BMJ Open

Association of community types and features in a case-control analysis of new onset type 2 diabetes across a diverse geography in Pennsylvania

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-043528
Article Type:	Original research
Date Submitted by the Author:	08-Aug-2020
Complete List of Authors:	Schwartz, BS; Johns Hopkins University Bloomberg School of Public Health, Department of Environmental Health and Engineering Pollak, Jonathan; Johns Hopkins University Bloomberg School of Public Health, Department of Environmental Health and Engineering Poulsen, Melissa; Geisinger, Population Health Sciences Bandeen-Roche, Karen; Johns Hopkins University Bloomberg School of Public Health, Biostatistics Moon, Katherine; Johns Hopkins Bloomberg School of Public Health Center for Teaching and Learning, Environmental Health & Engineering DeWalle, Joseph; Geisinger, Population Health Sciences Siegel, Karen; Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion Mercado, Carla; Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion Imperatore, Giuseppina ; Centers for Disease Control and Prevention, Division of Diabetes Tranlsation Hirsch, Annemarie; Geisinger Health System, Department of Epidemiology and Health Services Research
Keywords:	DIABETES & ENDOCRINOLOGY, EPIDEMIOLOGY, PUBLIC HEALTH

Association of community types and features in a case-control analysis of new onset type 2 diabetes across a diverse geography in Pennsylvania

Brian S. Schwartz,[1,2,4,5] Jonathan S. Pollak,[1] Melissa N. Poulsen,[5] Karen Bandeen-Roche,[3] Katherine A. Moon,[1] Joseph DeWalle,[5] Karen R. Siegel,[6] Carla I. Mercado,[6] Giuseppina Imperatore,[6] Annemarie G. Hirsch[1,5]

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[1] Department of Environmental Health and Engineering

[2] Department of Epidemiology

[3] Department of Biostatistics

Johns Hopkins School of Medicine, Baltimore, MD

[4] Department of Medicine

Geisinger, Danville, PA

[5] Department of Population Health Sciences

Centers for Disease Control and Prevention, Atlanta, GA

[6] Division of Diabetes Translation, National Center for Chronic Disease Prevention and Health Promotion

Corresponding author: Brian S. Schwartz, Department of Environmental Health and Engineering, 615 N. Wolfe Street, Room W7041, Baltimore, MD, 21205, voice 410-955-4158, email bschwar1@jhu.edu.

Word count: abstract = 287, manuscript = 3122 | **Tables and figures** = 4 | **Online supplement tables** = 6 | **References** = 35

Running title: Geography of type 2 diabetes in Pennsylvania

Abstract

Objectives: To evaluate associations of community types and features with new onset type 2 diabetes in diverse communities. Understanding the location and scale of geographic disparities can lead to community-level interventions.

Design: Nested case-control study within the open dynamic cohort of health system patients.

Setting: Large, integrated health system in 37 counties in central and northeastern Pennsylvania, USA.

Participants and analysis: We used electronic health records to identify persons with new-onset type 2 diabetes from 2008–2016 (n = 15,888). Persons with diabetes were age, sex, and year matched (1:5) to persons without diabetes (n = 79,435). We used generalized estimating equations to control for individual-level confounding variables, accounting for clustering of persons within communities. Communities were defined as 1) townships, boroughs, and city census tracts; 2) urbanized area (large metro), urban cluster (small cities and towns), and rural; 3) combination of the first two; and 4) county. Community socioeconomic deprivation and greenness were evaluated alone and in models stratified by community types.

Results: Borough and city census tract residence (vs. townships) were associated (odds ratio [95% confidence interval]) with higher odds of type 2 diabetes (1.10 [1.04-1.16] and 1.34 [1.25-1.44], respectively). Urbanized areas (vs. rural) also had increased odds of type 2 diabetes (1.14 [1.08-1.21]). In the combined definition, the strongest associations (vs. townships in rural areas) were city census tracts in urban clusters (1.41 [1.22-1.62]) and city census tracts in urbanized areas (1.33 [1.22-1.45]). Higher community socioeconomic deprivation and lower greenness were each associated with increased odds.

Conclusions: Urban residence was associated with higher odds of type 2 diabetes than for other areas. Higher community socioeconomic deprivation in city census tracts and lower greenness in all community types were also associated with type 2 diabetes.

Strengths and limitations of this study

- Type 2 diabetes, with a large sample size, was objectively documented and verified or excluded with extensive biomarker and medical data.
- Temporality was appropriate for all independent variables.
- We studied several approaches to community characterization at more relevant contextual scales than many prior studies in a range of communities from urban to rural.
- We did not measure behavioral mediators of the community definitions and features, such as physical activity or dietary intake.
- We could not account for residential selection bias, but the residential stability and general population representativeness of our study population may mitigate these concerns.

INTRODUCTION

Diabetes is a common and costly chronic disease; in the U.S. in 2018, over 34 million individuals had diabetes, with annual spending exceeding $320 billion [1]. Diabetes occurrence varies by race/ethnicity and also evidences geographic disparities [2, 3]; prevalence by county in the U.S. varies over a 7-fold range [4]. Studies report that diabetes is 17% more prevalent in rural than urban areas [5], consistent with rural health disparities for other chronic conditions [6, 7], attributed to sociodemographic factors (e.g., higher poverty, older populations) and barriers to health care access [8, 9].

Community characteristics that may underlie observed geographic disparities in type 2 diabetes include land use (e.g., walkable vs. automobile dependent), fitness, food, and social (e.g., deprivation, disorganization) environments; greenspace (i.e., natural environments); and air pollution. Some of these are diabetogenic and others protective [10-12]. Community characteristics co-occur in patterns that differ by **community type** (e.g., higher population density co-occurs with higher deprivation and food availability and lower automobile dependence and greenness). Simultaneously evaluation and control of these domains across community types can be problematic due to limited and non-overlapping distributions that make independent attribution of disease risk to specific domains difficult [13]. An alternative is to use carefully defined community types to first identify the **location** and **geographic scale** of type 2 diabetes risk. These community types should reduce within community variation and maximize between community differences. Subsequent analyses can then stratify by community type and evaluate well-characterized **community features** in relation to type 2 diabetes risk.

Residential development patterns reflect a continuum from rural to urban with variation by many community features [14]. The U.S. Census Bureau defines *urbanized areas* as dense settlements with 50,000 or more residents, *urban clusters* as areas with 2500–50,000 residents, and all others as *rural* [15]. In Pennsylvania, communities are defined administratively as townships, boroughs, and cities using census minor civil division boundaries [16]. In combination, these two definitions provide an opportunity to evaluate experientially and behaviorally relevant geographies as well as to further subdivide the broad category of "rural," which includes a range of communities that vary in their associations with health outcomes [17, 18].

We evaluated four definitions of community across a range of community types from rural to urban in a 37-county region of Pennsylvania, in relation to type 2 diabetes onset to inform more robust study of the community-level features that may underlie type 2 diabetes risk. Next, because higher community socioeconomic deprivation and lower greenness have been consistently associated with higher risk of type 2 diabetes [19, 20], we evaluated associations with these features overall and within community types.

METHODS

Study Population and Design

This study was conducted by Geisinger-Johns Hopkins Bloomberg School of Public Health, one of four academic research centers in the Diabetes LEAD (Location, Environmental Attributes, and Disparities) Network (http://diabetesleadnetwork.org/), a collaboration funded by the Centers for Disease Control and Prevention dedicated to providing scientific evidence to develop targeted interventions and policies to prevent type 2 diabetes and related health outcomes across the U.S. The study was approved by the Geisinger Institutional Review Board under waivers of consent and assent to use electronic health record (EHR) data.

Using previously reported methods [16], we used Geisinger EHR data from 1.6 million individuals to identify new onset type 2 diabetes from 2008–2016. Individuals represent the general population in the region with high residential stability [21]. The study area included 37 counties in Pennsylvania (**Figure 1**). These data were used in a nested case-control study.

Patient and Public Involvement

Patients and public representatives were not involved in the development of the study. Study results will be disseminated through Geisinger's Environmental Health Institute in its website (https://www.geisinger.edu/research/departments-and-centers/environmental-health-institute) and communications to Geisinger patients and the public.

Identification of New Onset Type 2 Diabetes Cases and Controls

Persons with type 2 diabetes (n = 15,888) were identified using diabetes encounter diagnoses, medication orders, and laboratory test results (**Online Supplement Table S1**). EHR algorithms can identify diabetes with high sensitivity, specificity, and positive predictive value [22, 23]. Controls (n = 79,435, with 65,084 unique persons), persons who never met any diabetes criteria, were randomly selected with replacement and frequency-matched to cases (5:1) on age, sex, and year of encounter. To ensure that we could identify diabetes if present, we required at least two encounters on different days with a primary care provider prior. To ensure diabetes was new onset, persons had to have at least one encounter with the health system at least two years prior without evidence of diabetes.

Community Types and Community Features

Addresses at last contact with the health system were geocoded using ArcGIS version 10.4 (ESRI Inc., Redlands, CA). We used four definitions of community to evaluate different spatial scales and a range of characterizations of the size and urbanicity of these areas (**Figure 2**). First, using minor civil divisions and census tract boundaries, we categorized study communities into townships, boroughs, and city census tracts, as previously reported [24], referred to as *administrative community type*. Townships range from agriculturally-focused rural areas to low density suburbs; boroughs are walkable small towns of 5,000 to 10,000 persons with a core area of gridded streets; and cities are medium-sized urban areas (largest is Scranton–Wilkes-Barre–Hazleton Metropolitan Statistical Area, 97th in U.S. by population). Second, we used U.S. Census Bureau's urbanized areas and urban clusters to define residential addresses as "major urban," "smaller urban," and "rural" [15], referred to as *urban/rural status*. Third, to evaluate community at a more granular level, we combined the first and second categorizations, referred to as *combined community type*. This resulted in eight groups (city census tract/rural had few residences so were combined with borough/rural; township/rural was reference group). Fourth, because most prior research of geographic disparities in diabetes evaluated counties, which are much larger geographies, we evaluated counties alone and after stratification by administrative community type.

We evaluated two time-varying community features. Peak (16-day composite in early July of each year) normalized difference vegetation index (NDVI, referred to as greenness) was evaluated in 1250m squares around residences in the prior year [25]. We measured community socioeconomic deprivation using a previously described scale [26], the sum of z-transformed values of six indicators identified from a factor analysis (proportion unemployed, less than a high school education, below poverty level, on public assistance, not in the workforce, and without a car), using data from the Decennial Census (2000 only) and American Community Survey (2006-2010, 2011-2015). The scale was assigned as the closest measure prior to the year of onset/encounter.

Statistical Analysis

The goals of the analysis were: 1) evaluate four definitions of community in relation to odds of type 2 diabetes onset; 2) evaluate two community features, community socioeconomic deprivation and greenness, in relation to type 2 diabetes onset in all communities; and 3) evaluate associations of the two community features after stratification by community type. Analysis controlled for key individual-level confounding variables and accounted for spatial clustering of persons within communities. Statistical analysis was completed using Stata-MP version 15.1 (StataCorp LLC, College Station, TX).

Logistic regression was used to estimate associations (odds ratios, 95% confidence intervals) using generalized estimating equations with robust standard errors and an exchangeable correlation structure within administrative community types. We adjusted for age (years; linear, quadratic, and cubic terms to allow for non-linearity), sex, race (white vs. all other races), ethnicity (Hispanic vs. non-Hispanic), and percent of time using Medical Assistance (surrogate for family socioeconomic status [≥ 50% vs. < 50%])

[27]. We did not include body mass index (BMI, kg/m^2) in models because this is likely a mediator of community associations (inclusion would attenuate or eliminate associations of interest). Models were first evaluated using all persons in all communities. We analyzed associations of the four definitions of community, community socioeconomic deprivation (quartiles; 4th quartile [worst deprivation] reference group), and greenness (tertiles) with diabetes status. Due to concerns about non-overlapping distributions resulting in extrapolation rather than adjustment (i.e., non-positivity [28]), we then stratified the community features models by community type.

In sensitivity analyses, to evaluate whether access to care – and thus higher likelihood of diabetes diagnosis – may have accounted for associations between community and diabetes, we examined the number of prior outpatient encounters (linear and quadratic terms) for study individuals by administrative community type and Medical Assistance status and added this variable to regression models.

RESULTS

Description of Study Population and Communities

Individuals were predominantly white and non-Hispanic; the majority had a primary care provider; and most cases were diagnosed with diabetes in an outpatient setting (**Table 1**). Individuals resided in 291 boroughs, 146 city census tracts, and 633 townships (**Online Supplement Table S2**). Over 40% of persons resided in rural areas (**Table 1**). Most borough residents were divided between urbanized areas and urban clusters. Approximately two-thirds of persons in townships resided in rural areas. A similar proportion of individuals in city census tracts resided in urbanized areas. On average, townships had higher greenness and lower community socioeconomic deprivation compared to boroughs and city census tracts (**Online Supplement Table S2**). Average racial and ethnic diversity and use of Medical Assistance for health insurance were highest in city census tracts. The mean total number of encounters with the health system before diabetes onset or the control selection date was high for all individuals, in all community types, regardless of Medical Assistance status (**Online Supplement Table S3**). Laboratory data confirmed that the categorization of diabetes cases and controls was valid (**Online Supplement Table S4**).

Associations of Communities with Type 2 Diabetes Onset

In the base model, controlling for age and sex, non-white race (vs. white), Hispanic ethnicity (vs. non-Hispanic), and Medical Assistance status were each associated with increased odds of type 2 diabetes onset. These associations did not substantively change as the community type and community features were added to the model. Odds ratios for non-white race (vs. white) ranged from 1.36 to 1.41, for Hispanic ethnicity (vs. non-Hispanic) from 1.46 to 1.52, and for Medical Assistance (≥ 50% of time vs. < 50%) from 1.71 to 1.74, with all confidence intervals excluding 1.0. Next, when administrative community type was added (townships as reference group), residing in boroughs and city census tracts was associated with significantly higher odds (**Table 2, Model 1**). Second, urban/rural status was added to the base model and residing in urbanized areas (vs. rural areas) had increased odds of diabetes onset (**Table 2, Model 2**). Third, the combined definition was added to the base model, and some categories (e.g., city census tracts in major urban and smaller urban areas highest, boroughs in these areas intermediate, vs. townships in rural areas as reference) were associated with increased odds of new onset diabetes (**Table 2, Model 3**). Finally, county was added to the base model, and seven counties were associated with reduced odds and two with increased odds of diabetes (**Table 2, Model 4**). We next evaluated community socioeconomic deprivation and greenness. When these community features were added to the base model, lower deprivation (**Table 2, Model 5**) and higher greenness (**Table 2, Model 6**) were associated with reduced odds of diabetes.

Models were next stratified by community type (only results for administrative community type shown). Race/ethnicity and Medical Assistance status were still associated with type 2 diabetes onset in the stratified models in all community types (**Online Supplement Table S5**). Associations of community socioeconomic deprivation with diabetes evidenced decreasing odds ratios across decreasing deprivation quartiles in all community types, but only crossed an inferential threshold in city census tracts, with approximately 25% lower odds in the 1st vs. 4th quartile. Higher greenness was associated with reduced odds of diabetes in all community types.

Even after stratification by administrative community type and adjustment for community socioeconomic deprivation, several counties were independently associated with increased or reduced odds of diabetes onset (**Online Supplement Table S6**). The number of significant associations (n = 18, nine each with reduced or increased odds) was somewhat larger than that expected due to chance (108 statistical tests performed), with most associations observed for residing in boroughs. In these models, associations with community socioeconomic deprivation were present in the 1st quartile (vs. 4th) in townships and boroughs and in all quartiles in city census tracts. In all community types, higher greenness was associated with lower odds of diabetes.

Sensitivity Analyses

Addition of total outpatient encounters before diagnosis/control selection date did not substantively change associations in non-stratified or stratified models (results not shown). Community socioeconomic deprivation and greenness were evaluated together in models in boroughs and townships. In boroughs, associations of greenness with type 2 diabetes onset were attenuated by 1-2% and associations with community socioeconomic deprivation were no longer present. In townships, there was no substantive change in associations or inferences for greenness and associations with community socioeconomic deprivation were no longer present. These variables could not be evaluated together in city census tracts due to insufficient overlap in distributions.

DISCUSSION

There is great interest in understanding geographic disparities in type 2 diabetes risk. If the primary causes of these differences were community-level factors, community-level interventions could have large impacts on diabetes risk. A strong theoretical basis, and growing empirical evidence, indicates that community features contribute to diabetes risk directly or through increased risk of obesity, such as social, built, and natural environments contributing to impacts on physical activity and stress [29-31]. The primary goal of this study was to evaluate geographic disparities in type 2 diabetes by evaluating four definitions of community across the full range from rural to urban. We then evaluated associations of community socioeconomic deprivation and greenness overall and in models stratified by community type, the latter greatly reducing the degree to which these associations could be confounded by other community features.

In the study region, the use of combined community type allowed us to carefully identify the location and scale of risk. Risk of new onset type 2 diabetes was highest in cities in smaller urban areas, followed by cities in major urban areas and boroughs in major and smaller urban areas. In addition, even after accounting for community type and features, county was independently associated with diabetes onset. While many prior studies have evaluated county differences in diabetes risk, none have also simultaneously evaluated communities. Our associations suggest that the risk factors that undergird U.S. geographic differences in diabetes likely exist at multiple, nested spatial scales. Some of the county associations were of high magnitude (e.g., exceeded 1.5 for protection or risk). Finally, there were consistent associations of higher community socioeconomic deprivation and lower greenness with higher diabetes risk, the former primarily in city census tracts, where average deprivation levels were higher, and the latter in all communities. We do not believe that the apparent lower diabetes risk in rural areas was due to less likely diagnosis due to lower access to health care, since, on average, individuals in the study, regardless of Medical Assistance status and community type, had high contact with the health care system.

We found several strong and consistent associations of individual-level characteristics. Non-white race, Hispanic ethnicity, and Medical Assistance status (a surrogate for low family socioeconomic status) were consistently associated with 1.3 to 1.7-fold increased odds of type 2 diabetes onset. Overall, the findings suggest that sociodemographic factors (race/ethnicity and individual-level socioeconomic status), urbanicity, higher community socioeconomic deprivation, and lower greenness, all of which co-occur in our region, were strong risk factors for type 2 diabetes.

Our findings on elevated risk of type 2 diabetes onset in urban areas is inconsistent with national studies that have reported higher crude prevalence estimates of type 2 diabetes in rural areas [32]. However, a study of the Behavioral Risk Factor Surveillance System found that after adjusting for individual-level socioeconomic measures, prevalence was higher in urban areas [33]. Geospatial predictors of diabetes risk likely vary by community and region; prior studies have reported, for example, that nine county-level measures of socioeconomic, race/ethnicity, and built environmental features explained up to 94% of the variation in type 2 diabetes prevalence in the Midwest, but very little variation in Pennsylvania [34].

The associations of greenness with diabetes were consistent with prior studies, but our results are the first to demonstrate robust findings across all types of communities while additionally controlling for county. The measurement of community features across community types may result in measures with different interpretations in different communities and regions; for example, agricultural, coniferous forest, and deciduous forest greenness are not evenly distributed and have different impacts on health [18].

Most prior studies of geographic disparities in diabetes have been cross-sectional, at the ecologic level, relying on self-reported diabetes, and focused on prevalent diabetes by county (too large and heterogeneous) or census tract (not experientially and behaviorally relevant). The current study avoided all these limitations. In addition, while many public health services are delivered at the county level, many potential interventions to address diabetes would need to be implemented at smaller scales and would not have county-wide impacts.

The study had some limitations. Although we adjusted for Medical Assistance health insurance as a surrogate for family socioeconomic status, there could still be residual confounding by individual-level income [27]. We did not measure behavioral mediators of the community definitions and features, such as physical activity or dietary intake. We could not account for residential selection bias, in which associations are due to reverse causation (if persons with individual-level risk factors for diabetes are more likely to reside in certain areas, by choice or opportunity). This can be a concern in studies of this type; social processes determine residence, so it can be difficult to distinguish individual-level characteristics from features of communities [35]. The residential stability and general population representativeness of our study population may mitigate these concerns.

The study had several strengths. Diabetes was objectively documented and verified with extensive biomarker and medical data. Temporality was appropriate for all independent variables. Study participants resided in a range of communities from urban to rural. We studied several approaches to community characterization at more relevant contextual scales than many prior studies and showed that smaller community contexts were associated with diabetes onset. Stratifying by community types limited bias from non-positivity [28].

The study findings provide important clues for the location (i.e., urban) and geographic scale (i.e., as localized as a square mile, the average area of boroughs and city census tracts) that identifies geospatial disparities in type 2 diabetes in

Pennsylvania. We speculate that, since risk was higher in urban areas, our findings may suggest a smaller role for the positive features of the food and physical activity environments present in these areas (e.g., greater access to grocery stores, more walkable neighborhoods, more commercial physical activity opportunity establishments) and a larger role for individual and community demographic and socioeconomic factors found in the same areas.

Author contributions

Manuscript authors contributed in the following ways: conception of work: BSS, MNP, KRS, CIM, GI, AGH; obtained funding: BSS, AGH; study design: BSS, JSP, KBR, AGH; data management and analysis: JSP, KBR, BSS, MNP, JD, KAM, AGH; results interpretation: BSS, MNP, KBR, JD, KAM, KRS, CIM, GI, AGH; initial manuscript writing: BSS, MNP, KAM, AGH; critical revision of manuscript, final approval, and accountable for their work: BSS, JSP, MNP, KBR, KAM, JD, KRS, CIM, GI, AGH.

Competing interests

All authors declared that they have no competing interests.

Funding

This publication was made possible by Cooperative Agreement Number DP006293 funded by the U.S. Centers for Disease Control and Prevention, Division of Diabetes Translation.

Data sharing

De-identified electronic health record data are available upon written request with IRB approval and a data use agreement. All community data are publicly available.

References

1. Centers for Disease Control and Prevention, *National Diabetes Statistics Report 2020: Estimtaes of Diabetes and Its Burden in the United States.*, U.S. Department of Health and Human Services, Editor. 2020, Centers for Disease Control and Prevention: Atlanta, GA.
2. Garcia, M.C., et al., *Reducing Potentially Excess Deaths from the Five Leading Causes of Death in the Rural United States.* MMWR Surveill Summ, 2017. **66**(2): p. 1-7.
3. Ford, E.S., et al., *Geographic variation in the prevalence of obesity, diabetes, and obesity-related behaviors.* Obes Res, 2005. **13**(1): p. 118-22.
4. Cunningham, S.A., et al., *County-level contextual factors associated with diabetes incidence in the United States.* Ann Epidemiol, 2018. **28**(1): p. 20-25 e2.
5. Rural Health Information Hub. *Why Diabetes is a Concern for Rural Communities* 2020 April 13, 2020]; Available from: https://www.ruralhealthinfo.org/toolkits/diabetes/1/rural-concerns.
6. Singh, G.K. and M. Siahpush, *Widening rural-urban disparities in all-cause mortality and mortality from major causes of death in the USA, 1969-2009.* J Urban Health, 2014. **91**(2): p. 272-92.
7. James, C.V., et al., *Racial/Ethnic Health Disparities Among Rural Adults - United States, 2012-2015.* MMWR Surveill Summ, 2017. **66**(23): p. 1-9.
8. Cosby, A.G., et al., *Growth and Persistence of Place-Based Mortality in the United States: The Rural Mortality Penalty.* Am J Public Health, 2019. **109**(1): p. 155-162.
9. Henning-Smith, C.E., et al., *Rural Counties With Majority Black Or Indigenous Populations Suffer The Highest Rates Of Premature Death In The US.* Health Aff (Millwood), 2019. **38**(12): p. 2019-2026.
10. Maier, W., et al., *Area level deprivation is an independent determinant of prevalent type 2 diabetes and obesity at the national level in Germany. Results from the National Telephone Health Interview Surveys 'German Health Update' GEDA 2009 and 2010.* PLoS One, 2014. **9**(2): p. e89661.
11. Muller, G., et al., *Regional and neighborhood disparities in the odds of type 2 diabetes: results from 5 population-based studies in Germany (DIAB-CORE consortium).* Am J Epidemiol, 2013. **178**(2): p. 221-30.
12. Jagai, J.S., et al., *Association between environmental quality and diabetes in the USA.* J Diabetes Investig, 2019.
13. Honold, J., et al., *Multiple environmental burdens and neighborhood-related health of city residents.* Journal of Environmental Psychology, 2012. **32**(4): p. 305-317.
14. Bennett, K.J., et al., *What Is Rural? Challenges And Implications Of Definitions That Inadequately Encompass Rural People And Places.* Health Aff (Millwood), 2019. **38**(12): p. 1985-1992.
15. Census Bureau. *Geography program: 2010 census urban and rural classification and urban area criteria.* . 2018 [cited 2020 January 5, 2020]; Available from: https://www.census.gov/programssurveys/geography/guidance/geoareas/urban-rural/2010-urbanrural.html.

16. Hirsch, A.G., et al., *Associations of Four Community Factors With Longitudinal Change in Hemoglobin A1c Levels in Patients With Type 2 Diabetes.* Diabetes Care, 2018. **41**(3): p. 461-468.

17. Cohen, S.A., et al., *A Closer Look at Rural-Urban Health Disparities: Associations Between Obesity and Rurality Vary by Geospatial and Sociodemographic Factors.* J Rural Health, 2017. **33**(2): p. 167-179.

18. James, W.L., *All rural places are not created equal: revisiting the rural mortality penalty in the United States.* Am J Public Health, 2014. **104**(11): p. 2122-9.

19. Astell-Burt, T., X. Feng, and G.S. Kolt, *Is neighborhood green space associated with a lower risk of type 2 diabetes? Evidence from 267,072 Australians.* Diabetes Care, 2014. **37**(1): p. 197-201.

20. Muller, G., et al., *Inner-city green space and its association with body mass index and prevalent type 2 diabetes: a cross-sectional study in an urban German city.* BMJ Open, 2018. **8**(1): p. e019062.

21. Casey, J.A., et al., *Unconventional Natural Gas Development and Birth Outcomes in Pennsylvania, USA.* Epidemiology, 2016. **27**(2): p. 163-72.

22. Lawrence, J.M., et al., *Validation of pediatric diabetes case identification approaches for diagnosed cases by using information in the electronic health records of a large integrated managed health care organization.* Am J Epidemiol, 2014. **179**(1): p. 27-38.

23. Zhong, V.W., et al., *Use of administrative and electronic health record data for development of automated algorithms for childhood diabetes case ascertainment and type classification: the SEARCH for Diabetes in Youth Study.* Pediatr Diabetes, 2014. **15**(8): p. 573-84.

24. Schwartz, B.S., et al., *Body mass index and the built and social environments in children and adolescents using electronic health records.* Am J Prev Med, 2011. **41**(4): p. e17-28.

25. Casey, J.A., et al., *Greenness and Birth Outcomes in a Range of Pennsylvania Communities.* Int J Environ Res Public Health, 2016. **13**(3).

26. Nau, C., et al., *Community socioeconomic deprivation and obesity trajectories in children using electronic health records.* Obesity (Silver Spring), 2015. **23**(1): p. 207-12.

27. Casey, J.A., et al., *Measures of SES for Electronic Health Record-based Research.* Am J Prev Med, 2018. **54**(3): p. 430-439.

28. Petersen, M.L., et al., *Diagnosing and responding to violations in the positivity assumption.* Stat Methods Med Res, 2012. **21**(1): p. 31-54.

29. Cox, M., et al., *Locality deprivation and Type 2 diabetes incidence: a local test of relative inequalities.* Soc Sci Med, 2007. **65**(9): p. 1953-64.

30. Maier, W., et al., *The impact of regional deprivation and individual socio-economic status on the prevalence of Type 2 diabetes in Germany. A pooled analysis of five population-based studies.* Diabet Med, 2013. **30**(3): p. e78-86.

31. James, P., et al., *A Review of the Health Benefits of Greenness.* Curr Epidemiol Rep, 2015. **2**(2): p. 131-142.

32. National Center for Chronic Disease Prevention and Health Promotion. *Division of Diabetes Translation At A Glance.* 2019 January 20, 2020]; Available from: https://www.cdc.gov/chronicdisease/resources/publications/aag/diabetes.htm.

33. O'Connor, A. and G. Wellenius, *Rural-urban disparities in the prevalence of diabetes and coronary heart disease.* Public Health, 2012. **126**(10): p. 813-20.

34. Hipp, J.A. and N. Chalise, *Spatial analysis and correlates of county-level diabetes prevalence, 2009-2010.* Prev Chronic Dis, 2015. **12**: p. E08.

35. Macintyre, S. and A. Ellaway, *Ecological approaches: rediscovering the role of the physical and social environment.*, in *Social Epidemiology.*, I. Kawachi and L. Berkman, Editors. 2000, Oxford University Press: New York. p. 332-348.

Table 1. Selected characteristics of individuals with diabetes and controls, frequency-matched to cases (5:1) on age, sex, and year of diagnosis or control selection date.

Variable	**Cases**	**Controls**	**p-value***
Unique persons	15,888	65,084	NA
Number	15,888	79,435	NA
Sex, female, n (COL %)	7798 (49.1)	38,988 (49.1)	matched
Age at diagnosis or control selection date, years, mean (SD)	54.9 (15.1)	54.9 (15.3)	matched
Age, years, categories, n (COL %)			matched
10 to < 20 years	304 (1.9)	1520 (1.9)	
20 to < 30 years	628 (4.0)	3140 (4.0)	
30 to < 40 years	1611 (10.1)	8055 (10.1)	
40 to < 50 years	3086 (19.4)	15,429 (19.4)	
50 to < 60 years	4286 (27.0)	21,428 (27.0)	
60 to < 70 years	3510 (22.1)	17,548 (22.1)	
70 to < 80 years	1737 (10.9)	8685 (10.9)	
80 to < 90 years	645 (4.1)	3225 (4.1)	
≥ 90 years	81 (0.5)	405 (0.5)	
Race, white, n (COL %)	15,429 (97.1)	77,867 (98.0)	< 0.001
Hispanic ethnicity, n (COL %)	369 (2.3)	1094 (1.4)	< 0.001
Primary care provider†, yes, n (%)	11,884 (74.8)	61,042 (76.9)	< 0.001
Year of diagnosis/encounter, n (COL %)			matched
2008	1761 (11.1)	8805 (11.1)	
2009	2019 (12.7)	10,095 (12.7)	
2010	1747 (11.0)	8735 (11.0)	
2011	1675 (10.5)	8373 (10.5)	
2012	1716 (10.8)	8579 (10.8)	
2013	1842 (11.6)	9209 (11.6)	
2014	1844 (11.6)	9220 (11.6)	
2015	1734 (10.9)	8669 (10.9)	
2016	1550 (9.8)	7750 (9.8)	
Setting of diagnosis/encounter, n (COL %)			< 0.001
Outpatient	12,068 (76.0)	73,998 (93.2)	
Medication order	1632 (10.3)	0 (0.0)	
Urgent care	165 (1.0)	2116 (2.7)	
Emergency department	1526 (9.6)	3068 (3.9)	
Inpatient	498 (3.1)	252 (0.3)	
Outpatient encounters in year before diagnosis or control selection date, mean (SD)	4.4 (5.1)	3.5 (4.1)	< 0.001
Outpatient encounters, total before diagnosis or control selection date, mean (SD)	35.9 (34.8)	35.2 (32.5)	0.01
Medical Assistance, % of time receiving, n (COL %)			< 0.001
< 50%	14,921 (93.9)	76,705 (83.7)	
≥ 50%	967 (6.1)	2730 (3.4)	
Outpatient encounters before diagnosis/encounter, mean (SD), by % of time receiving Medical Assistance			< 0.001

Variable	Cases	Controls	p-value*
0%	35.5 (34.1)	34.9 (32.1)	
0.1-24.9%	45.2 (40.7)	42.8 (38.3)	
25.0-74.9%	33.9 (35.8)	35.2 (33.6)	
75+%	29.1 (26.9)	27.7 (26.0)	
Duration from first contact with health system to diagnosis/control selection date, years, n (%)			0.72
Quartile 1 (2 to < 5 years)	1860 (11.7)	9466 (11.9)	
Quartile 2 (5 to < 8 years)	2571 (16.2)	12,646 (15.9)	
Quartile 3 (8 to < 12 years)	4700 (29.6)	23,665 (29.8)	
Quartile 4 (≥ 12 years)	6757 (42.5)	33,658 (42.4)	
Community socioeconomic deprivation, n (COL %)‡			< 0.001
Quartile 1	3001 (18.9)	17,329 (21.8)	
Quartile 2	4300 (27.1)	23,172 (29.2)	
Quartile 3	4217 (26.5)	20.328 (25.6)	
Quartile 4	4370 (27.5)	18,606 (23.4)	
Greenness, peak NDVI, in buffer, n (COL %) §			< 0.001
Tertile 1	5894 (37.1)	25,894 (32.6)	
Tertile 2	5023 (31.6)	26.751 (33.7)	
Tertile 3	4971 (31.3)	26,790 (33.7)	
Administrative community type of residence, n (COL %)			< 0.001
Borough	4621 (29.1)	21,756 (27.4)	
Census tract in city	1806 (11.4)	6548 (8.2)	
Township	9461 (59.6)	51,131 (64.4)	
Setting of residence, n (COL %)			< 0.001
Rural	6513 (41.0)	34,984 (44.0)	
Urbanized area	4906 (30.9)	23,423 (29.5)	
Urban cluster	4469 (28.1)	21,028 (26.5)	

Abbreviations: COL = column; NDVI = normalized difference vegetation index; SD = standard deviation.

* Because controls could be in these comparisons more than once, methods were used for significance testing that accounted for this, including inverse-probability weighted regression for time-invariant characteristics, mixed-effect regression for time-varying continuous (linear), binary (logistic), and count (Poisson) characteristics, and multinomial logistic regression with robust standard errors for polytomous time-varying characteristics. In the weighted analyses, weights were the number of appearances in the analysis (implemented with a dataset having only one record per person).

† According to Geisinger's primary care provider lists.

‡ Quartile cutoffs were defined within the three time periods; the range of values for Q1, Q2, Q3, and Q4 were -18.33 to -1.96; -1.99 to -0.015; 0.005 to 2.05; and 2.11 to 12.4.

§ The range of values in T1, T2, and T3 were 0.07 to 0.627, 0.63 to 0.756, and 0.76 to 0.94, respectively.

Table 2. Adjusted* associations of community and community feature variables **from separate models** with new onset type 2 diabetes status.

Variable	OR (95% CI)
Community types	
Model 1: Administrative community type	
Township	1.0
Borough	1.10 (1.04, 1.16)
City census tract	1.34 (1.25, 1.44)
Model 2: Residential location, urban/rural	
Rural	1.0
Urbanized area	1.14 (1.08, 1.21)
Urban cluster	1.04 (0.98, 1.11)
Model 3: Combined location†	
TS/rural	1.0
TS/UC	1.00 (0.92, 1.08)
TS/UA	1.06 (0.98, 1.16)
B+CCT/rural	1.04 (0.95, 1.15)
B/UC	1.09 (1.01, 1.18)
B/UA	1.15 (1.06, 1.25)
CCT/UC	1.41 (1.22, 1.62)
CCT/UA	1.33 (1.22, 1.45)
Model 4: County ‡	
Luzerne	1.0
Blair	0.73 (0.57, 0.95)
Centre	0.84 (0.75, 0.94)
Juniata	1.19 (1.00, 1.40)
Lackawanna	1.19 (1.07, 1.31)
Lebanon	0.39 (0.16, 0.93)
Monroe	0.78 (0.69, 0.88)
Schuylkill	0.85 (0.78, 0.92)
Sullivan	0.60 (0.45, 0.81)
Union	0.77 (0.64, 0.93)
Community features, all communities combined	
Model 5: community socioeconomic deprivation, quartiles §	
1	0.82 (0.76, 0.88)
2	0.87 (0.81, 0.93)
3	0.89 (0.83, 0.96)
4	1.0
Model 6: greenness (normalized difference vegetation index) ||	
1	1.0
2	0.88 (0.85, 0.93)
3	0.84 (0.80, 0.88)

* Logistic regression models using generalized estimating equations with robust standard errors; one community or community feature variable was in the model at a time; models adjusted for sex, race (white vs. non-white), ethnicity (Hispanic vs. non-Hispanic), age (age, age^2, age^3), and Medical Assistance status.

† This is a combination of administrative community type and residential location (urban/rural); TS = township, B = borough, CCT = city census tract, UA = urbanized area, UC = urban cluster; the few persons in CCT/rural were combined with B/rural.

‡ Only counties with confidence interval excluding 1.0 are shown in table. Luzerne County was selected as the reference group because it is the most populous county in the study region.

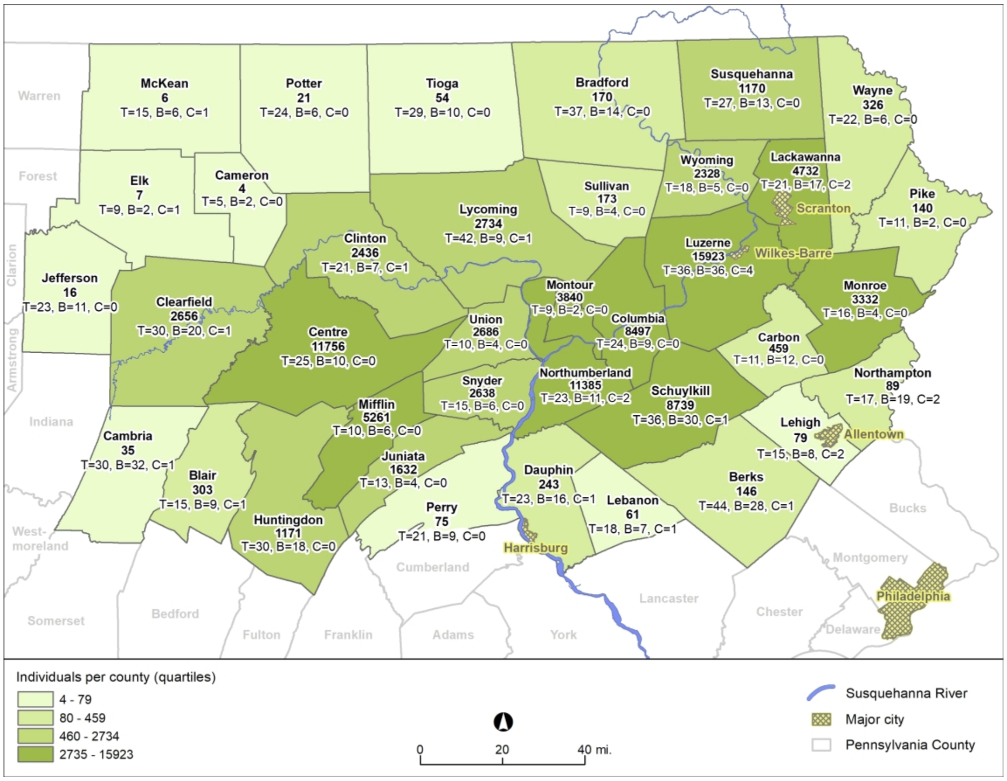
§ Quartile cutoffs were defined within the three time periods; the range of values for persons in Q1, Q2, Q3, and Q4 were -25.06 to -1.82; -1.99 to 0.10; 0.005 to 2.05; and 1.89 to 12.4, respectively.

|| The range of values in T1, T2, and T3 were 0.07 to 0.627, 0.63 to 0.756, and 0.76 to 0.94, respectively.

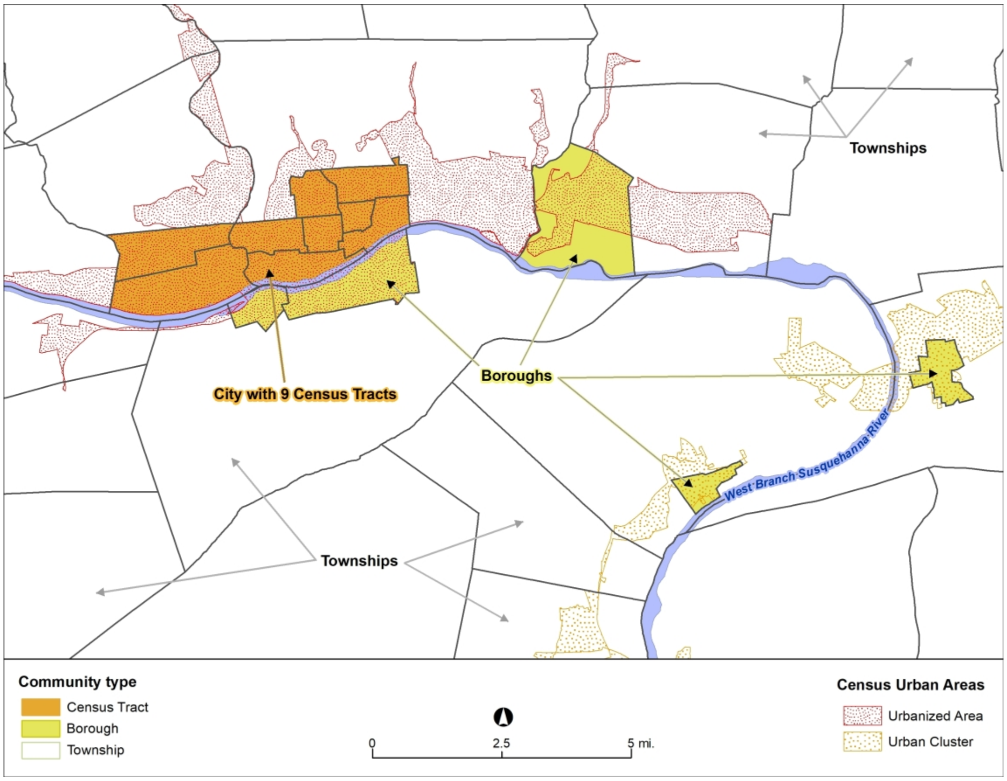
Figure Captions

Figure 1. Distribution of study individuals and administrative community types by county in study region. The bolded number is the number of individuals; T, B, and C identify the number of townships, boroughs, and city census tracts within each county that were included in the analysis.

Figure 2. Areas along the Susquehanna River in Lycoming County, Pennsylvania from Williamsport (city) and South Williamsport (borough) to Montoursville (borough), Muncy (borough), and Montgomery (borough), showing relations between administrative community types (townships, boroughs, and city census tracts) and urbanized areas, urban clusters, and rural areas. Both sets of these administrative boundaries were used in the analysis.



279x215mm (300 x 300 DPI)



279x215mm (300 x 300 DPI)

Online Supplement

Table S1. Diabetes case finding using EHR data.

Must meet at least one of the following criteria:

1. At least two separate encounter dates (inpatient, outpatient, emergency department) with type 2 diabetes diagnosis codes (ICD-9, ICD-10, or electronic diagnosis group [EDG]).
   a. Excluded if had ≥ 10 years of type 1 diagnoses and < five years with type 2 diagnoses.
   b. Excluded if < 10 years of age at first diabetes diagnosis.
2. At least one diabetes medication order, other than metformin or acarbose if female. Metformin combination medications were included.
   a. Excluded if first diabetes medication order was prior to age 10 years.
3. At least one encounter with type 2 diabetes diagnosis and an abnormal laboratory value (random glucose ≥ 200 mg/dl; fasting glucose ≥ 126 mg/dl; or hemoglobin A1c ≥ 6.5%).
   a. Excluded if had ≥ 10 years of type 1 diagnoses and < five years with type 2 diagnoses.

- The date of onset was assigned as the earliest date with any evidence of diabetes (e.g., had generic diabetes diagnoses that were not used for definition #1, or had abnormal laboratory value that was not accompanied by a diagnosis so did not meet definition #3).

Notes:
a) To meet criteria #2 or #3, criterion had to occur > 9 months prior to or > 1 month after delivery of child (to avoid gestational diabetes). Gestational diabetes was not an exclusion if the individual subsequently developed type 2 diabetes. Date of onset was assigned as when the person met the type 2 diabetes criterion; and
b) EDG codes are used in Epic EHR software (Epic Systems Corporation, Verona, WI) and often have higher specificity and greater detail.
c) Of the 15,888 diabetes cases: 11,944 met criterion 1; 10,183 met criterion 2; 12,552 met criterion 3; 7008 met all three; and 4775 met at least two.
d) Because metformin can be used for pre-diabetes, we evaluated how many persons could have had this diagnosis instead of diabetes in our diabetes onset definition. Of the 1579 men who met only definition #2, between 544 (3.4%) and 1207 (7.6%) may have had pre-diabetes instead of diabetes, depending on how longitudinal information on diagnoses, medications, medication indications, and abnormal laboratory results were used and interpreted.

Table S2. Selected characteristics of study individuals and communities by administrative community type.

Variables	Borough	Census Tract	Township
By community type (n = 1070 communities)			
Number (%), total	291 (27.2)	146 (13.6)	633 (59.2)
Number (%), among cases	224 (27.6)	107 (13.2)	482 (59.3)
Number (%), among controls	278 (26.9)	137 (13.2)	620 (59.9)
Counties with at least one resident in community type, n	35	16	37
Counties with at least 20 residents in community type, n	27	9	32
Community measures, by community type (n = 1070 communities)			
Area, square miles, mean (SD)	1.72 (2.32)	1.20 (3.52)	29.4 (18.1)
Community socioeconomic deprivation, mean (SD)	-0.09 (2.99)	4.17 (3.80)	-1.15 (2.71)
Population density, persons per square mile, mean (SD)	2094.7 (1642.3)	6594.5 (5014.6)	157.5 (279.4)
Developed land, % (SD)	37.2 (22.6)	72.6 (23.0)	3.66 (7.35)
Intersection density per square mile, mean (SD)	120.6 (86.1)	208.5 (117.0)	13.34 (14.77)
By participant (n = 95,323 individuals)			
Cases, n (%) (total = 15,888)	4621 (29.1)	1806 (11.4)	9461 (59.5)
Controls, n (%) (total = 79,435)	21,756 (27.4)	6548 (8.2)	51,131 (64.4)
Age at diabetes onset or control selection date, years, mean (SD)	54.4 (15.9)	52.7 (16.1)	55.3 (14.8)
Sex, female, n (%)	13,329 (50.2)	4449 (53.3)	29,098 (48.0)
Race, white, n (%)	245,963 (98.4)	7873 (94.2)	59,460 (98.1)
Ethnicity, Hispanic, n (%)	353 (1.3)	430 (5.2)	680 (1.1)
Body mass index, kg/m^2, mean (SD)	30.6 (7.47)	30.9 (7.96)	30.3 (6.94)
Medical Assistance, % of time, mean (SD)	5.9 (17.9)	10.3 (23.2)	3.3 (13.5)
Medical Assistance, ever*, n (%)	3311 (12.6)	1692 (20.3)	4311 (7.1)
Contact with health system before diagnosis/control selection date, years, mean (SD)	12.7 (4.37)	12.1 (4.57)	12.9 (4.34)
Charlson index, mean (SD)	1.75 (1.83)	1.64 (1.78)	1.76 (1.78)
Greenness, peak NDVI, in buffer, mean (SD)	0.61 (0.11)	0.51 (0.10)	0.73 (0.10)
Urban status by UA and UC boundaries, col %			
Rural	11.5	0.1	63.5
Urbanized area (UA)	43.3	64.8	19.0
Urban cluster (UC)	45.3	35.1	17.5

Abbreviations: NDVI = normalized difference vegetation index; SD = standard deviation.
* At least one encounter that used Medical Assistance for health insurance.

Table S3. Mean outpatient encounters among cases and controls by community type and Medical Assistance status.

	Cases, n = 15,888			Controls, n = 79,435		
Variable	**Boroughs n = 4621**	**City Census Tracts n = 1806**	**Townships n = 9461**	**Boroughs n = 21,756**	**City Census Tracts n = 6548**	**Townships n = 51,131**
Outpatient encounters, total before diagnosis, mean (SD)	35.9 (34.8)	31.6 (32.1)	36.8 (35.2)	35.7 (33.8)	33.5 (32.8)	35.2 (31.8)
Outpatient encounters before diagnosis, mean (SD), by Medical Assistance status (% time receiving)						
0%	35.2 (33.9)	30.9 (31.2)	36.3 (34.6)	35.1 (33.1)	32.9 (32.1)	35.0 (31.7)
0.1-24.9%	47.7 (41.3)	41.8 (44.3)	44.7 (39.0)	44.2 (40.7)	40.0 (39.9)	42.6 (36.1)
25.0-74.9%	32.5 (34.5)	29.3 (25.4)	37.1 (40.2)	37.3 (36.8)	33.6 (32.4)	34.2 (31.4)
75+%	30.6 (28.9)	34.2 (21.0)	30.7 (28.0)	27.7 (28.5)	27.9 (28.4)	27.7 (22.6)
SD = standard deviation						

Medical Profile of Cases and Controls

To evaluate our categorization of diabetes cases and controls, we examined a number of biomarkers and other measures of relevance to diabetes, dysglycemia, and other cardio-metabolic risk factors development that were available in the EHR, including hemoglobin A1c (HbA1c), lipids (cholesterol and triglycerides), blood glucose (fasting and unspecified), and body mass index (BMI) (**Online Supplement Table S4**). Fasting blood glucose was measured in the year before the diabetes onset or control dates in 24% of cases and 29% of controls. Interestingly, the mean value was higher in the year before diagnosis in persons who would develop diabetes compared to those who would not, 108.5 vs. 95.8 mg/dL ($p < 0.001$). In the year after diagnosis or control dates, fasting blood glucose was available in 58% of cases and 30% of controls, and mean levels were much higher in cases compared to controls (147.9 vs. 95.9, $p < 0.001$). HbA1c, triglycerides, unspecified blood glucose, and BMI all evidenced similar patterns (**Online Supplement Table S4**). In the year before and after diagnosis, most cases and controls had BMI measured, with a much higher mean in cases compared to controls before and after diagnosis.

Table S4. Selected laboratory and other biometric values comparing new onset type 2 diabetes cases and controls without diabetes.

Variable	Cases	Controls
Number	15,888	79,435
Hemoglobin A1c (HbA1c)		
# in year before diagnosis or control selection date per person, number of persons (%) with		
0 values	13,618 (85.7)	75,731 (95.3)
1 value	1801 (11.3)	3257 (4.1)
2+ values	469 (3.0)	447 (0.6)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	2270 (14.3)	3704 (4.7)
HbA1c %, mean (SD)	5.9 (0.4)	5.6 (0.4)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	11,990 (75.5)	3839 (4.8)
HbA1c %, mean (SD)	7.5 (2.0)	5.6 (0.4)
LDL cholesterol		
# in year before diagnosis or index date per person, number of persons (%) with		
0 values	10,155 (63.9)	46,485 (58.5)
1 value	4068 (25.6)	23,737 (29.9)
2+ values	1665 (10.5)	9213 (11.6)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	5733 (36.1)	32,950 (41.5)
LDL-cholesterol, mg/dL, mean (SD)	107.2 (35.6)	109.6 (33.0)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	11,726 (73.8)	34,223 (43.1)
LDL-cholesterol, mg/dL, mean (SD)	108.5 (36.7)	111.1 (33.7)
Triglycerides		
# in year before diagnosis or index date per person, number of persons (%) with		
0 values	10,529 (66.3)	48,714 (61.3)
1 value	3869 (24.4)	22,585 (28.4)
2+ values	1490 (9.4)	8136 (10.2)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	5359 (33.7)	30,721 (38.7)
Triglycerides, mg/dL, mean (SD)	188.7 (131.7)	133.7 (81.2)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	11,207 (70.5)	31,663 (39.9)
Triglycerides, mg/dL, mean (SD)	216.5 (244.8)	135.0 (86.8)
Glucose, fasting		
# in year before diagnosis or index date per person, # of persons (%) with		
0 values	12,139 (76.4)	56,198 (70.8)
1 value	2968 (18.7)	19,023 (24.0)
2+ values	781 (5.0)	4214 (5.3)

Variable	Cases	Controls
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	3749 (23.6)	23,237 (29.3)
Glucose, mg/dL, mean (SD)	108.5 (11.8)	95.8 (9.3)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	9259 (58.3)	24,105 (30.3)
Glucose, mg/dL, mean (SD)	147.9 (60.9)	95.9 (9.3)
Glucose, unspecified		
# in year before diagnosis or index date per person, # persons (%) with		
0 values	9913 (62.4)	54,258 (68.3)
1 value	3115 (19.6)	15,293 (19.3)
2+ values	2860 (18.0)	9884 (12.4)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	5975 (37.6)	25,177 (31.7)
Glucose, mg/dL, mean (SD)	124.6 (28.2)	97.7 (15.5)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	10,833 (68.2)	27,779 (35.0)
Glucose, mg/dL, mean (SD)	170.7 (95.2)	98.4 (16.5)
Body mass index (BMI)		
# in year before diagnosis or index date per person, mean (SD)	3.1 (4.1)	2.4 (3.2)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	11,237 (70.7)	54,733 (68.9)
BMI, kg/m^2, mean (SD)	36.2 (8.4)	29.3 (6.4)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	13,957 (87.9)	65,084 (81.9)
BMI, kg/m^2, mean (SD)	36.0 (8.4)	29.3 (6.4)

Table S5. Adjusted* associations of selected independent variables with type 2 diabetes status stratified by administrative community type.

	Stratified by Administrative Community Type			**Stratified by Administrative Community Type**		
	Boroughs	**City Census Tracts**	**Townships**	**Boroughs**	**City Census Tracts**	**Townships**
Variable	**Model 1a** **OR (95% CI)**	**Model 1b** **OR (95% CI)**	**Model 1c** **OR (95% CI)**	**Model 2a** **OR (95% CI)**	**Model 2b** **OR (95% CI)**	**Model 2c** **OR (95% CI)**
Race						
White	1.0	1.0	1.0	1.0	1.0	1.0
All others	1.44 (1.12, 1.94)	1.30 (1.05, 1.60)	1.36 (1.14, 1.61)	1.43 (1.12, 1.84)	1.28 (1.04, 1.58)	1.35 (1.14, 1.61)
Ethnicity						
Non-Hispanic	1.0	1.0	1.0	1.0	1.0	1.0
Hispanic	1.50 (1.16, 1.94)	1.33 (1.02, 1.72)	1.52 (1.16, 1.97)	1.50 (1.16, 1.94)	1.32 (1.02, 1.71)	1.52 (1.17. 1.97)
Medical Assistance						
< 50% of time	1.0	1.0	1.0	1.0	1.0	1.0
50+% of time	1.66 (1.47, 1.86)	1.46 (1.26, 1.70)	1.83 (1.61, 2.09)	1.66 (1.48, 1.86)	1.48 (1.27, 1.72)	1.83 (1.61, 2.09)
CSD **						
Q1	0.88 (0.77, 1.01)	0.75 (0.56, 1.00)	0.93 (0.84, 1.02)			
Q2	0.96 (0.84, 1.08)	0.77 (0.63, 0.94)	0.97 (0.89, 1.06)			
Q3	0.98 (0.87, 1.10)	0.78 (0.67, 0.91)	0.98 (0.89, 1.07)			
Q4	1.0	1.0	1.0			
NDVI, 1250x1250m †						
T1				1.0	1.0	1.0
T2				0.93 (0.87, 0.99)	0.76 (0.64, 0.90)	0.93 (0.87, 0.99)
T3				0.85 (0.76, 0.96)	0.76 (0.50, 1.17)	0.90 (0.84, 0.96)

Abbreviations: CSD = community socioeconomic deprivation; NDVI = normalized difference vegetation index;

* Logistic regression models using generalized estimating equations with robust standard errors; also adjusted for sex and age (age, age^2, age^3).

** Quartile cutoffs were defined within the three time periods; the range of values for persons in Q1, Q2, Q3, and Q4 were -25.06 to -1.82; -1.99 to 0.10; 0.005 to 2.05; and 1.89 to 12.4, respectively.

† The range of values in T1, T2, and T3 were 0.07 to 0.627, 0.63 to 0.756, and 0.76 to 0.94, respectively.

Table S6. Adjusted* associations of selected independent variables with type 2 diabetes status stratified by administrative community type with county and community socioeconomic deprivation ***OR*** greenness.

	Stratified by Administrative Community Type		
	Boroughs	City Census Tracts	Townships
Variable	Model 1 OR (95% CI)	Model 1 OR (95% CI)	Model 1 OR (95% CI)
Model 1 – with county and community socioeconomic deprivation (CSD)			
Race			
White	1.0	1.0	1.0
All others	1.45 (1.13, 1.86)	1.31 (1.06, 1.62)	1.39 (1.16, 1.66)
Ethnicity			
Non-Hispanic	1.0	1.0	1.0
Hispanic	1.49 (1.15, 1.92)	1.32 (1.02, 1.71)	1.55 (1.18, 2.04)
Medical Assistance			
< 50% of time	1.0	1.0	1.0
50+% of time	1.66 (1.47, 1.87)	1.48 (1.28, 1.72)	1.85 (1.62, 2.11)
Community socioeconomic deprivation, quartiles			
Q1	0.87 (0.76, 0.996)	0.71 (0.52, 0.95)	0.91 (0.82, 0.99)
Q2	0.93 (0.83, 1.06)	0.78 (0.65, 0.95)	0.96 (0.88, 1.05)
Q3	0.97 (0.87, 1.09)	0.79 (0.67, 0.93)	0.98 (0.90, 1.07)
Q4	1.0	1.0	1.0
County			
Luzerne	1.0	1.0	1.0
Blair	**0.64** (0.51, 0.81)	0.62 (0.23, 1.64)	0.86 (0.61, 1.21)
Clearfield	1.00 (0.82, 1.24)	**0.76** (0.66, 0.87)	0.97 (0.82, 1.15)
Dauphin	0.90 (0.56, 1.45)	**2.81** (1.47, 5.37)	1.43 (0.96, 2.15)
Juniata	**1.68** (1.22, 2.31)	NA†	1.18 (0.99, 1.41)
Lackawanna	1.12 (0.96, 1.37)	**1.23** (1.06, 1.43)	1.13 (0.93, 1.38)
Lehigh	**18.2** (2.00, 165.1)	2.00 (0.85, 4.68)	0.66 (0.26, 1.65)
Mifflin	**1.20** (1.00, 1.43)	NA	1.06 (0.93, 1.21)
Monroe	**0.73** (0.59, 0.91)	NA	**0.85** (0.74, 0.98)
Perry	**3.16** (1.34, 7.47)	NA	0.96 (0.51, 1.83)
Potter	**4.90** (4.42, 5.43)	NA	0.71 (0.15, 3.31)
Schuylkill	0.91 (0.80, 1.02)	0.93 (0.80, 1.07)	**0.82** (0.73, 0.91)
Snyder	**0.84** (0.72, 0.98)	NA	1.01 (0.88, 1.16)
Sullivan	0.63 (0.38, 1.07)	NA	**0.65** (0.47, 0.90)
Union	0.84 (0.53, 1.34)	NA	**0.80** (0.66, 0.98)
Wayne	**3.36** (1.83, 6.16)	NA	0.96 (0.59, 1.58)
Wyoming	**0.86** (0.76, 0.96)	NA	**1.15** (1.00, 1.32)
Model 2 – same as Model 1, but with NDVI not CSD, with county; only NDVI associations are shown			
Normalized difference vegetation index (NDVI)			
T1	1.0	1.0	1.0
T2	0.91 (0.85, 0.98)	0.77 (0.64, 0.92)	0.93 (0.87, 0.99)
T3	0.85 (0.75, 0.97)	0.76 (0.48, 1.19)	0.90 (0.84, 0.97)

* Logistic regression models using generalized estimating equations with robust standard errors; also adjusted for sex and age (age, age^2, age^3). Counties with at least one association that excluded 1.0 in confidence interval included in table (37 counties were included in total; 36 county indicators vs. Luzerne County as reference).

† NA = these counties did not have city minor civil divisions or did not converge due to small numbers.

STROBE Statement—Checklist of items that should be included in reports of ***case-control studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	**3**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	**3**
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	**5-6**
Objectives	3	State specific objectives, including any prespecified hypotheses	**6, 9**
Methods			
Study design	4	Present key elements of study design early in the paper	**6, 7**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	**7**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls	**7, 8**
		(*b*) For matched studies, give matching criteria and the number of controls per case	**7**
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	**7-9**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	**7-9**
Bias	9	Describe any efforts to address potential sources of bias	**9, 10**
Study size	10	Explain how the study size was arrived at	**7**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	**9, 10**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	**9, 10**
		(*b*) Describe any methods used to examine subgroups and interactions	**9, 10**
		(*c*) Explain how missing data were addressed	**9, 10**
		(*d*) If applicable, explain how matching of cases and controls was addressed	**9, 10**
		(*e*) Describe any sensitivity analyses	**9, 10**
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	**7**
		(b) Give reasons for non-participation at each stage	**NA**
		(c) Consider use of a flow diagram	**NA**
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	**21, 22**
		(b) Indicate number of participants with missing data for each variable of interest	**NA**
Outcome data	15*	Report numbers in each exposure category, or summary measures of exposure	**21, 22**

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	**23, 24**
		(*b*) Report category boundaries when continuous variables were categorized	**23, 24**
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	**NA**
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	**13**
Discussion			
Key results	18	Summarise key results with reference to study objectives	**13, 14**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	**16**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	**16, 17**
Generalisability	21	Discuss the generalisability (external validity) of the study results	**16, 17**
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	**2, 6**

*Give information separately for cases and controls.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

BMJ Open

Association of community types and features in a case-control analysis of new onset type 2 diabetes across a diverse geography in Pennsylvania

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-043528.R1
Article Type:	Original research
Date Submitted by the Author:	17-Dec-2020
Complete List of Authors:	Schwartz, BS; Johns Hopkins University Bloomberg School of Public Health, Department of Environmental Health and Engineering Pollak, Jonathan; Johns Hopkins University Bloomberg School of Public Health, Department of Environmental Health and Engineering Poulsen, Melissa; Geisinger, Population Health Sciences Bandeen-Roche, Karen; Johns Hopkins University Bloomberg School of Public Health, Biostatistics Moon, Katherine; Johns Hopkins Bloomberg School of Public Health Center for Teaching and Learning, Environmental Health & Engineering DeWalle, Joseph; Geisinger, Population Health Sciences Siegel, Karen; Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion Mercado, Carla; Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion Imperatore, Giuseppina ; Centers for Disease Control and Prevention, Division of Diabetes Tranlsation Hirsch, Annemarie; Geisinger Health System, Department of Epidemiology and Health Services Research
<b>Primary Subject Heading</b>:	Diabetes and endocrinology
Secondary Subject Heading:	Epidemiology
Keywords:	DIABETES & ENDOCRINOLOGY, EPIDEMIOLOGY, PUBLIC HEALTH

Association of community types and features in a case-control analysis of new onset type 2 diabetes across a diverse geography in Pennsylvania

Brian S. Schwartz,[1,2,4,5] Jonathan S. Pollak,[1] Melissa N. Poulsen,[5] Karen Bandeen-Roche,[3] Katherine A. Moon,[1] Joseph DeWalle,[5] Karen R. Siegel,[6] Carla I. Mercado,[6] Giuseppina Imperatore,[6] Annemarie G. Hirsch[1,5]

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[1] Department of Environmental Health and Engineering

[2] Department of Epidemiology

[3] Department of Biostatistics

Johns Hopkins School of Medicine, Baltimore, MD

[4] Department of Medicine

Geisinger, Danville, PA

[5] Department of Population Health Sciences

Centers for Disease Control and Prevention, Atlanta, GA

[6] Division of Diabetes Translation, National Center for Chronic Disease Prevention and Health Promotion

Corresponding author: Brian S. Schwartz, Department of Environmental Health and Engineering, 615 N. Wolfe Street, Room W7041, Baltimore, MD, 21205, voice 410-955-4158, email bschwar1@jhu.edu.

Word count: abstract = 287, manuscript = 3122 | **Tables and figures** = 4 | **Online supplement tables** = 6 | **References** = 35

Running title: Geography of type 2 diabetes in Pennsylvania

Abstract

Objectives: To evaluate associations of community types and features with new onset type 2 diabetes in diverse communities. Understanding the location and scale of geographic disparities can lead to community-level interventions.

Design: Nested case-control study within the open dynamic cohort of health system patients.

Setting: Large, integrated health system in 37 counties in central and northeastern Pennsylvania, USA.

Participants and analysis: We used electronic health records to identify persons with new-onset type 2 diabetes from 2008–2016 (n = 15,888). Persons with diabetes were age, sex, and year matched (1:5) to persons without diabetes (n = 79,435). We used generalized estimating equations to control for individual-level confounding variables, accounting for clustering of persons within communities. Communities were defined as 1) townships, boroughs, and city census tracts; 2) urbanized area (large metro), urban cluster (small cities and towns), and rural; 3) combination of the first two; and 4) county. Community socioeconomic deprivation and greenness were evaluated alone and in models stratified by community types.

Results: Borough and city census tract residence (vs. townships) were associated (odds ratio [95% confidence interval]) with higher odds of type 2 diabetes (1.10 [1.04-1.16] and 1.34 [1.25-1.44], respectively). Urbanized areas (vs. rural) also had increased odds of type 2 diabetes (1.14 [1.08-1.21]). In the combined definition, the strongest associations (vs. townships in rural areas) were city census tracts in urban clusters (1.41 [1.22-1.62]) and city census tracts in urbanized areas (1.33 [1.22-1.45]). Higher community socioeconomic deprivation and lower greenness were each associated with increased odds.

Conclusions: Urban residence was associated with higher odds of type 2 diabetes than for other areas. Higher community socioeconomic deprivation in city census tracts and lower greenness in all community types were also associated with type 2 diabetes.

Strengths and limitations of this study

- Type 2 diabetes, with a large sample size, was objectively documented and verified or excluded with extensive biomarker and medical data.
- Temporality was appropriate for all independent variables.
- We studied several approaches to community characterization at more relevant contextual scales than many prior studies in a range of communities from urban to rural.
- We did not measure behavioral mediators of the community definitions and features, such as physical activity or dietary intake.
- We could not account for residential selection bias, but the residential stability and general population representativeness of our study population may mitigate these concerns.

INTRODUCTION

Diabetes is a common and costly chronic disease; in the U.S. in 2018, over 34 million individuals had diabetes, with annual spending exceeding $320 billion [1]. Diabetes occurrence varies by race/ethnicity and also evidences geographic disparities [2, 3]; prevalence by county in the U.S. varies over a 7-fold range [4]. Studies report that diabetes is 17% more prevalent in rural than urban areas [5], consistent with rural health disparities for other chronic conditions [6, 7], attributed to sociodemographic factors (e.g., higher poverty, older populations) and barriers to health care access [8, 9].

Community characteristics that may underlie observed geographic disparities in type 2 diabetes include land use (e.g., walkable vs. automobile dependent), fitness, food, and social (e.g., deprivation, disorganization) environments; greenspace (i.e., natural environments); and air pollution. Some of these are diabetogenic and others protective [10-12]. Community characteristics co-occur in patterns that differ by **community type** (e.g., higher population density co-occurs with higher deprivation and food availability and lower automobile dependence and greenness). Simultaneously evaluation and control of these domains across community types can be problematic due to limited and non-overlapping distributions that make independent attribution of disease risk to specific domains difficult [13]. An alternative is to use carefully defined community types to first identify the **location** and **geographic scale** of type 2 diabetes risk [14-17]. These community types should reduce within community variation and maximize between community differences. Subsequent analyses can then stratify by community type and evaluate well-characterized **community features** in relation to type 2 diabetes risk.

Residential development patterns reflect a continuum from rural to urban with variation by many community features [18]. The U.S. Census Bureau defines *urbanized areas* as dense settlements with 50,000 or more residents, *urban clusters* as areas with 2500–50,000 residents, and all others as *rural* [19]. In Pennsylvania, communities are defined administratively as townships, boroughs, and cities using census minor civil division boundaries [20]. In combination, these two definitions provide an opportunity to evaluate experientially and behaviorally relevant geographies as well as to further subdivide the broad category of "rural," which includes a range of communities that vary in their associations with health outcomes [21, 22].

We evaluated four definitions of community across a range of community types from rural to urban in a 37-county region of Pennsylvania, in relation to type 2 diabetes onset to inform more robust study of the community-level features that may underlie type 2 diabetes risk. Next, because higher community socioeconomic deprivation and lower greenness have been consistently associated with higher risk of type 2 diabetes [23, 24], we evaluated associations with these features overall and within community types.

METHODS

Study Population and Design

This study was conducted by Geisinger-Johns Hopkins Bloomberg School of Public Health, one of four academic research centers in the Diabetes LEAD (Location, Environmental Attributes, and Disparities) Network (http://diabetesleadnetwork.org/), a collaboration funded by the Centers for Disease Control and Prevention dedicated to providing scientific evidence to develop targeted interventions and policies to prevent type 2 diabetes and related health outcomes across the U.S. The study was approved by the Geisinger Institutional Review Board under waivers of consent and assent to use electronic health record (EHR) data.

Using previously reported methods [20], we used Geisinger EHR data from 1.6 million individuals to identify new onset type 2 diabetes from 2008–2016. Individuals represent the general population in the region with high residential stability [25]. The study area included 37 counties in Pennsylvania (**Figure 1**). These data were used in a nested case-control study.

Patient and Public Involvement

Patients and public representatives were not involved in the development of the study. Study results will be disseminated through Geisinger's Environmental Health Institute in its website (https://www.geisinger.edu/research/departments-and-centers/environmental-health-institute) and communications to Geisinger patients and the public.

Identification of New Onset Type 2 Diabetes Cases and Controls

Persons with type 2 diabetes (n = 15,888) were identified using diabetes encounter diagnoses, medication orders, and laboratory test results (**Online Supplement Table S1**). EHR algorithms can identify diabetes with high sensitivity, specificity, and positive predictive value [26, 27]. Controls (n = 79,435, with 65,084 unique persons), persons who never met any of the diabetes criteria used for cases, were randomly selected with replacement and frequency-matched to cases (5:1) on age, sex, and year of encounter. To ensure that we could identify diabetes if present, we required at least two encounters on different days with a primary care provider prior. To ensure diabetes was new onset, persons had to have at least one encounter with the health system at least two years prior without evidence of diabetes.

Community Types and Community Features

Addresses at last contact with the health system were geocoded using ArcGIS version 10.4 (ESRI Inc., Redlands, CA). We used four definitions of community, defined as *administrative community type*, *urban/rural status*, *combined community type*, and *county*, to evaluate different spatial scales and a range of characterizations of the size and urbanicity of these areas (**Figure 2**). First, using minor civil divisions and census tract boundaries, we categorized study communities into townships, boroughs, and city census tracts, as previously reported [28], referred to as *administrative community type*. Townships range from agriculturally-focused rural areas to low density suburbs; boroughs are walkable small towns of 5,000 to 10,000 persons with a core area of gridded streets; and cities are medium-sized urban areas (largest is Scranton–Wilkes-Barre–Hazleton Metropolitan Statistical Area, 97th in U.S. by population). Second, we used U.S. Census Bureau's urbanized areas and urban clusters to define residential addresses as "major urban," "smaller urban," and "rural" [19], referred to as *urban/rural status*. Third, to evaluate community at a more granular level, we combined the first and second categorizations, referred to as *combined community type*. This resulted in eight groups (city census tract/rural had few residences so were combined with borough/rural; township/rural was the reference group). Fourth, because most prior research of geographic disparities in diabetes evaluated counties, which are much larger geographies, we evaluated counties alone and after stratification by administrative community type.

We evaluated two time-varying community features. Peak (16-day composite in early July of each year) normalized difference vegetation index (NDVI, referred to as greenness) was evaluated in 1250m squares around residences in the prior year [29]. We measured community socioeconomic deprivation using a previously described scale [30], the sum of z-transformed values of six indicators identified from a factor analysis (proportion unemployed, less than a high school education, below poverty level, on public assistance, not in the workforce, and without a car), using data from the Decennial Census (2000 only) and American Community Survey (2006-2010, 2011-2015). The scale was assigned as the closest measure prior to the year of onset/encounter.

Statistical Analysis

The goals of the analysis were: 1) evaluate four definitions of community in relation to odds of type 2 diabetes onset; 2) evaluate two community features, community socioeconomic deprivation and greenness, in relation to type 2 diabetes onset in all communities; and 3) evaluate associations of the two community features after stratification by community type. Analysis controlled for key individual-level confounding variables and accounted for spatial clustering of persons within communities. Statistical analysis was completed using Stata-MP version 15.1 (StataCorp LLC, College Station, TX).

Logistic regression was used to estimate associations (odds ratios, 95% confidence intervals) using generalized estimating equations with robust standard errors and an exchangeable correlation structure within administrative community types. We adjusted for age (years; linear, quadratic, and cubic terms to allow for non-linearity), sex, race (white vs. all other races), ethnicity (Hispanic vs. non-Hispanic), and percent of time using Medical Assistance (surrogate for family socioeconomic status [≥ 50% vs. < 50%]) [31]. We did not include body mass index (BMI, kg/m^2) in models because this is likely a mediator of community associations (inclusion would attenuate or eliminate associations of interest). Models were first evaluated using all persons in all communities. We analyzed associations of the four definitions of community, community socioeconomic deprivation (quartiles; 4th quartile [worst deprivation] reference group), and greenness (tertiles) with diabetes status. Due to concerns about non-overlapping distributions resulting in extrapolation rather than adjustment (i.e., non-positivity [32]), we then stratified the community features models by community type.

In sensitivity analyses, to evaluate whether access to care – and thus higher likelihood of diabetes diagnosis – may have accounted for associations between community and diabetes, we examined the number of prior outpatient encounters (linear and quadratic terms) for study individuals by administrative community type and Medical Assistance status and added this variable to regression models.

RESULTS

Description of Study Population and Communities

Individuals were predominantly white and non-Hispanic; the majority had a primary care provider; and most cases were diagnosed with diabetes in an outpatient setting (**Table 1**). Individuals resided in 291 boroughs, 146 city census tracts, and 633 townships (**Online Supplement Table S2**). Over 40% of persons resided in rural areas (**Table 1**). Most borough residents were divided between urbanized areas and urban clusters. Approximately two-thirds of persons in townships resided in rural areas. A similar proportion of individuals in city census tracts resided in urbanized areas. On average, townships had higher greenness and lower community socioeconomic deprivation compared to boroughs and city census tracts (**Online Supplement Table S2**). Average racial and ethnic diversity and use of Medical Assistance for health insurance were highest in city census tracts. The mean total number of encounters with the health system before diabetes onset or the control selection date was high for all individuals, in all community types, regardless of Medical Assistance status (**Online Supplement Table S3**). Laboratory data confirmed that the categorization of diabetes cases and controls was valid (**Online Supplement Table S4**).

Associations of Communities with Type 2 Diabetes Onset

In the base model, controlling for age and sex, non-white race (vs. white), Hispanic ethnicity (vs. non-Hispanic), and Medical Assistance status were each associated with increased odds of type 2 diabetes onset. These associations did not substantively change as the community type and community features were added to the model. Odds ratios for non-white race (vs. white) ranged from 1.36 to 1.41, for Hispanic ethnicity (vs. non-Hispanic) from 1.46 to 1.52, and for Medical Assistance (≥ 50% of time vs. < 50%) from 1.71 to 1.74, with all confidence intervals excluding 1.0. Next, when administrative community type was added (townships as reference group), residing in boroughs and city census tracts was associated with significantly higher odds (**Table 2, Model 1**). Second, urban/rural status was added to the base model and residing in urbanized areas (vs. rural areas) had increased odds of diabetes onset (**Table 2, Model 2**). Third, the combined definition was added to the base model, and some categories (e.g., city census tracts in major urban and smaller urban areas highest, boroughs in these areas intermediate, vs. townships in rural areas as reference) were associated with increased odds of new onset diabetes (**Table 2, Model 3**). Finally, county was added to the base model, and seven counties were associated with reduced odds and two with increased odds of diabetes (**Table 2, Model 4**). We next evaluated community socioeconomic deprivation and greenness. When these community features were added to the base model, lower deprivation (**Table 2, Model 5**) and higher greenness (**Table 2, Model 6**) were associated with reduced odds of diabetes.

Models were next stratified by community type (only results for administrative community type shown). Race/ethnicity and Medical Assistance status were still associated with type 2 diabetes onset in the stratified models in all administrative community types (**Online Supplement Table S5**). Associations of community socioeconomic deprivation with diabetes evidenced decreasing odds ratios across decreasing deprivation quartiles in all community types, but only crossed an inferential threshold in city census tracts, with approximately 25% lower odds in the 1st vs. 4th quartile. Higher greenness was associated with reduced odds of diabetes in all community types.

Even after stratification by administrative community type and adjustment for community socioeconomic deprivation, several counties were independently associated with increased or reduced odds of diabetes onset (**Online Supplement Table S6**). The number of significant associations (n = 18, nine each with reduced or increased odds) was somewhat larger than that expected due to chance (108 statistical tests performed), with most associations observed for residing in boroughs. In these models, associations with community socioeconomic deprivation were present in the 1st quartile (vs. 4th) in townships and boroughs and in all quartiles in city census tracts. In all community types, higher greenness was associated with lower odds of diabetes.

Sensitivity Analyses

Addition of total outpatient encounters before diagnosis/control selection date did not substantively change associations in non-stratified or stratified models (results not shown). Community socioeconomic deprivation and greenness were evaluated together in models in boroughs and townships. In boroughs, associations of greenness with type 2 diabetes onset were attenuated by 1-2% and associations with community socioeconomic deprivation were no longer present. In townships, there was no substantive change in associations or inferences for greenness and associations with community socioeconomic deprivation were no longer present. These variables could not be evaluated together in city census tracts due to insufficient overlap in distributions.

DISCUSSION

There is great interest in understanding geographic disparities in type 2 diabetes risk. If the primary causes of these differences were community-level factors, community-level interventions could have large impacts on diabetes risk. A strong theoretical basis, and growing empirical evidence, indicates that community features contribute to diabetes risk directly or through increased risk of obesity, such as social, built, and natural environments contributing to impacts on physical activity and stress [33-35]. The primary goal of this study was to evaluate geographic disparities in type 2 diabetes by evaluating four definitions of community across the full range from rural to urban. We then evaluated associations of community socioeconomic deprivation and greenness overall and in models stratified by community type, the latter greatly reducing the degree to which these associations could be confounded by other community features.

In the study region, the use of combined community type allowed us to carefully identify the location and scale of risk. Risk of new onset type 2 diabetes was highest in cities in smaller urban areas, followed by cities in major urban areas and boroughs in major and smaller urban areas. In addition, even after accounting for community type and features, county was independently associated with diabetes onset. While many prior studies have evaluated county differences in diabetes risk [4, 36-38], none have also simultaneously evaluated communities. Our associations suggest that the risk factors that undergird U.S. geographic differences in diabetes likely exist at multiple, nested spatial scales. Some of the county associations were of high magnitude (e.g., exceeded 1.5 for protection or risk). Finally, there were consistent associations of higher community socioeconomic deprivation and lower greenness with higher diabetes risk, the former primarily in city census tracts, where average deprivation levels were higher, and the latter in all communities. We do not believe that the apparent lower diabetes risk in rural areas was due to less likely diagnosis due to lower access to health care, since, on average, individuals in the study, regardless of Medical Assistance status and community type, had high contact with the health care system.

We found several strong and consistent associations of individual-level characteristics. Non-white race, Hispanic ethnicity, and Medical Assistance status (a surrogate for low family socioeconomic status) were consistently associated with 1.3 to

1.7-fold increased odds of type 2 diabetes onset. Overall, the findings suggest that sociodemographic factors (race/ethnicity and individual-level socioeconomic status), urbanicity, higher community socioeconomic deprivation, and lower greenness, all of which co-occur in our region, were strong risk factors for type 2 diabetes.

Our findings on elevated risk of type 2 diabetes onset in urban areas is inconsistent with national studies that have reported higher crude prevalence estimates of type 2 diabetes in rural areas [39]. However, a study of the Behavioral Risk Factor Surveillance System found that after adjusting for individual-level socioeconomic measures, prevalence was higher in urban areas [40]. Geospatial predictors of diabetes risk likely vary by community and region; prior studies have reported, for example, that nine county-level measures of socioeconomic, race/ethnicity, and built environmental features explained up to 94% of the variation in type 2 diabetes prevalence in the Midwest, but very little variation in Pennsylvania [36].

The associations of greenness with diabetes were consistent with prior studies, but our results are the first to demonstrate robust findings across all types of communities while additionally controlling for county. The measurement of community features across community types may result in measures with different interpretations in different communities and regions; for example, agricultural, coniferous forest, and deciduous forest greenness are not evenly distributed and have different impacts on health [22].

Most prior studies of geographic disparities in diabetes have been cross-sectional, at the ecologic level, relying on self-reported diabetes, and focused on prevalent diabetes by county (too large and heterogeneous) or census tract (not experientially and behaviorally relevant). The current study avoided all these limitations. In addition, while many public health services are delivered at the county level, many potential interventions to address diabetes would need to be implemented at smaller scales and would not have county-wide impacts.

The study had some limitations. Although we adjusted for Medical Assistance health insurance as a surrogate for family socioeconomic status, there could still be residual confounding by individual-level income [31]. We did not measure behavioral mediators of the community definitions and features, such as physical activity or dietary intake. We could not account for residential selection bias, in which associations are due to reverse causation (if persons with individual-level risk factors for diabetes are more likely to reside in certain areas, by choice or opportunity). This can be a concern in studies of this type; social processes determine residence, so it can be difficult to distinguish individual-level characteristics from features of communities [41]. The residential stability and general population representativeness of our study population may mitigate these concerns. Although we used four definitions of community, all used administrative boundaries and thus may not represent how residents view the communities in which they reside and could still present edge and boundary effects and the modifiable areal unit problem [42-44].

The study had several strengths. Diabetes was objectively documented and verified with extensive biomarker and medical data. Temporality was appropriate for all independent variables. Study participants resided in a range of communities from urban to rural. We studied several approaches to community characterization at more relevant contextual scales than many prior studies and showed that smaller community contexts were associated with diabetes onset. Stratifying by community types limited bias from non-positivity [32].

The study findings provide important clues for the location (i.e., urban) and geographic scale (i.e., as localized as a square mile, the average area of boroughs and city census tracts) that identifies geospatial disparities in type 2 diabetes in Pennsylvania. We speculate that, since risk was higher in urban areas, our findings may suggest a smaller role for the positive features of the food and physical activity environments present in these areas (e.g., greater access to grocery stores, more walkable neighborhoods, more commercial physical activity opportunity establishments) and a larger role for individual and community demographic and socioeconomic factors found in the same areas.

Author contributions

Manuscript authors contributed in the following ways: conception of work: BSS, MNP, KRS, CIM, GI, AGH; obtained funding: BSS, AGH; study design: BSS, JSP, KBR, AGH; data management and analysis: JSP, KBR, BSS, MNP, JD, KAM, AGH; results interpretation: BSS, MNP, KBR, JD, KAM, KRS, CIM, GI, AGH; initial manuscript writing: BSS, MNP, KAM, AGH; critical revision of manuscript, final approval, and accountable for their work: BSS, JSP, MNP, KBR, KAM, JD, KRS, CIM, GI, AGH.

Competing interests

All authors declared that they have no competing interests.

Funding

This publication was made possible by Cooperative Agreement Number DP006293 funded by the U.S. Centers for Disease Control and Prevention, Division of Diabetes Translation.

Data sharing

De-identified electronic health record data are available upon written request with IRB approval and a data use agreement. All community data are publicly available.

References

1. Centers for Disease Control and Prevention, *National Diabetes Statistics Report 2020: Estimtaes of Diabetes and Its Burden in the United States.*, U.S. Department of Health and Human Services, Editor. 2020, Centers for Disease Control and Prevention: Atlanta, GA.
2. Garcia, M.C., et al., *Reducing Potentially Excess Deaths from the Five Leading Causes of Death in the Rural United States.* MMWR Surveill Summ, 2017. **66**(2): p. 1-7.
3. Ford, E.S., et al., *Geographic variation in the prevalence of obesity, diabetes, and obesity-related behaviors.* Obes Res, 2005. **13**(1): p. 118-22.
4. Cunningham, S.A., et al., *County-level contextual factors associated with diabetes incidence in the United States.* Ann Epidemiol, 2018. **28**(1): p. 20-25 e2.
5. Rural Health Information Hub. *Why Diabetes is a Concern for Rural Communities* 2020 April 13, 2020]; Available from: https://www.ruralhealthinfo.org/toolkits/diabetes/1/rural-concerns.
6. Singh, G.K. and M. Siahpush, *Widening rural-urban disparities in all-cause mortality and mortality from major causes of death in the USA, 1969-2009.* J Urban Health, 2014. **91**(2): p. 272-92.
7. James, C.V., et al., *Racial/Ethnic Health Disparities Among Rural Adults - United States, 2012-2015.* MMWR Surveill Summ, 2017. **66**(23): p. 1-9.
8. Cosby, A.G., et al., *Growth and Persistence of Place-Based Mortality in the United States: The Rural Mortality Penalty.* Am J Public Health, 2019. **109**(1): p. 155-162.
9. Henning-Smith, C.E., et al., *Rural Counties With Majority Black Or Indigenous Populations Suffer The Highest Rates Of Premature Death In The US.* Health Aff (Millwood), 2019. **38**(12): p. 2019-2026.
10. Maier, W., et al., *Area level deprivation is an independent determinant of prevalent type 2 diabetes and obesity at the national level in Germany. Results from the National Telephone Health Interview Surveys 'German Health Update' GEDA 2009 and 2010.* PLoS One, 2014. **9**(2): p. e89661.
11. Muller, G., et al., *Regional and neighborhood disparities in the odds of type 2 diabetes: results from 5 population-based studies in Germany (DIAB-CORE consortium).* Am J Epidemiol, 2013. **178**(2): p. 221-30.
12. Jagai, J.S., et al., *Association between environmental quality and diabetes in the USA.* J Diabetes Investig, 2019.
13. Honold, J., et al., *Multiple environmental burdens and neighborhood-related health of city residents.* Journal of Environmental Psychology, 2012. **32**(4): p. 305-317.
14. Mavoa, S., et al., *How Do Neighbourhood Definitions Influence the Associations between Built Environment and Physical Activity?* Int J Environ Res Public Health, 2019. **16**(9).
15. Dekker, L.H., R.H. Rijnks, and G.J. Navis, *Regional variation in type 2 diabetes: evidence from 137 820 adults on the role of neighbourhood body mass index.* Eur J Public Health, 2020. **30**(1): p. 189-194.
16. Stafford, M., O. Duke-Williams, and N. Shelton, *Small area inequalities in health: are we underestimating them?* Soc Sci Med, 2008. **67**(6): p. 891-9.
17. Tuson, M., et al., *Incorporating geography into a new generalized theoretical and statistical framework addressing the modifiable areal unit problem.* Int J Health Geogr, 2019. **18**(1): p. 6.
18. Bennett, K.J., et al., *What Is Rural? Challenges And Implications Of Definitions That Inadequately Encompass Rural People And Places.* Health Aff (Millwood), 2019. **38**(12): p. 1985-1992.
19. Census Bureau. *Geography program: 2010 census urban and rural classification and urban area criteria.* . 2018 [cited 2020 January 5, 2020]; Available from:

https://www.census.gov/programssurveys/geography/guidance/geoareas/urban-rural/2010-urbanrural.html.

20. Hirsch, A.G., et al., *Associations of Four Community Factors With Longitudinal Change in Hemoglobin A1c Levels in Patients With Type 2 Diabetes.* Diabetes Care, 2018. **41**(3): p. 461-468.

21. Cohen, S.A., et al., *A Closer Look at Rural-Urban Health Disparities: Associations Between Obesity and Rurality Vary by Geospatial and Sociodemographic Factors.* J Rural Health, 2017. **33**(2): p. 167-179.

22. James, W.L., *All rural places are not created equal: revisiting the rural mortality penalty in the United States.* Am J Public Health, 2014. **104**(11): p. 2122-9.

23. Astell-Burt, T., X. Feng, and G.S. Kolt, *Is neighborhood green space associated with a lower risk of type 2 diabetes? Evidence from 267,072 Australians.* Diabetes Care, 2014. **37**(1): p. 197-201.

24. Muller, G., et al., *Inner-city green space and its association with body mass index and prevalent type 2 diabetes: a cross-sectional study in an urban German city.* BMJ Open, 2018. **8**(1): p. e019062.

25. Casey, J.A., et al., *Unconventional Natural Gas Development and Birth Outcomes in Pennsylvania, USA.* Epidemiology, 2016. **27**(2): p. 163-72.

26. Lawrence, J.M., et al., *Validation of pediatric diabetes case identification approaches for diagnosed cases by using information in the electronic health records of a large integrated managed health care organization.* Am J Epidemiol, 2014. **179**(1): p. 27-38.

27. Zhong, V.W., et al., *Use of administrative and electronic health record data for development of automated algorithms for childhood diabetes case ascertainment and type classification: the SEARCH for Diabetes in Youth Study.* Pediatr Diabetes, 2014. **15**(8): p. 573-84.

28. Schwartz, B.S., et al., *Body mass index and the built and social environments in children and adolescents using electronic health records.* Am J Prev Med, 2011. **41**(4): p. e17-28.

29. Casey, J.A., et al., *Greenness and Birth Outcomes in a Range of Pennsylvania Communities.* Int J Environ Res Public Health, 2016. **13**(3).

30. Nau, C., et al., *Community socioeconomic deprivation and obesity trajectories in children using electronic health records.* Obesity (Silver Spring), 2015. **23**(1): p. 207-12.

31. Casey, J.A., et al., *Measures of SES for Electronic Health Record-based Research.* Am J Prev Med, 2018. **54**(3): p. 430-439.

32. Petersen, M.L., et al., *Diagnosing and responding to violations in the positivity assumption.* Stat Methods Med Res, 2012. **21**(1): p. 31-54.

33. Cox, M., et al., *Locality deprivation and Type 2 diabetes incidence: a local test of relative inequalities.* Soc Sci Med, 2007. **65**(9): p. 1953-64.

34. Maier, W., et al., *The impact of regional deprivation and individual socio-economic status on the prevalence of Type 2 diabetes in Germany. A pooled analysis of five population-based studies.* Diabet Med, 2013. **30**(3): p. e78-86.

35. James, P., et al., *A Review of the Health Benefits of Greenness.* Curr Epidemiol Rep, 2015. **2**(2): p. 131-142.

36. Hipp, J.A. and N. Chalise, *Spatial analysis and correlates of county-level diabetes prevalence, 2009-2010.* Prev Chronic Dis, 2015. **12**: p. E08.

37. Geiss, L.S., et al., *Changes in diagnosed diabetes, obesity, and physical inactivity prevalence in US counties, 2004-2012.* PLoS One, 2017. **12**(3): p. e0173428.

38. Liese, A.D., et al., *Evaluating geographic variation in type 1 and type 2 diabetes mellitus incidence in youth in four US regions.* Health Place, 2010. **16**(3): p. 547-56.

39. National Center for Chronic Disease Prevention and Health Promotion. *Division of Diabetes Translation At A Glance.* 2019 January 20, 2020]; Available from: https://www.cdc.gov/chronicdisease/resources/publications/aag/diabetes.htm.

40. O'Connor, A. and G. Wellenius, *Rural-urban disparities in the prevalence of diabetes and coronary heart disease.* Public Health, 2012. **126**(10): p. 813-20.

41. Macintyre, S. and A. Ellaway, *Ecological approaches: rediscovering the role of the physical and social environment.*, in *Social Epidemiology.*, I. Kawachi and L. Berkman, Editors. 2000, Oxford University Press: New York. p. 332-348.

42. Openshaw, S., *The Modifiable Areal Unit Problem.* 1984, Norwich, CT.: GeoBooks.

43. Wong, D., *The modifiable areal unit problem (MAUP).* , in *The SAGE Handbook of Spatial Analysis.* , A.S. Fotheringham and P.A. Rogerson, Editors. 2009, SAGE Publications: London. p. 105-124.

44. Sadler, R.C., J.A. Gilliland, and G. Arku, *An application of the edge effect in measuring accessibility to multiple food retailer types in southwestern Ontario, Canada.* Int J Health Geogr, 2011. **10**: p. 34.

Table 1. Selected characteristics of individuals with diabetes and controls, frequency-matched to cases (5:1) on age, sex, and year of diagnosis or control selection date.

Variable	Cases	Controls	p-value*
Unique persons	15,888	65,084	NA
Number	15,888	79,435	NA
Sex, female, n (COL %)	7798 (49.1)	38,988 (49.1)	matched
Age at diagnosis or control selection date, years, mean (SD)	54.9 (15.1)	54.9 (15.3)	matched
Age, years, categories, n (COL %)			matched
10 to < 20 years	304 (1.9)	1520 (1.9)	
20 to < 30 years	628 (4.0)	3140 (4.0)	
30 to < 40 years	1611 (10.1)	8055 (10.1)	
40 to < 50 years	3086 (19.4)	15,429 (19.4)	
50 to < 60 years	4286 (27.0)	21,428 (27.0)	
60 to < 70 years	3510 (22.1)	17,548 (22.1)	
70 to < 80 years	1737 (10.9)	8685 (10.9)	
80 to < 90 years	645 (4.1)	3225 (4.1)	
≥ 90 years	81 (0.5)	405 (0.5)	
Race, white, n (COL %)	15,429 (97.1)	77,867 (98.0)	< 0.001
Hispanic ethnicity, n (COL %)	369 (2.3)	1094 (1.4)	< 0.001
Primary care provider†, yes, n (%)	11,884 (74.8)	61,042 (76.9)	< 0.001
Year of diagnosis/encounter, n (COL %)			matched
2008	1761 (11.1)	8805 (11.1)	
2009	2019 (12.7)	10,095 (12.7)	
2010	1747 (11.0)	8735 (11.0)	
2011	1675 (10.5)	8373 (10.5)	
2012	1716 (10.8)	8579 (10.8)	
2013	1842 (11.6)	9209 (11.6)	
2014	1844 (11.6)	9220 (11.6)	
2015	1734 (10.9)	8669 (10.9)	
2016	1550 (9.8)	7750 (9.8)	
Setting of diagnosis/encounter, n (COL %)			< 0.001
Outpatient	12,068 (76.0)	73,998 (93.2)	
Medication order	1632 (10.3)	0 (0.0)	
Urgent care	165 (1.0)	2116 (2.7)	
Emergency department	1526 (9.6)	3068 (3.9)	
Inpatient	498 (3.1)	252 (0.3)	
Outpatient encounters in year before diagnosis or control selection date, mean (SD)	4.4 (5.1)	3.5 (4.1)	< 0.001
Outpatient encounters, total before diagnosis or control selection date, mean (SD)	35.9 (34.8)	35.2 (32.5)	0.01
Medical Assistance, % of time receiving, n (COL %)			< 0.001
< 50%	14,921 (93.9)	76,705 (83.7)	
≥ 50%	967 (6.1)	2730 (3.4)	
Outpatient encounters before diagnosis/encounter, mean (SD), by % of time receiving Medical Assistance			< 0.001

Variable	Cases	Controls	p-value*
0%	35.5 (34.1)	34.9 (32.1)	
0.1-24.9%	45.2 (40.7)	42.8 (38.3)	
25.0-74.9%	33.9 (35.8)	35.2 (33.6)	
75+%	29.1 (26.9)	27.7 (26.0)	
Duration from first contact with health system to diagnosis/control selection date, years, n (%)			0.72
Quartile 1 (2 to < 5 years)	1860 (11.7)	9466 (11.9)	
Quartile 2 (5 to < 8 years)	2571 (16.2)	12,646 (15.9)	
Quartile 3 (8 to < 12 years)	4700 (29.6)	23,665 (29.8)	
Quartile 4 (≥ 12 years)	6757 (42.5)	33,658 (42.4)	
Community socioeconomic deprivation, n (COL %)‡			< 0.001
Quartile 1	3001 (18.9)	17,329 (21.8)	
Quartile 2	4300 (27.1)	23,172 (29.2)	
Quartile 3	4217 (26.5)	20.328 (25.6)	
Quartile 4	4370 (27.5)	18,606 (23.4)	
Greenness, peak NDVI, in buffer, n (COL %) §			< 0.001
Tertile 1	5894 (37.1)	25,894 (32.6)	
Tertile 2	5023 (31.6)	26.751 (33.7)	
Tertile 3	4971 (31.3)	26,790 (33.7)	
Administrative community type of residence, n (COL %)			< 0.001
Borough	4621 (29.1)	21,756 (27.4)	
Census tract in city	1806 (11.4)	6548 (8.2)	
Township	9461 (59.6)	51,131 (64.4)	
Setting of residence, n (COL %)			< 0.001
Rural	6513 (41.0)	34,984 (44.0)	
Urbanized area	4906 (30.9)	23,423 (29.5)	
Urban cluster	4469 (28.1)	21,028 (26.5)	

Abbreviations: COL = column; NDVI = normalized difference vegetation index; SD = standard deviation.

* Because controls could be in these comparisons more than once, methods were used for significance testing that accounted for this, including inverse-probability weighted regression for time-invariant characteristics, mixed-effect regression for time-varying continuous (linear), binary (logistic), and count (Poisson) characteristics, and multinomial logistic regression with robust standard errors for polytomous time-varying characteristics. In the weighted analyses, weights were the number of appearances in the analysis (implemented with a dataset having only one record per person).

† According to Geisinger's primary care provider lists.

‡ Quartile cutoffs were defined within the three time periods; the range of values for Q1, Q2, Q3, and Q4 were -18.33 to -1.96; -1.99 to -0.015; 0.005 to 2.05; and 2.11 to 12.4.

§ The range of values in T1, T2, and T3 were 0.07 to 0.627, 0.63 to 0.756, and 0.76 to 0.94, respectively.

Table 2. Adjusted* associations of community and community feature variables **from separate models** with new onset type 2 diabetes status.

Variable	OR (95% CI)
Community types	
Model 1: Administrative community type	
Township	1.0
Borough	1.10 (1.04, 1.16)
City census tract	1.34 (1.25, 1.44)
Model 2: Residential location, urban/rural	
Rural	1.0
Urbanized area	1.14 (1.08, 1.21)
Urban cluster	1.04 (0.98, 1.11)
Model 3: Combined location†	
Township / rural	1.0
Township / urban cluster	1.00 (0.92, 1.08)
Township / urbanized area	1.06 (0.98, 1.16)
Borough + city census tract / rural	1.04 (0.95, 1.15)
Borough / urban cluster	1.09 (1.01, 1.18)
Borough / urbanized area	1.15 (1.06, 1.25)
City census tract / urban cluster	1.41 (1.22, 1.62)
City census tract / urbanized area	1.33 (1.22, 1.45)
Model 4: County ‡	
Luzerne	1.0
Blair	0.73 (0.57, 0.95)
Centre	0.84 (0.75, 0.94)
Juniata	1.19 (1.00, 1.40)
Lackawanna	1.19 (1.07, 1.31)
Lebanon	0.39 (0.16, 0.93)
Monroe	0.78 (0.69, 0.88)
Schuylkill	0.85 (0.78, 0.92)
Sullivan	0.60 (0.45, 0.81)
Union	0.77 (0.64, 0.93)
Community features, all communities combined	
Model 5: community socioeconomic deprivation, quartiles §	
1	0.82 (0.76, 0.88)
2	0.87 (0.81, 0.93)
3	0.89 (0.83, 0.96)
4	1.0
Model 6: greenness (normalized difference vegetation index) ||	
1	1.0
2	0.88 (0.85, 0.93)
3	0.84 (0.80, 0.88)

* Logistic regression models using generalized estimating equations with robust standard errors; one community or community feature variable was in the model at a time; models adjusted for sex, race (white vs. non-white), ethnicity (Hispanic vs. non-Hispanic), age (age, age^2, age^3), and Medical Assistance status.

† This is a combination of administrative community type and residential location (urban/rural); the few persons in city census tract / rural were combined with borough / rural.

‡ Only counties with confidence interval excluding 1.0 are shown in table. Luzerne County was selected as the reference group because it is the most populous county in the study region.

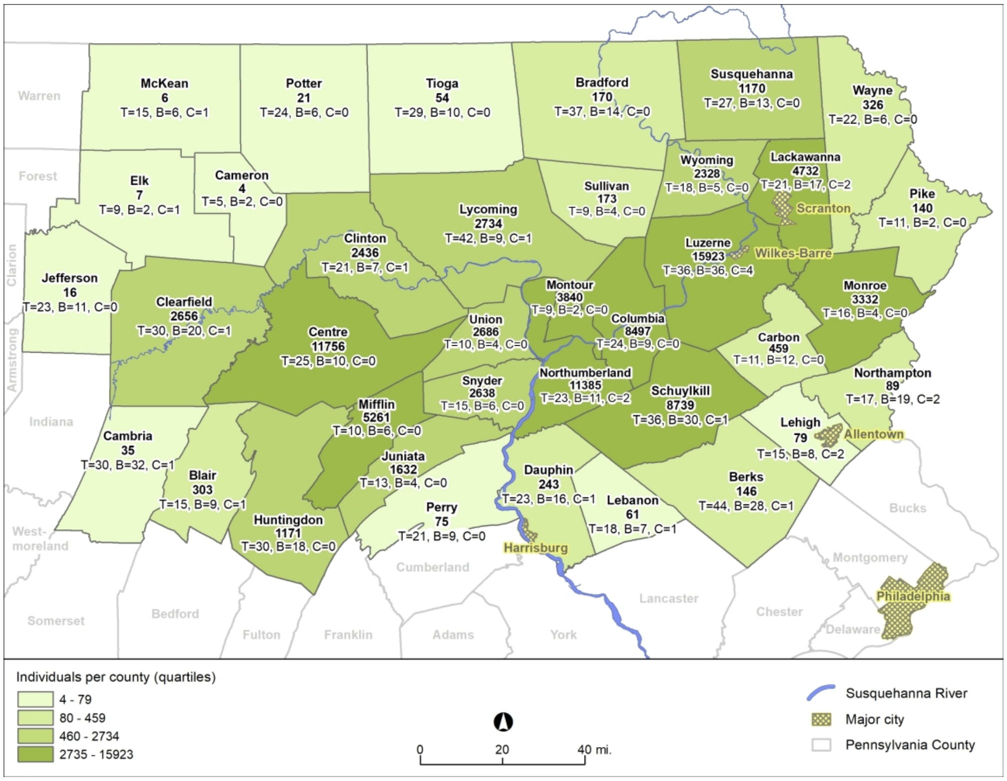
§ Quartile cutoffs were defined within the three time periods; the range of values for persons in Q1, Q2, Q3, and Q4 were -25.06 to -1.82; -1.99 to 0.10; 0.005 to 2.05; and 1.89 to 12.4, respectively.

|| The range of values in T1, T2, and T3 were 0.07 to 0.627, 0.63 to 0.756, and 0.76 to 0.94, respectively.

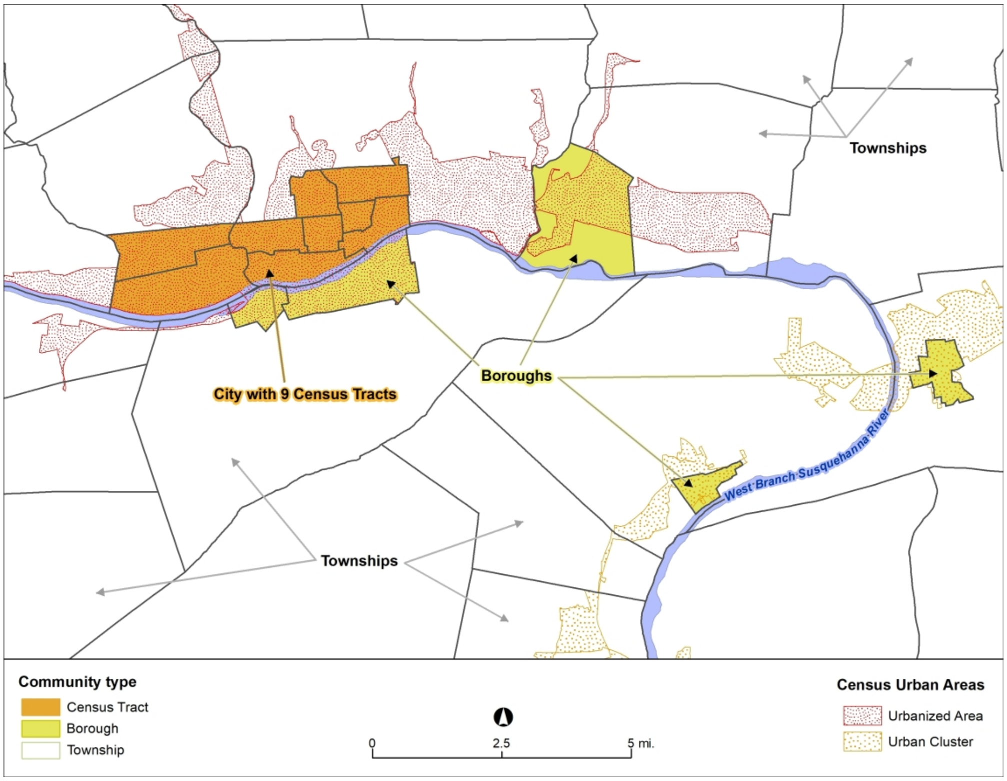
Figure Captions

Figure 1. Distribution of study individuals and administrative community types by county in study region. The bolded number is the number of individuals; T, B, and C identify the number of townships, boroughs, and city census tracts within each county that were included in the analysis.

Figure 2. Areas along the Susquehanna River in Lycoming County, Pennsylvania from Williamsport (city) and South Williamsport (borough) to Montoursville (borough), Muncy (borough), and Montgomery (borough), showing relations between administrative community types (townships, boroughs, and city census tracts) and urbanized areas, urban clusters, and rural areas. Both sets of these administrative boundaries were used in the analysis.


Warren
Forest
Clarion
Armstrong
Indiana
West-moreland
Somerset
Bedford
Fulton
Franklin
Adams
York
Cumberland
Lancaster
Chester
Delaware
Montgomery
Bucks
McKean 6 T=15, B=6, C=1
Potter 21 T=24, B=6, C=0
Tioga 54 T=29, B=10, C=0
Bradford 170 T=37, B=14, C=0
Susquehanna 1170 T=27, B=13, C=0
Wayne 326 T=22, B=6, C=0
Elk 7 T=9, B=2, C=1
Cameron 4 T=5, B=2, C=0
Lycoming 2734 T=42, B=9, C=1
Sullivan 173 T=9, B=4, C=0
Wyoming 2328 T=18, B=5, C=0
Lackawanna 4732 T=21, B=17, C=2
Scranton
Pike 140 T=11, B=2, C=0
Clinton 2436 T=21, B=7, C=1
Luzerne 15923 T=36, B=36, C=4
Wilkes-Barre
Jefferson 16 T=23, B=11, C=0
Clearfield 2656 T=30, B=20, C=1
Centre 11756 T=25, B=10, C=0
Montour 3840 T=9, B=2, C=0
Union 2686 T=10, B=4, C=0
Columbia 8497 T=24, B=9, C=0
Monroe 3332 T=16, B=4, C=0
Carbon 459 T=11, B=12, C=0
Northampton 89 T=17, B=19, C=2
Snyder 2638 T=15, B=6, C=0
Northumberland 11385 T=23, B=11, C=2
Schuylkill 8739 T=36, B=30, C=1
Mifflin 5261 T=10, B=6, C=0
Cambria 35 T=30, B=32, C=1
Lehigh 79 T=15, B=8, C=2
Allentown
Juniata 1632 T=13, B=4, C=0
Blair 303 T=15, B=9, C=1
Dauphin 243 T=23, B=16, C=1
Berks 146 T=44, B=28, C=1
Perry 75 T=21, B=9, C=0
Lebanon 61 T=18, B=7, C=1
Huntingdon 1171 T=30, B=18, C=0
Harrisburg
Philadelphia
Individuals per county (quartiles)
4 - 79
80 - 459
460 - 2734
2735 - 15923
0 20 40 mi.
Susquehanna River
Major city
Pennsylvania County

279x215mm (300 x 300 DPI)



279x215mm (300 x 300 DPI)

Online Supplement

Table S1. Diabetes case finding using EHR data.

Must meet at least one of the following criteria:

1. At least two separate encounter dates (inpatient, outpatient, emergency department) with type 2 diabetes diagnosis codes (ICD-9, ICD-10, or electronic diagnosis group [EDG]).
   a. Excluded if had ≥ 10 years of type 1 diagnoses and < five years with type 2 diagnoses.
   b. Excluded if < 10 years of age at first diabetes diagnosis.
2. At least one diabetes medication order, other than metformin or acarbose if female. Metformin combination medications were included.
   a. Excluded if first diabetes medication order was prior to age 10 years.
3. At least one encounter with type 2 diabetes diagnosis and an abnormal laboratory value (random glucose ≥ 200 mg/dl; fasting glucose ≥ 126 mg/dl; or hemoglobin A1c ≥ 6.5%).
   a. Excluded if had ≥ 10 years of type 1 diagnoses and < five years with type 2 diagnoses.

- The date of onset was assigned as the earliest date with any evidence of diabetes (e.g., had generic diabetes diagnoses that were not used for definition #1, or had abnormal laboratory value that was not accompanied by a diagnosis so did not meet definition #3).

Notes:
a) To meet criteria #2 or #3, criterion had to occur > 9 months prior to or > 1 month after delivery of child (to avoid gestational diabetes). Gestational diabetes was not an exclusion if the individual subsequently developed type 2 diabetes. Date of onset was assigned as when the person met the type 2 diabetes criterion; and
b) EDG codes are used in Epic EHR software (Epic Systems Corporation, Verona, WI) and often have higher specificity and greater detail.
c) Of the 15,888 diabetes cases: 11,944 met criterion 1; 10,183 met criterion 2; 12,552 met criterion 3; 7008 met all three; and 4775 met at least two.
d) Because metformin can be used for pre-diabetes, we evaluated how many persons could have had this diagnosis instead of diabetes in our diabetes onset definition. Of the 1579 men who met only definition #2, between 544 (3.4%) and 1207 (7.6%) may have had pre-diabetes instead of diabetes, depending on how longitudinal information on diagnoses, medications, medication indications, and abnormal laboratory results were used and interpreted.

Table S2. Selected characteristics of study individuals and communities by administrative community type.

Variables	Borough	Census Tract	Township
By community type (n = 1070 communities)			
Number (%), total	291 (27.2)	146 (13.6)	633 (59.2)
Number (%), among cases	224 (27.6)	107 (13.2)	482 (59.3)
Number (%), among controls	278 (26.9)	137 (13.2)	620 (59.9)
Counties with at least one resident in community type, n	35	16	37
Counties with at least 20 residents in community type, n	27	9	32
Community measures, by community type (n = 1070 communities)			
Area, square miles, mean (SD)	1.72 (2.32)	1.20 (3.52)	29.4 (18.1)
Community socioeconomic deprivation, mean (SD)	-0.09 (2.99)	4.17 (3.80)	-1.15 (2.71)
Population density, persons per square mile, mean (SD)	2094.7 (1642.3)	6594.5 (5014.6)	157.5 (279.4)
Developed land, % (SD)	37.2 (22.6)	72.6 (23.0)	3.66 (7.35)
Intersection density per square mile, mean (SD)	120.6 (86.1)	208.5 (117.0)	13.34 (14.77)
By participant (n = 95,323 individuals)			
Cases, n (%) (total = 15,888)	4621 (29.1)	1806 (11.4)	9461 (59.5)
Controls, n (%) (total = 79,435)	21,756 (27.4)	6548 (8.2)	51,131 (64.4)
Age at diabetes onset or control selection date, years, mean (SD)	54.4 (15.9)	52.7 (16.1)	55.3 (14.8)
Sex, female, n (%)	13,329 (50.2)	4449 (53.3)	29,098 (48.0)
Race, white, n (%)	245,963 (98.4)	7873 (94.2)	59,460 (98.1)
Ethnicity, Hispanic, n (%)	353 (1.3)	430 (5.2)	680 (1.1)
Body mass index, kg/m^2, mean (SD)	30.6 (7.47)	30.9 (7.96)	30.3 (6.94)
Medical Assistance, % of time, mean (SD)	5.9 (17.9)	10.3 (23.2)	3.3 (13.5)
Medical Assistance, ever*, n (%)	3311 (12.6)	1692 (20.3)	4311 (7.1)
Contact with health system before diagnosis/control selection date, years, mean (SD)	12.7 (4.37)	12.1 (4.57)	12.9 (4.34)
Charlson index, mean (SD)	1.75 (1.83)	1.64 (1.78)	1.76 (1.78)
Greenness, peak NDVI, in buffer, mean (SD)	0.61 (0.11)	0.51 (0.10)	0.73 (0.10)
Urban status by UA and UC boundaries, n (col %)			
Rural	3031 (11.5)	10 (0.1)	38,456 (63.5)
Urbanized area (UA)	11,409 (43.3)	5414 (64.8)	11,506 (19.0)
Urban cluster (UC)	11,937 (45.3)	2930 (35.1)	10,630 (17.5)

Abbreviations: NDVI = normalized difference vegetation index; SD = standard deviation.
* At least one encounter that used Medical Assistance for health insurance.

Table S3. Mean outpatient encounters among cases and controls by community type and Medical Assistance status.

Variable	Cases, n = 15,888			Controls, n = 79,435		
	Boroughs n = 4621	City Census Tracts n = 1806	Townships n = 9461	Boroughs n = 21,756	City Census Tracts n = 6548	Townships n = 51,131
Outpatient encounters, total before diagnosis, mean (SD)	35.9 (34.8)	31.6 (32.1)	36.8 (35.2)	35.7 (33.8)	33.5 (32.8)	35.2 (31.8)
Outpatient encounters before diagnosis, mean (SD), by Medical Assistance status (% time receiving)						
0%	35.2 (33.9)	30.9 (31.2)	36.3 (34.6)	35.1 (33.1)	32.9 (32.1)	35.0 (31.7)
0.1-24.9%	47.7 (41.3)	41.8 (44.3)	44.7 (39.0)	44.2 (40.7)	40.0 (39.9)	42.6 (36.1)
25.0-74.9%	32.5 (34.5)	29.3 (25.4)	37.1 (40.2)	37.3 (36.8)	33.6 (32.4)	34.2 (31.4)
75+%	30.6 (28.9)	34.2 (21.0)	30.7 (28.0)	27.7 (28.5)	27.9 (28.4)	27.7 (22.6)
SD = standard deviation						

Medical Profile of Cases and Controls

To evaluate our categorization of diabetes cases and controls, we examined a number of biomarkers and other measures of relevance to diabetes, dysglycemia, and other cardio-metabolic risk factors development that were available in the EHR, including hemoglobin A1c (HbA1c), lipids (cholesterol and triglycerides), blood glucose (fasting and unspecified), and body mass index (BMI) (**Online Supplement Table S4**). Fasting blood glucose was measured in the year before the diabetes onset or control dates in 24% of cases and 29% of controls. Interestingly, the mean value was higher in the year before diagnosis in persons who would develop diabetes compared to those who would not, 108.5 vs. 95.8 mg/dL ($p < 0.001$). In the year after diagnosis or control dates, fasting blood glucose was available in 58% of cases and 30% of controls, and mean levels were much higher in cases compared to controls (147.9 vs. 95.9, $p < 0.001$). HbA1c, triglycerides, unspecified blood glucose, and BMI all evidenced similar patterns (**Online Supplement Table S4**). In the year before and after diagnosis, most cases and controls had BMI measured, with a much higher mean in cases compared to controls before and after diagnosis.

Table S4. Selected laboratory and other biometric values comparing new onset type 2 diabetes cases and controls without diabetes.

Variable	Cases	Controls
Number	15,888	79,435
Hemoglobin A1c (HbA1c)		
# in year before diagnosis or control selection date per person, number of persons (%) with		
0 values	13,618 (85.7)	75,731 (95.3)
1 value	1801 (11.3)	3257 (4.1)
2+ values	469 (3.0)	447 (0.6)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	2270 (14.3)	3704 (4.7)
HbA1c %, mean (SD)	5.9 (0.4)	5.6 (0.4)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	11,990 (75.5)	3839 (4.8)
HbA1c %, mean (SD)	7.5 (2.0)	5.6 (0.4)
LDL cholesterol		
# in year before diagnosis or index date per person, number of persons (%) with		
0 values	10,155 (63.9)	46,485 (58.5)
1 value	4068 (25.6)	23,737 (29.9)
2+ values	1665 (10.5)	9213 (11.6)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	5733 (36.1)	32,950 (41.5)
LDL-cholesterol, mg/dL, mean (SD)	107.2 (35.6)	109.6 (33.0)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	11,726 (73.8)	34,223 (43.1)
LDL-cholesterol, mg/dL, mean (SD)	108.5 (36.7)	111.1 (33.7)
Triglycerides		
# in year before diagnosis or index date per person, number of persons (%) with		
0 values	10,529 (66.3)	48,714 (61.3)
1 value	3869 (24.4)	22,585 (28.4)
2+ values	1490 (9.4)	8136 (10.2)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	5359 (33.7)	30,721 (38.7)
Triglycerides, mg/dL, mean (SD)	188.7 (131.7)	133.7 (81.2)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	11,207 (70.5)	31,663 (39.9)
Triglycerides, mg/dL, mean (SD)	216.5 (244.8)	135.0 (86.8)
Glucose, fasting		
# in year before diagnosis or index date per person, # of persons (%) with		
0 values	12,139 (76.4)	56,198 (70.8)
1 value	2968 (18.7)	19,023 (24.0)
2+ values	781 (5.0)	4214 (5.3)

Variable	Cases	Controls
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	3749 (23.6)	23,237 (29.3)
Glucose, mg/dL, mean (SD)	108.5 (11.8)	95.8 (9.3)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	9259 (58.3)	24,105 (30.3)
Glucose, mg/dL, mean (SD)	147.9 (60.9)	95.9 (9.3)
Glucose, unspecified		
# in year before diagnosis or index date per person, # persons (%) with		
0 values	9913 (62.4)	54,258 (68.3)
1 value	3115 (19.6)	15,293 (19.3)
2+ values	2860 (18.0)	9884 (12.4)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	5975 (37.6)	25,177 (31.7)
Glucose, mg/dL, mean (SD)	124.6 (28.2)	97.7 (15.5)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	10,833 (68.2)	27,779 (35.0)
Glucose, mg/dL, mean (SD)	170.7 (95.2)	98.4 (16.5)
Body mass index (BMI)		
# in year before diagnosis or index date per person, mean (SD)	3.1 (4.1)	2.4 (3.2)
Closest value in year prior to diagnosis or index date		
Persons with value, n (%)	11,237 (70.7)	54,733 (68.9)
BMI, kg/m^2, mean (SD)	36.2 (8.4)	29.3 (6.4)
Closest value in year after diagnosis or index date		
Persons with value, n (%)	13,957 (87.9)	65,084 (81.9)
BMI, kg/m^2, mean (SD)	36.0 (8.4)	29.3 (6.4)

Table S5. Adjusted* associations of selected independent variables with type 2 diabetes status stratified by administrative community type.

	Stratified by Administrative Community Type			**Stratified by Administrative Community Type**		
	Boroughs	**City Census Tracts**	**Townships**	**Boroughs**	**City Census Tracts**	**Townships**
	Model 1a	**Model 1b**	**Model 1c**	**Model 2a**	**Model 2b**	**Model 2c**
Variable	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**
Race						
White	1.0	1.0	1.0	1.0	1.0	1.0
All others	1.44 (1.12, 1.94)	1.30 (1.05, 1.60)	1.36 (1.14, 1.61)	1.43 (1.12, 1.84)	1.28 (1.04, 1.58)	1.35 (1.14, 1.61)
Ethnicity						
Non-Hispanic	1.0	1.0	1.0	1.0	1.0	1.0
Hispanic	1.50 (1.16, 1.94)	1.33 (1.02, 1.72)	1.52 (1.16, 1.97)	1.50 (1.16, 1.94)	1.32 (1.02, 1.71)	1.52 (1.17. 1.97)
Medical Assistance						
< 50% of time	1.0	1.0	1.0	1.0	1.0	1.0
50+% of time	1.66 (1.47, 1.86)	1.46 (1.26, 1.70)	1.83 (1.61, 2.09)	1.66 (1.48, 1.86)	1.48 (1.27, 1.72)	1.83 (1.61, 2.09)
CSD **						
Q1	0.88 (0.77, 1.01)	0.75 (0.56, 1.00)	0.93 (0.84, 1.02)			
Q2	0.96 (0.84, 1.08)	0.77 (0.63, 0.94)	0.97 (0.89, 1.06)			
Q3	0.98 (0.87, 1.10)	0.78 (0.67, 0.91)	0.98 (0.89, 1.07)			
Q4	1.0	1.0	1.0			
NDVI, 1250x1250m †						
T1				1.0	1.0	1.0
T2				0.93 (0.87, 0.99)	0.76 (0.64, 0.90)	0.93 (0.87, 0.99)
T3				0.85 (0.76, 0.96)	0.76 (0.50, 1.17)	0.90 (0.84, 0.96)

Abbreviations: CSD = community socioeconomic deprivation; NDVI = normalized difference vegetation index;

* Logistic regression models using generalized estimating equations with robust standard errors; also adjusted for sex and age (age, age^2, age^3).

** Quartile cutoffs were defined within the three time periods; the range of values for persons in Q1, Q2, Q3, and Q4 were -25.06 to -1.82; -1.99 to 0.10; 0.005 to 2.05; and 1.89 to 12.4, respectively.

† The range of values in T1, T2, and T3 were 0.07 to 0.627, 0.63 to 0.756, and 0.76 to 0.94, respectively.

Table S6. Adjusted* associations of selected independent variables with type 2 diabetes status stratified by administrative community type with county and community socioeconomic deprivation ***OR*** greenness.

	Stratified by Administrative Community Type		
	Boroughs	**City Census Tracts**	**Townships**
	Model 1	**Model 1**	**Model 1**
Variable	**OR (95% CI)**	**OR (95% CI)**	**OR (95% CI)**
Model 1 – with county and community socioeconomic deprivation (CSD)			
Race			
White	1.0	1.0	1.0
All others	1.45 (1.13, 1.86)	1.31 (1.06, 1.62)	1.39 (1.16, 1.66)
Ethnicity			
Non-Hispanic	1.0	1.0	1.0
Hispanic	1.49 (1.15, 1.92)	1.32 (1.02, 1.71)	1.55 (1.18, 2.04)
Medical Assistance			
< 50% of time	1.0	1.0	1.0
50+% of time	1.66 (1.47, 1.87)	1.48 (1.28, 1.72)	1.85 (1.62, 2.11)
Community socioeconomic deprivation, quartiles			
Q1	0.87 (0.76, 0.996)	0.71 (0.52, 0.95)	0.91 (0.82, 0.99)
Q2	0.93 (0.83, 1.06)	0.78 (0.65, 0.95)	0.96 (0.88, 1.05)
Q3	0.97 (0.87, 1.09)	0.79 (0.67, 0.93)	0.98 (0.90, 1.07)
Q4	1.0	1.0	1.0
County			
Luzerne	1.0	1.0	1.0
Blair	**0.64** (0.51, 0.81)	0.62 (0.23, 1.64)	0.86 (0.61, 1.21)
Clearfield	1.00 (0.82, 1.24)	**0.76** (0.66, 0.87)	0.97 (0.82, 1.15)
Dauphin	0.90 (0.56, 1.45)	**2.81** (1.47, 5.37)	1.43 (0.96, 2.15)
Juniata	**1.68** (1.22, 2.31)	NA†	1.18 (0.99, 1.41)
Lackawanna	1.12 (0.96, 1.37)	**1.23** (1.06, 1.43)	1.13 (0.93, 1.38)
Lehigh	**18.2** (2.00, 165.1)	2.00 (0.85, 4.68)	0.66 (0.26, 1.65)
Mifflin	**1.20** (1.00, 1.43)	NA	1.06 (0.93, 1.21)
Monroe	**0.73** (0.59, 0.91)	NA	**0.85** (0.74, 0.98)
Perry	**3.16** (1.34, 7.47)	NA	0.96 (0.51, 1.83)
Potter	**4.90** (4.42, 5.43)	NA	0.71 (0.15, 3.31)
Schuylkill	0.91 (0.80, 1.02)	0.93 (0.80, 1.07)	**0.82** (0.73, 0.91)
Snyder	**0.84** (0.72, 0.98)	NA	1.01 (0.88, 1.16)
Sullivan	0.63 (0.38, 1.07)	NA	**0.65** (0.47, 0.90)
Union	0.84 (0.53, 1.34)	NA	**0.80** (0.66, 0.98)
Wayne	**3.36** (1.83, 6.16)	NA	0.96 (0.59, 1.58)
Wyoming	**0.86** (0.76, 0.96)	NA	**1.15** (1.00, 1.32)
Model 2 – same as Model 1, but with NDVI not CSD, with county; only NDVI associations are shown			
Normalized difference vegetation index (NDVI)			
T1	1.0	1.0	1.0
T2	0.91 (0.85, 0.98)	0.77 (0.64, 0.92)	0.93 (0.87, 0.99)
T3	0.85 (0.75, 0.97)	0.76 (0.48, 1.19)	0.90 (0.84, 0.97)

* Logistic regression models using generalized estimating equations with robust standard errors; also adjusted for sex and age (age, age^2, age^3). Counties with at least one association that excluded 1.0 in confidence interval included in table (37 counties were included in total; 36 county indicators vs. Luzerne County as reference).

† NA = these counties did not have city minor civil divisions or did not converge due to small numbers.

STROBE Statement—Checklist of items that should be included in reports of ***case-control studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	3
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	3
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	5-6
Objectives	3	State specific objectives, including any prespecified hypotheses	6, 9
Methods			
Study design	4	Present key elements of study design early in the paper	6, 7
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	7
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls	7, 8
		(*b*) For matched studies, give matching criteria and the number of controls per case	7
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	7-9
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	7-9
Bias	9	Describe any efforts to address potential sources of bias	9, 10
Study size	10	Explain how the study size was arrived at	7
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	9, 10
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	9, 10
		(*b*) Describe any methods used to examine subgroups and interactions	9, 10
		(*c*) Explain how missing data were addressed	9, 10
		(*d*) If applicable, explain how matching of cases and controls was addressed	9, 10
		(*e*) Describe any sensitivity analyses	9, 10
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	7
		(b) Give reasons for non-participation at each stage	NA
		(c) Consider use of a flow diagram	NA
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	21, 22
		(b) Indicate number of participants with missing data for each variable of interest	NA
Outcome data	15*	Report numbers in each exposure category, or summary measures of exposure	21, 22

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	**23, 24**
		(*b*) Report category boundaries when continuous variables were categorized	**23, 24**
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	**NA**
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	**13**
Discussion			
Key results	18	Summarise key results with reference to study objectives	**13, 14**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	**16**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	**16, 17**
Generalisability	21	Discuss the generalisability (external validity) of the study results	**16, 17**
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	**2, 6**

*Give information separately for cases and controls.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.