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

B S Schwartz [ID],[1] Jonathan Pollak,[1] Melissa N Poulsen,[2] Karen Bandeen-Roche,[3] Katherine Moon [ID],[1] Joseph DeWalle,[2] Karen Siegel,[4] Carla Mercado,[4] Giuseppina Imperatore,[4] Annemarie G Hirsch [ID] [2]

For numbered affiliations see end of article.

Correspondence to
Professor B S Schwartz;
bschwar1@jhu.edu

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 to 2016 (n=15 888). Persons with diabetes were age, sex and year matched (1:5) to persons without diabetes (n=79 435). We used generalised 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) urbanised 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 (OR (95% CI)) with higher odds of type 2 diabetes (1.10 (1.04 to 1.16) and 1.34 (1.25 to 1.44), respectively). Urbanised areas (vs rural) also had increased odds of type 2 diabetes (1.14 (1.08 to 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 to 1.62)) and city census tracts in urbanised areas (1.33 (1.22 to 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.

INTRODUCTION

Diabetes is a common and costly chronic disease; in the USA in 2018, over 34 million individuals had diabetes, with annual

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 characterisation at more relevant contextual scales than many prior studies in a range of communities from urban to rural.
► We did not measure behavioural 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.

spending exceeding \$320 billion.[1] Diabetes occurrence varies by race/ethnicity and also evidences geographic disparities[2 3]; prevalence by county in the USA varies over a sevenfold 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 (eg, higher poverty, older populations) and barriers to healthcare access.[8 9]

Community characteristics that may underlie observed geographic disparities in type 2 diabetes include land use (eg, walkable vs automobile dependent), fitness, food and social (eg, deprivation, disorganisation) environments; greenspace (ie, 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* (eg, higher population density co-occurs with

higher deprivation and food availability and lower automobile dependence and greenness). Simultaneous 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 maximise between community differences. Subsequent analyses can then stratify by community type and evaluate well-characterised *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 US Census Bureau defines *urbanised 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 behaviourally 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 centres 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 USA.

Using previously reported methods,[20] we used Geisinger electronic health record (EHR) data from 1.6 million individuals to identify new onset type 2 diabetes from 2008 to 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
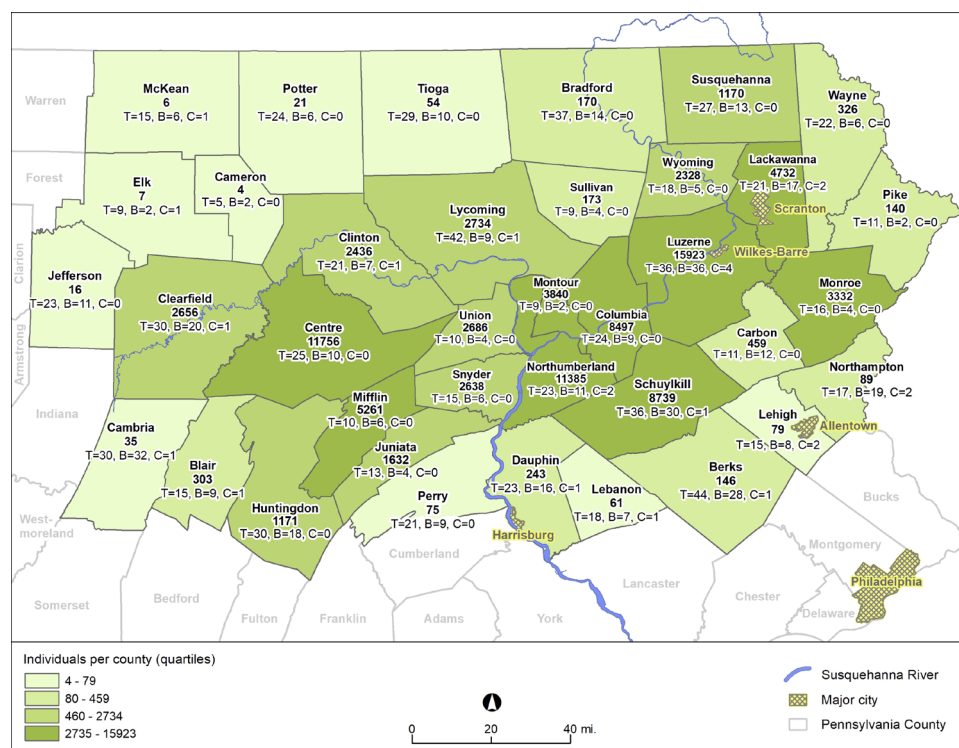

Figure 1 Distribution of study individuals and administrative community types by county in study region. The bold number is the number of individuals; T, B and C identify the number of townships, boroughs and city census tracts within each county that were included in the analysis.

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 supplemental 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 2 years prior without evidence of diabetes.

Community types and community features

Addresses at last contact with the health system were geocoded using ArcGIS V.10.4 (ESRI, Redlands, California, USA). 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 characterisations of the size and urbanicity of these areas (figure 2). First, using minor civil divisions and census tract boundaries, we categorised 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 5000–10 000 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 USA by population). Second, we used US Census Bureau's urbanised 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 categorisations, 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) normalised difference vegetation index (referred to as greenness) was evaluated in 1250 x 1250 m^2 around residences in the prior year.[29] We measured community socioeconomic deprivation using a previously described scale,[30] the sum

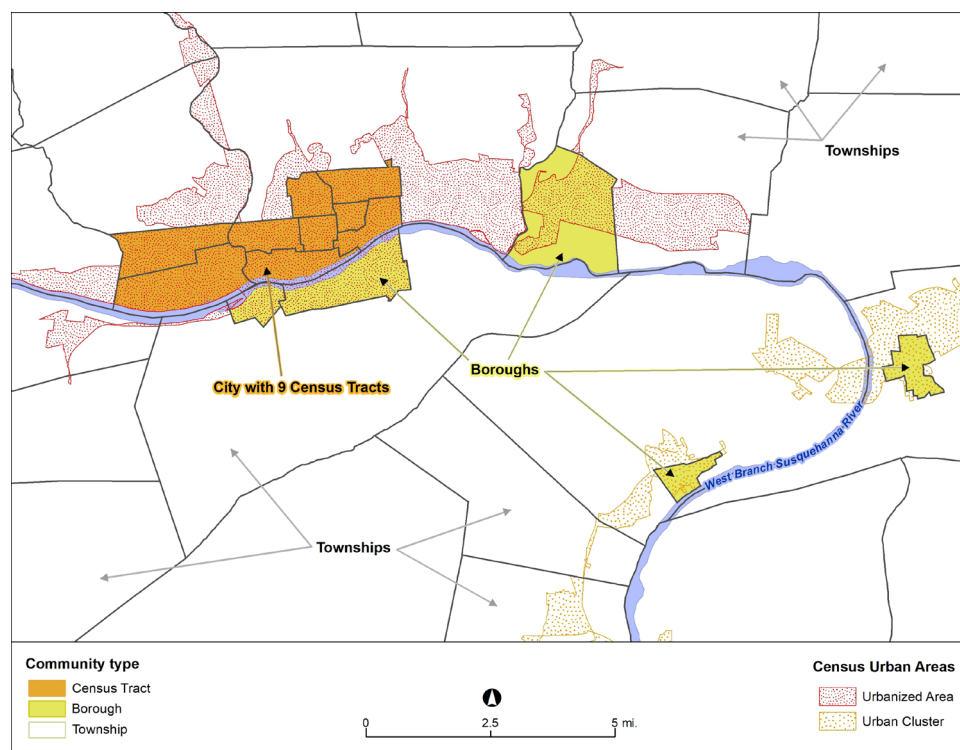**Figure 2** Areas along the Susquehanna River in Lycoming County, Pennsylvania from Williamsport (city) and South Williamsport (borough) to Montoursville (borough), Muncy (borough) and Montgomery (borough), showing relations between administrative community types (townships, boroughs and city census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

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 V.15.1.

Logistic regression was used to estimate associations (ORs, 95% CIs) using generalised estimating equations with robust SEs 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 per cent of time using Medical Assistance (surrogate for family socioeconomic status ($\geq$50% vs <50%)).[31] We did not include body mass index (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 analysed associations of the four definitions of community, community socioeconomic deprivation (quartiles; fourth quartile (worst deprivation) reference group) and greenness (tertiles) with diabetes status. Due to concerns about non-overlapping distributions resulting in extrapolation rather than adjustment (ie, 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 supplemental

table S2). Over 40% of persons resided in rural areas (table 1). Most borough residents were divided between urbanised 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 urbanised areas. On average, townships had higher greenness and lower community socioeconomic deprivation compared with boroughs and city census tracts (online supplemental 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 supplemental table S3). Laboratory data confirmed that the categorisation of diabetes cases and controls was valid (online supplemental 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. ORs 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 ($\geq$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 urbanised 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 (eg, city census tracts in major urban and smaller urban areas highest, boroughs in these areas intermediate, versus 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 supplemental table S5). Associations of community socioeconomic deprivation with diabetes evidenced decreasing ORs across decreasing deprivation quartiles in all community

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–<20 years	304 (1.9)	1520 (1.9)	
20–<30 years	628 (4.0)	3140 (4.0)	
30–<40 years	1611 (10.1)	8055 (10.1)	
40–<50 years	3086 (19.4)	15 429 (19.4)	
50–<60 years	4286 (27.0)	21 428 (27.0)	
60–<70 years	3510 (22.1)	17 548 (22.1)	
70–<80 years	1737 (10.9)	8685 (10.9)	
80–<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
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)	

Continued

Table 1 Continued

Variable	Cases	Controls	P value*
Duration from first contact with health system to diagnosis/control selection date, years, n (%)			0.72
Quartile 1 (2–<5 years)	1860 (11.7)	9466 (11.9)	
Quartile 2 (5–<8 years)	2571 (16.2)	12 646 (15.9)	
Quartile 3 (8–<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)	
Urbanised area	4906 (30.9)	23 423 (29.5)	
Urban cluster	4469 (28.1)	21 028 (26.5)	

*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 SEs 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.
COL, column; NDVI, normalised difference vegetation index.

types, but only crossed an inferential threshold in city census tracts, with approximately 25% lower odds in the first versus fourth 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 supplemental 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 first quartile (vs fourth) 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.

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

*Logistic regression models using generalised estimating equations with robust SEs; 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, age², age³) 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 CI 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.

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 US geographic differences in diabetes likely exist at multiple, nested spatial scales. Some of the county associations were of high magnitude (eg, 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 healthcare, since, on average, individuals in the study, regardless of Medical Assistance status and community type, had high contact with the healthcare 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 ecological level, relying on self-reported diabetes and focused on prevalent diabetes by county (too large and heterogeneous) or census tract (not experientially and behaviourally 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 behavioural 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 characterisation 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 (ie, urban) and geographic scale (ie, as localised 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 (eg, greater access to grocery stores, more walkable neighbourhoods, more commercial physical activity opportunity establishments) and a larger role for individual and community demographic and socioeconomic factors found in the same areas.

Author affiliations
[1]Environmental Health and Engineering, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA
[2]Population Health Sciences, Geisinger, Danville, Pennsylvania, USA
[3]Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA
[4]National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Map disclaimer The depiction of boundaries on this map does not imply the expression of any opinion whatsoever on the part of *BMJ* (or any member of its group) concerning the legal status of any country, territory, jurisdiction or area or of its authorities. This map is provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The study was approved by the Geisinger Institutional Review Board under waivers of consent and assent to use electronic health record (EHR) data.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Deidentified data are available upon request with IRB approval and a data use agreement.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
B S Schwartz http://orcid.org/0000-0002-0739-9865
Katherine Moon http://orcid.org/0000-0002-7604-075X
Annemarie G Hirsch http://orcid.org/0000-0001-7699-2171

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*. Atlanta, GA: Centers for Disease Control and Prevention, 2020.

2 Garcia MC, Faul M, Massetti G, *et al*. Reducing potentially excess deaths from the five leading causes of death in the rural United States. *MMWR Surveill Summ* 2017;66:1–7.

3 Ford ES, Mokdad AH, Giles WH, *et al*. Geographic variation in the prevalence of obesity, diabetes, and obesity-related behaviors. *Obes Res* 2005;13:118–22.

4 Cunningham SA, Patel SA, Beckles GL, *et al*. County-level contextual factors associated with diabetes incidence in the United States. *Ann Epidemiol* 2018;28:e2:20–5.

5 Rural Health Information Hub. Why diabetes is a concern for rural communities 2020 April 13, 2020. Available: https://www.ruralhealthinfo.org/toolkits/diabetes/1/rural-concerns

6 Singh GK, Siahpush M. Widening rural-urban disparities in all-cause mortality and mortality from major causes of death in the USA, 1969-2009. *J Urban Health* 2014;91:272–92.

7 James CV, Moonesinghe R, Wilson-Frederick SM, *et al*. Racial/Ethnic Health Disparities Among Rural Adults - United States, 2012-2015. *MMWR Surveill Summ* 2017;66:1–9.

8 Cosby AG, McDoom-Echebiri MM, James W, *et al*. Growth and persistence of Place-Based mortality in the United States: the rural mortality penalty. *Am J Public Health* 2019;109:155–62.

9 Henning-Smith CE, Hernandez AM, Hardeman RR, *et al*. Rural counties with majority black or Indigenous populations suffer the highest rates of premature death in the US. *Health Aff* 2019;38:2019–26.

10 Maier W, Scheidt-Nave C, Holle R, *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:e89661.

11 Müller G, Kluttig A, Greiser KH, *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;178:221–30.

12 Jagai JS, Krajewski AK, Shaikh S, *et al*. Association between environmental quality and diabetes in the USA. *J Diabetes Investig* 2020;11:315–24.

13 Honold J, Beyer R, Lakes T, *et al*. Multiple environmental burdens and neighborhood-related health of City residents. *J Environ Psychol* 2012;32:305–17.

14 Mavoa S, Bagheri N, Koohsari MJ, *et al*. How do neighbourhood definitions influence the associations between built environment and physical activity? *Int J Environ Res Public Health* 2019;16. doi:10.3390/ijerph16091501. [Epub ahead of print: 28 Apr 2019].

15 Dekker LH, Rijnks RH, Navis GJ. Regional variation in type 2 diabetes: evidence from 137 820 adults on the role of neighbourhood body mass index. *Eur J Public Health* 2020;30:189–94.

16 Stafford M, Duke-Williams O, Shelton N. Small area inequalities in health: are we underestimating them? *Soc Sci Med* 2008;67:891–9.

17 Tuson M, Yap M, Kok MR, *et al*. Incorporating geography into a new generalized theoretical and statistical framework addressing the modifiable areal unit problem. *Int J Health Geogr* 2019;18:6.

18 Bennett KJ, Borders TF, Holmes GM, *et al*. What is rural? challenges and implications of definitions that inadequately encompass rural people and places. *Health Aff* 2019;38:1985–92.

19 Census Bureau. Geography program: 2010 census urban and rural classification and urban area criteria, 2018. Available: https://www.census.gov/programssurveys/geography/guidance/geoareas/urban-rural/2010-urbanrural.html [Accessed 5 Jan 2020].

20 Hirsch AG, Durden TE, Nordberg C, *et al*. Associations of Four Community Factors With Longitudinal Change in Hemoglobin A$_{1c}$ Levels in Patients With Type 2 Diabetes. *Diabetes Care* 2018;41:461–8.

21 Cohen SA, Cook SK, Kelley L, *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;33:167–79.

22 James WL. All rural places are not created equal: revisiting the rural mortality penalty in the United States. *Am J Public Health* 2014;104:2122–9.

23 Astell-Burt T, Feng X, Kolt GS. Is neighborhood green space associated with a lower risk of type 2 diabetes? Evidence from 267,072 Australians. *Diabetes Care* 2014;37:197–201.

24 Müller G, Harhoff R, Rahe C, *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:e019062.

25 Casey JA, Savitz DA, Rasmussen SG, *et al*. Unconventional natural gas development and birth outcomes in Pennsylvania, USA. *Epidemiology* 2016;27:163–72.

26 Lawrence JM, Black MH, Zhang JL, *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;179:27–38.

27 Zhong VW, Pfaff ER, Beavers DP, *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;15:573–84.

28 Schwartz BS, Stewart WF, Godby 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:e17–28.

29 Casey JA, James P, Rudolph KE, *et al*. Greenness and birth outcomes in a range of Pennsylvania communities. *Int J Environ Res Public Health* 2016;13. doi:10.3390/ijerph13030311. [Epub ahead of print: 11 Mar 2016].

30 Nau C, Schwartz BS, Bandeen-Roche K, *et al*. Community socioeconomic deprivation and obesity trajectories in children using electronic health records. *Obesity* 2015;23:207–12.

31 Casey JA, Pollak J, Glymour MM, *et al*. Measures of Ses for electronic health Record-based research. *Am J Prev Med* 2018;54:430–9.

32 Petersen ML, Porter KE, Gruber S, *et al*. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res* 2012;21:31–54.

33 Cox M, Boyle PJ, Davey PG, *et al*. Locality deprivation and type 2 diabetes incidence: a local test of relative inequalities. *Soc Sci Med* 2007;65:1953–64.

34 Maier W, Holle R, Hunger M, *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:e78–86.

35 James P, Banay RF, Hart JE, *et al*. A review of the health benefits of Greenness. *Curr Epidemiol Rep* 2015;2:131–42.

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

37 Geiss LS, Kirtland K, Lin J, *et al*. Changes in diagnosed diabetes, obesity, and physical inactivity prevalence in US counties, 2004-2012. *PLoS One* 2017;12:e0173428.

38 Liese AD, Lawson A, Song H-R, *et al*. Evaluating geographic variation in type 1 and type 2 diabetes mellitus incidence in youth in four us regions. *Health Place* 2010;16:547–56.

39 National Center for Chronic Disease Prevention and Health Promotion. Division of diabetes translation at a glance, 2019. Available: https://www.cdc.gov/chronicdisease/resources/publications/aag/diabetes.htm

40 O'Connor A, Wellenius G. Rural-Urban disparities in the prevalence of diabetes and coronary heart disease. *Public Health* 2012;126:813–20.

41 Macintyre S, Ellaway A. Ecological approaches: rediscovering the role of the physical and social environment. In: Kawachi I, Berkman L, eds. *Social epidemiology*. New York: Oxford University Press, 2000: 332–48.

42 Openshaw S. *The modifiable areal unit problem*. Norwich, CT: GeoBooks, 1984.

43 Wong D. The modifiable areal unit problem (MAUP). In: Fotheringham AS, Rogerson PA, eds. *The SAGE Handbook of spatial analysis*. London: SAGE Publications, 2009: 105–24.

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