labor force participation. The most important improvement would come in the form of improved statistical robustness, especially through properly accounting for the uncertainty present in ACS and BRFSS estimates. More sophisticated statistical models could test for non-linear relationships between the primary variables, and weighting methods could be used to ensure the appropriateness of the secondary variables present in the model. Refined versions of this project could re-examine diabetes and labor force participation in Kentucky counties, whether across the same 2015-2019 time frame studied here or others. Studies could examine changes in diabetes prevalence and labor force participation over time to determine how earlier changes in one variable predict later variation in the other. Researchers could also expand the focus of the current project to different states, regions, or even nationwide. Quasi-experimental methods such as difference-in-differences, regression discontinuity, and instrumental variables could investigate the causal directionality between diabetes and labor force outcomes.

CONCLUSION

For counties in Kentucky, higher diabetes prevalence is correlated with significantly lower labor force participation rates, with the primary analysis estimating a correlation coefficient of -0.425 (95% CI: -0.779,-0.072). This would correspond to a 0.425% reduction in a county's labor force participation rate for every 1% increase in its prevalence of diabetes. The strength of the correlation varied across secondary analyses intended to account for the statistical shortcomings of the projects, ranging from 0.282 (-0.720, +0.156) to -0.541 (-0.982, -0.0.99). Sensitivity analyses using the upper and lower limits of the ACS and BRFSS estimates employed in this project showed that large over-estimations of county labor force participation rates would diminish the strength and significance of the observed correlation. These findings support prior research connecting diabetes to worsened economic outcomes and offer new Kentucky-specific estimates not previously available. The project was limited by its unsophisticated statistical methods, principally a lack of adequate error propagation. Future research is needed to explore the causal relationships between diabetes and labor force participation in Kentucky and beyond, as well as to identify and design interventions that can improve both the health of the Commonwealth's residents and the strength of the state's economy.

BIBLIOGRAPHY

Adu, M. D., Malabu, U. H., Malau-Aduli, A. E. O., & Malau-Aduli, B. S. (2019). Enablers and barriers to effective diabetes self-management: A multi-national investigation. *PloS one*, 14(6)

American Diabetes Association. Standard of Care in Diabetes – 2023. (2023). *Diabetes Care,* 46(1)

American Diabetes Association. (2018). Economic costs of diabetes in the U.S. in 2017. *Diabetes Care*, 41, 917- 929.

Bouldin, Erin D., et al. "Associations between having an informal caregiver, social support, and self-care among low-income adults with poorly controlled diabetes." *Chronic illness* 13.4 (2017): 239-250.

Clark, M. Minier, J. Courtemanche, C. Paris, B. Childress, M. (2019). The Economic Impact of Diabetes in Kentucky. Center for Business and Economic Research, Gatton College of Business and Economics.

Fang M. (2018). Trends in the Prevalence of Diabetes Among U.S. Adults: 1999-2016. *American Journal of Preventive Medicine*, 55(4), 497–505. https://doi.org/10.1016/j.amepre.2018.05.018

FRED (2019). Federal Reserve Economic Data. Federal Reserve Bank of St. Louis. Accessed at https://fred.stlouisfed.org/release/tables?rid=446&eid=784070&od=2019-01-01#

Herber, GC., Ruijsbroek, A., Koopmanschap, M. et al (2019). Single transitions and persistence of unemployment are associated with poor health outcomes. *BMC Public Health* 19, 740. https://doi.org/10.1186/s12889-019-7059-8

James, Frederick. "Monte Carlo theory and practice." *Reports on progress in Physics* 43.9 (1980): 1145.

Keeter, S., Hatley, N., Kennedy, C., & Lau, A. (2017). What low response rates mean for telephone surveys. *Pew Research Center*, 15(1), 1-39.

KY Cabinet for Health and Family Services, Department for Medicaid Services, Department for Public Health, Office of Health Data and Analytics, and KY Personnel Cabinet, Department of Employee Insurance. (2021) 2021 Kentucky Diabetes Report. Accessed March 2023 at https://www.chfs.ky.gov/agencies/dph/dpqi/cdpb/dpcp/2021DiabetesReport.pdf

Maggio, C. A., & Pi-Sunyer, F. X. (2003). Obesity and type 2 diabetes. *Endocrinology and metabolism clinics of North America*, 32(4), 805–viii. https://doi.org/10.1016/s0889-8529(03)00071-9

Minor T. (2013) An investigation into the effect of type I and type II diabetes duration on employment and wages. *Econ Hum Biol*. 11:534–544.

Mogre, V., Johnson, N. A., Tzelepis, F., & Paul, C. (2019). Barriers to diabetic self-care: A qualitative study of patients' and healthcare providers' perspectives. *Journal of clinical nursing*, 28(11-12), 2296–2308.

Ng YC, Jacobs P, Johnson JA. (2001) Productivity losses associated with diabetes in the U. S *Diabetes Care*. 24(2):257–61.

Ozougwu, J. C., Obimba, K. C., Belonwu, C. D., & Unakalamba, C. B. (2013). The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus. *J Physiol Pathophysiol*, 4(4), 46-57.

Pedron, S., Emmert-Fees, K., Laxy, M. et al (2019). The impact of diabetes on labour market participation: a systematic review of results and methods. *BMC Public Health* 19, 25. https://doi.org/10.1186/s12889-018-6324-6

Pierannunzi, C., Hu, S.S. & Balluz, L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC Med Res Methodol* 13, 49 (2013). https://doi.org/10.1186/1471-2288-13-49

Seuring, T., Archangelidi, O., & Suhrcke, M. (2015). The economic costs of type 2 diabetes: a global systematic review. *Pharmacoeconomics*, 33(8), 811-831.

Smed, S., Tetens, I., Lund, T. B., Holm, L., & Nielsen, A. L. (2018). The consequences of unemployment on diet composition and purchase behaviour: a longitudinal study from Denmark. *Public health nutrition*, 21(3), 580-592.

Spielman, S. E., Folch, D., & Nagle, N. (2014). Patterns and causes of uncertainty in the American Community Survey. *Applied geography*, 46, 147-157

Tunceli, K., Zeng, H., Habib, Z. A., & Williams, L. K. (2009). Long-term projections for diabetes-related work loss and limitations among U.S. adults. *Diabetes Research and Clinical Practice*, 83(1), e23-e25.

Tranmer, M., Murphy, J., Elliot, M., and Pampaka, M. (2020) Multiple Linear Regression (2nd Edition); Cathie Marsh Institute Working Paper 2020-01. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/2020-1- multiple-linear-regression.pdf

U.S. Bureau of Labor Statistics. (2023). Kentucky Labor Force Data. https://www.bls.gov/eag/eag.ky.htm

United States. Centers for Disease Control and Prevention. (2022). By the Numbers: Diabetes in America. National Diabetes Statistics Report, 2022.

United States. Centers for Disease Control and Prevention (2022). Diabetes Basics

United States. Centers for Disease Control and Prevention (2022). Diabetes Mortality by State

United States. Centers for Disease Control and Prevention (2022). Prevalence of Both Diagnosed and Undiagnosed Diabetes

United States. Centers for Disease Control and Prevention (2022). Stats of the States.

United States. Centers for Disease Control and Prevention (2000-2018) Behavioral Risk Factor Surveillance System Data.

Ying, Xue. "An overview of overfitting and its solutions." *Journal of physics: Conference series*. Vol. 1168. IOP Publishing, 2019.

Zhang X, Zhao X, Harris A. (2009) Chronic diseases and labour force participation in Australia. *J Health Econ*.28:91–108.

APPENDIX I: IRB COMPLIANCE

The author of this project completed all University of Kentucky mandated training for research ethics, including CITI certification for Responsible Conduct of Research and Human Research. This project received an official exemption from IRB review on the grounds that it did not constitute human research. No subjects were recruited, intervened upon, or interacted with; no individually identifiable information was collected or reviewed; and all data used was publicly available and de-identified. A copy of the NHR exemption determination is available upon request at rcmo229@g.uky.edu

APPENDIX II: TABLES

Table 4: Variables used in multivariable regression model

Variable	*Variable "Tag"*	*Category*	*Variable Type*	*Data Source*	*Description*
Labor force participation rate	LaborForce	Primary: Dependent	Continuous (%)	US Census Bureau: American Community Survey 2015-2019 5 Year Estimates	Proportion of non-institutionalized civilian population 16+ years old that is either working or looking for work
Diabetes prevalence	Diabetes	Primary: Independent	Continuous (%)	Centers for Disease Control: Behavioral Risk Factor Surveillance System, yearly estimates from 2015-2019.	Average of yearly 2015-2019 BRFSS point-estimates of diabetes prevalence for each county
Urban/Rural classification	UrbanRural	Control: Independent	Categorical (1-6)	National Center for Health Statistics: 2013 Urban-Rural Classification Scheme for Counties	Categorization scheme with 6 levels describing how rural or urban a county is. For the purpose of this analysis, the higher the score, the more urban the county 1: Noncore 2: Micropolitan 3: Small Metro 4: Medium Metro 5: Large Fringe Metro 6: Large Central Metro
Mean age	Age	Control: Independent	Continuous (years)	US Census Bureau: American Community Survey 2015-2019 5 Year Estimates	Mean age among the total population of each county
Male sex ratio	Male	Control: Independent	Continuous (%)	US Census Bureau: American Community Survey 2015-2019 5 Year Estimates	Proportion of the total county population that identifies as male.
Proportion 65 years or older	SixtyFivePlus	Control: Independent	Continuous (%)	US Census Bureau: American Community Survey 2015-2019 5 Year Estimates	Proportion of the total county population aged 65 years or older
Proportion mono-racial non-Hispanic white	White	Control: Independent	Continuous (%)	US Census Bureau: American Community Survey 2015-2019 5 Year Estimates	Proportion of the total population that identifies as white, non-Hispanic, non-Latino, and does not identify as bi-racial or multi-racial.
Mean educational attainment	Education	Control: Independent	Continuous (1-5)	US Census Bureau: American Community Survey 2015-2019 5 Year Estimates	Average of the highest educational attainment among those 25+ years old. Categories within the ACS were converted to a 5 point scale, and the average county-level attainment calculated as a number between 1 and 5. 1: Less than High School Diploma or Equivalent 2: High School Diploma or Equivalent 3: Some college or Associate's Degree 4: Bachelor's Degree 5: Graduate or Professional Degree
Mean household income	Income	Control: Independent	Continuous ($/year)	US Census Bureau: American Community Survey 2015-2019 5 Year Estimates	The mean annual income earned by households in each county.
Poverty rate	Poverty	Control: Independent	Continuous (%)	US Census Bureau: American Community Survey 2015-2019 5 Year Estimates	Proportion of households that fell below 100% of the poverty threshold within the 12 months preceding their response to the ACS.
Disability rate	Disability	Control: Independent	Continuous (%)	US Census Bureau: American Community Survey 2015-2019 5 Year Estimates	Percentage of individuals with any disability among the civilian non-institutionalized population.

Table 5: Average Characteristics of Kentucky Counties (2015-2019)

Variable	*Mean Value*	*Standard Deviation*	*Minimum*	*Maximum*	*Notes*
Labor force participation rate	53.3%	8.6%	27.6% (Elliot)	70.2% (Boone)	

Diabetes prevalence	10.9%	1.7%	7.9% (Scott)	16.7% (Letcher)	
Urban/Rural classification	2.1	1.4	1 (59 counties)	6 (Jefferson)	The average Kentucky county would be characterized as "Micropolitan" on the NCHS 2013 Urban-Rural Classification Scheme for Counties.
Mean age	41.0	3.0	28.3 (Christian)	50.6 (Lyon)	
Male sex ratio	98.9	7.0	83.3 (Owsley)	131.7 (Elliot)	The average Kentucky county has about 99 males for every 100 females.
Proportion 65 years or older	17.4	2.4	11.9 (Scott)	26.7 (Lyon)	
Proportion mono-racial non-Hispanic white	91.5	6.2	65.6 (Christian)	99 (Estill)	In the average Kentucky county, 91.5% of the population identifies as white race/ethnicity and does not identify as biracial, Hispanic, or Latino.
Mean educational attainment	2.5	0.2	2.1 (Clay)	3.3 (Oldham)	In the average Kentucky county, the mean highest educational attainment achieved by residents falls between a "High School Diploma or Equivalent" and "Some college or Associate's Degree."
Mean household income	$44,900/year	$12,000/year	$24,600/year (Wolfe)	$99,100/year (Oldham)	
Poverty rate	20.3%	7.2%	5.9% (Oldham)	38.4% (Clay)	
Disability rate	20.9%	5.5%	8.8% (Oldham)	36.8% (Wolfe)	

Table 6: Summary of all regression results

Regression	***Regression Name***	***Dependent Variable***	***Primary Independent Variable***	***Control Independent Variables***	***Coefficient of Primary Independent Variable (95% CI)***	***R^2 Value***
Primary regression:	Primary	Labor Force	Diabetes Prevalence	UrbanRural	-0.425 (-0.779, -0.072)	0.907

point estimates + all ten control variables		Participation Rate		Age Male SixtyFivePlus White Education Income Poverty Disability		
Excluding disability rates	Secondary 1	Labor Force Participation Rate	Diabetes Prevalence	UrbanRural Age Male SixtyFivePlus White Education Income Poverty	-0.525 (-0.873, -0.176)	0.903
Excluding household income	Secondary 2	Labor Force Participation Rate	Diabetes Prevalence	UrbanRural Age Male SixtyFivePlus White Education Poverty Disability	-0.409 (-0.753, -0.065)	0.907
Excluding mean age	Secondary 3	Labor Force Participation Rate	Diabetes Prevalence	UrbanRural Male SixtyFivePlus White Education Income Poverty Disability	-0.429 (-0.778, -0.080)	0.907
Excluding age, male, and white	Secondary 4	Labor Force Participation Rate	Diabetes Prevalence	UrbanRural SixtyFivePlus Education Income Poverty Disability	-0.282 (-0.720, +0.156)	0.843
Excluding age and income	Secondary 5	Labor Force Participation Rate	Diabetes Prevalence	UrbanRural Male SixtyFivePlus White Education Poverty Disability	-0.410 (-0.752, -0.067)	0.907
Excluding counties within top/bottom 5.5% for labor force participation or diabetes prevalence	Secondary 6	Labor Force Participation Rate	Diabetes Prevalence	UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.541 (-0.982, -0.0.99)	0.872
UpperLaborForce vs UpperDiabetes	Sensitivity 1	Labor Force (Upper Limit)	Diabetes Prevalence (Upper Limit)	UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.623 (-0.997, -0.249)	0.894
LaborForce vs UpperDiabetes	Sensitivity 2	Labor Force (Point Estimate)	Diabetes Prevalence (Upper Limit)	UrbanRural Age Male SixtyFivePlus White Education Income Poverty	-0.433 (-0.808, -0.058)	0.907

LowerLaborForce vs UpperDiabetes	Sensitivity 3	Labor Force (Lower Limit)	Diabetes Prevalence (Upper Limit)	Disability UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.243 (-0.637, +0.150)	0.911
UpperLaborForce vs Diabetes	Sensitivity 4	Labor Force (Upper Limit)	Diabetes Prevalence (Point Estimate)	UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.658 (-1.00, -0.307)	0.897
LaborForce vs Diabetes (identical to primary analysis)	N/A	Labor Force (Point Estimate)	Diabetes Prevalence (Point Estimate)	UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.425 (-0.779,-0.072)	0.907
LowerLaborForce vs Diabetes	Sensitivity 5	Labor Force (Lower Limit)	Diabetes Prevalence (Point Estimate)	UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.195 (-0.568, +0.178)	0.910
UpperLaborForce vs LowerDiabetes	Sensitivity 6	Labor Force (Upper Limit)	Diabetes Prevalence (Lower Limit)	UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.666 (-1.00, -0.331)	0.898
LaborForce vs LowerDiabetes	Sensitivity 7	Labor Force (Point Estimate)	Diabetes Prevalence (Lower Limit)	UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.424 (-0.765, -0.082)	0.907
LowerLaborForce vs LowerDiabetes	Sensitivity 8	Labor Force (Lower Limit)	Diabetes Prevalence (Lower Limit)	UrbanRural Age Male SixtyFivePlus White Education Income Poverty Disability	-0.181 (-0.542, +0.179)	0.910

Table 7: Testing assumptions of the linear regression model

Assumption	***Description***	***Test(s) conducted***	***Conclusions***
Independence of observations	Observations included the regression should not be biased by shared	N/A	Impossible to test. The exact methods of data collection and

	temporal, spatial, or other relationships arising from sampling errors or flawed data collection methods.		analysis used by the Census Bureau and CDC are unknown, and genuine temporal and spatial relationships are likely to exist within the data.
Linearity	The dependent variable should exhibit a linear relationship with each of the independent variables on their own, as well as with the independent variables collectively.	Graph matrix visualization RESET (Ramsey Regression Equation Specification Error Test)	Age and Male, and to a lesser degree White appear non-linear with LaborForce. Significant p value from RESET suggests possible non-linear relationship between LaborForce and the independent variables as a collective.
Normality	The residuals of the regression should be approximately normally distributed.	Visualization of residual-versus-fitted plot Shapiro-Wilk test for normality	Residual-versus-fitted plot appears to support normality. Prob>z = 0.429: Shapiro-Wilk test indicates that the residuals may *not* be normally distributed.
Homoscedasticity	The residuals of the regression should have equal variance for every value of the fitted values and of the predictors in order for the regression to generate accurate standard errors.	Breusch-Pagan test	Prob>chi2 = 0.0057. Residuals appear to be homoscedastic.
Independence	There should be no remaining association between observations not accounted for by the predictors in the regression model.	N/A	Labor force participation rates almost certainly ARE influenced by many different variables not accounted for in the regression model. It would be impossible to perfectly predict such a highly endogenous economic measure.
No Multicollinearity	Each independent variable should make a unique contribution to predicting the dependent variable, without any independent variable being redundant or containing a significant amount of information contained in another independent variable.	VIF (Variance Inflation Factor) Test	Income, Age, Poverty, 65+, and Education have VIF > 5. Income has VIF > 10 Several variables in the regression may exhibit multicollinearity, with household income being the worst offender.
No outlier effects	The results of the regression should not be distorted by extreme values.	Excluding counties at the 5.5^{th} percentile for highest or lowest labor force participation or diabetes prevalence. Clay, Oldham, and Wolfe counties were specifically excluded as potential outlier counties (see **Table 5**).	The strength of the correlation *increased* after potential outlier counties were removed, suggesting that the model's findings are not over-exaggerated by outlier effects.