Explaining spatial accessibility to high-quality nursing home care in the US using machine learning

Abstract

In this study we measure and map the system-wide spatial accessibility to good quality nursing home care for all counties in the contiguous United States, and use an 'imputed post-lasso' machine learning technique to systematically examine this accessibility measure's associations with a broad range of county-level socio-demographic variables. Both steps were carried out using publicly available datasets. Analyses found clear evidence of spatial patterning in accessibility, particularly by population density, state and the populations of specific racial minorities. This has implications for outcomes that extend beyond the care homes and we highlight a number of policy measures that may help to address these shortcomings. The 'out-of-sample' predictive performance of the machine learning approach highlights the method's usefulness in identifying systematic differences in accessibility to services.

Keywords

accessibility; data science; equity; health economics; machine learning; nursing homes

1. Introduction

Understanding the social determinants of health is crucial to understanding disparities in health outcomes (Marmot 2005). A substantial literature has demonstrated the relationships between health and economic disadvantage, minority status and geographic isolation (Edward and

Biddle 2017, Marmot and Bell 2012), and subsequently how these influence access (Joseph and Phillips 1984, Millman 1993) which in turn affects healthcare utilisation (Haynes *et al.* 1999, Jones *et al.* 2008) and hence health outcomes (Astell-Burt *et al.* 2011, Kirby *et al.* 2017, Yang *et al.* 2006). In examining such relationships, several authors have employed a simplified dichotomous approach ('financial vs other' (Joseph and Phillips 1984, Millman 1993) or 'spatial vs aspatial' (Khan 1992)). While parsimonious, these fail to capture the complex inter-relationship between determinants. These relationships, alongside the impracticalities of performing experimental approaches in such circumstances (Petticrew *et al.* 2005), can make it difficult to isolate the underlying causal mechanisms by which they influence health. As a result, advances typically rest upon the gradual accumulation of further evidence from natural experiments and subsequently require careful interpretation (Kelly *et al.* 2010). As large datasets for analysis become increasingly available, it can be more and more difficult to identify which variables should be investigated as predictive (if not causative) of health outcomes. Machine learning allows a principled and systematic approach to such variable selection.

Spatial accessibility is of particular importance for nursing homes as location is the most frequently cited factor in the choice of home by residents (Shugarman and Brown 2006). McIntyre et al. (2009) group determinants into three broad categories of accessibility: availability (whether appropriate services are available where and when they are needed), affordability (the ability to pay and consideration of the opportunity costs of doing so) and acceptability (cultural perspectives/conditions that empower patients to use services and to 'fit' with provider attitudes). These categories can interact with and through each other, better capturing complex endogenous relationships and permitting a fuller picture of the connections between disadvantage, access and outcomes to be constructed.

Previous studies have shown that overall care quality has measurable impacts on nursing home (NH) residents’ health (Cornell *et al.* 2019, Unroe *et al.* 2012). Nursing homes located in areas with high levels of poverty (Park and Martin 2018) or in rural areas (Bowblis *et al.* 2013, Yuan *et al.* 2018) appear to provide statistically worse care, as do NHs that have a high composition of Medicaid-funded patients (Mor *et al.* 2004), due to increased fiscal stress (Park and Martin 2018). Patients simultaneously eligible for both Medicare and Medicaid are discharged from hospital to NHs with poorer care quality than those eligible for Medicare alone (Rahman *et al.* 2014). Minorities receive statistically worse care (Allsworth *et al.* 2005, Bliss *et al.* 2015), in homes which are often highly segregated by race (Mor, Zinn, Angelelli, Teno and Miller 2004). In fact, residents seem to seek out homes with a majority of their own race, even at the cost of proximity and care quality (Rahman and Foster 2015). Residents of socioeconomically disadvantaged counties (Yuan, Louis, Cabral, Schneider, Ryan and Kazis 2018) and poor neighbourhoods (Tamara Konetzka *et al.* 2015) have greater difficulties in accessing high-quality nursing homes.

We sought to identify counties that displayed poorer spatial accessibility to good quality nursing home care, and to seek to understand the characteristics of these counties in order to identify potential equity concerns. This spatial accessibility broadly corresponds to the

McIntyre’s “availability” category. It is known that certain racial groups have significantly worse health states (particularly Native Americans (Davis 2005)) when they enter nursing home facilities; we hypothesised that spatial barriers to accessing nursing home care (similarly to other health services) for certain groups could lead to underutilisation by these groups, and could therefore partially explain such differences in functional impairment. Such issues are not necessarily confined to populations of racial groups alone, and we sought to consider a broad range of further socio-demographic data. Quantifying spatial accessibility at a county level had the advantage of allowing the incorporation of a large number of variables from the American Community Survey (ACS) into our analyses, and furthermore facilitated the mapping of results. Because there are over 3000 counties and over 15000 homes, we did not wish to assume that simple accessibility measures (such as each county’s ratio of elderly residents to NHs) would suffice, so we used a more sensitive “gravity potential” model to do so. We thereafter created a predictive model that would help to identify socio-demographic factors that might explain the distribution of spatial accessibility around the country and, partially because of the sheer quantity of ACS variables, applied a machine learning approach to identify which variables had greater statistical power. This paper builds upon the prior literature through its two objectives:

1. to measure and map US county-level nationwide spatial accessibility to high quality nursing home care, and

2. to discover the most relevant socio-demographic variables associated with these accessibility levels.

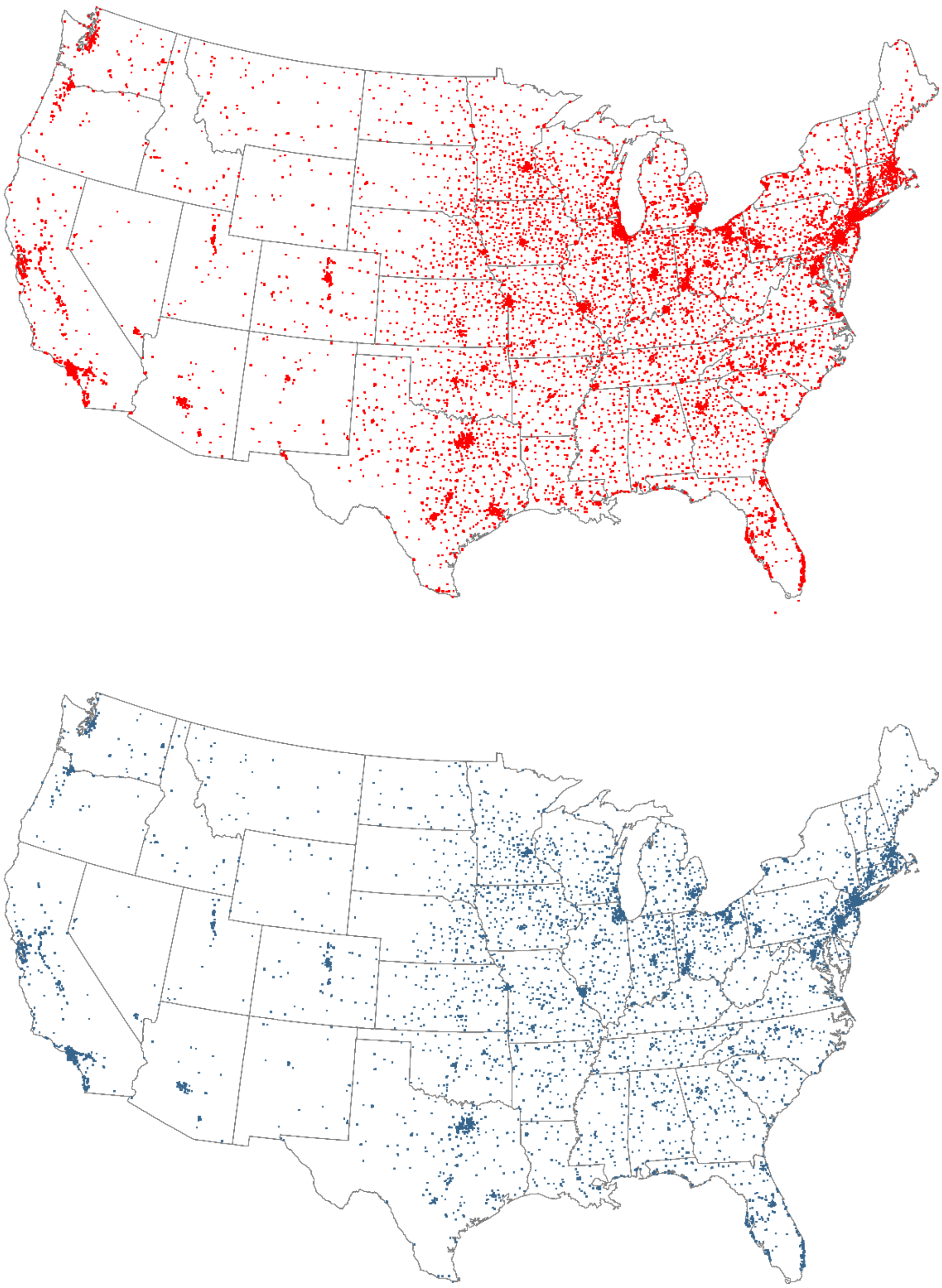
2. Methods

2.1 Data

The US Centers for Medicare and Medicaid Services (CMS) publish “Nursing Home Compare” (NHC) star ratings for all Medicare/Medicaid certified nursing homes in the US, alongside all underlying data used to construct these ratings. NHs are given a score of one (low) to five (high) stars based on their inspection records, staffing levels and quality measures; these are also thereafter combined into a single star rating of overall performance. We defined ‘high-quality’ NHs as those given an overall rating of 4 or 5 stars, as in the literature (e.g., Lutfiyya *et al.* (2013)), and data were obtained for Q3 of 2016. The address for each NH was geocoded (i.e. converted to a latitude-longitude coordinate) in R using the *geocode* package, and where this failed, using the *googleway* package. The counties in which nursing homes were located (which were rarely included in NH addresses) were then derived from this geolocation, alongside the associated Federal Information Processing Standard (FIPS) code, allowing further data linkage. Ignoring the small number of NHs where ratings are not yet available, addresses that could not be geocoded or are apparent duplicate records, 6,904 (45%) out of

15,215 NHs were classified as high-quality in 2016.Figure1 shows the locations of all geocoded homes and, separately, those that were classified as high quality.

Figure 1 – Geolocations of all 15,215 NHs identifiable in 2016 NHC data (top) and. 4 and 5-star ('high quality') NHs (bottom)



County-level estimates of socio-demographic data were available from the American Community Survey (ACS) (US Census Bureau) over a five-year window (2012-2016). The datasets included were “Comparative economic characteristics” (cp03), “ACS demographic and housing estimates” (dp05), “Age and sex” (s0101) and “Households and families” (s1101).

Aside from the ACS and NHC data, a small number of further variables were incorporated or derived for use in analyses: the counties’ population density, which we calculated based upon each county’s total population and county land area (obtained from the US Census gazetteer files (US Census Bureau 2018)); dummy variables representing which state the county is in; and counties’ Rural-Urban Continuum Codes (RUCC), as published by the US Department of Agriculture (2013). These describes how metropolitan a county is as one of nine categories; a binary urban/rural variable was also included that we derived based upon it (as in Yuan et al. (2018)).

2.2 Accessibility measures

For each county *i*, we calculated the population-weighted system-wide spatial accessibility to high-quality nursing home care based upon a “gravity potential” model (Talen and Anselin 1998). Such gravity models have been shown to be the most sensitive techniques for explaining population access to services (Song 1996). This accessibility measure took into account nearby NHs located across county boundaries, while ensuring that far away home NHs are given negligible weight. The model was similar to that used by Kalogirou and Foley (2006), whereby the spatial accessibility of each county *i* is calculated by

$$\mathrm{SA_i} = \sum_j \left[\left(n_j / (p_i) \right) \left(1/\max\left(d_{ij}^2, 1\right) \right) \right]$$

Where:

- n_j is the number of beds in each 'high-quality' nursing home *j*;
- p_i is the population over 65 in county *i* (taken from the ACS "DP05" dataset); and
- d_{ij} is the distance in kilometres between the centroid of county *i* to nursing home *j*. Distances less than 1km were set to this level to avoid attaching disproportionate weight to NHs close to the county centroid (resulting from calculating $1/d_{ij}^2$).

Greater values for SA indicate greater spatial access. For instance, a county *i* with a population over 65 of 100, in a country with only two high quality NHs, located 10km and 20km from county *i* respectively and containing 50 and 60 beds respectively, the spatial accessibility for county *i* would be given by $SA_i = \frac{50}{100*10^2} + \frac{60}{100*20^2} = 0.0065$.

Accessibility measures were calculated for all counties in the contiguous US except Ogala County, South Dakota, for which all ACS data was missing.

2.3 Descriptive analyses

An exploratory model was developed to identify the associations between measures of disadvantage and access to good quality care. Variables relating to the total/male/female population over 65 were excluded from the model, as the total had been used in the calculation of the accessibility measure. Some datasets included “margin of error” variables as well as point estimates (e.g. the estimated male population for Autauga County, Alabama was 26877, and the margin of error (recorded as a separate variable) was 120) – only point estimates were included in the model as it was felt that making predictions based on the latter would be difficult to interpret and would lack face validity. Variables present in multiple datasets (such as total population and population by race) were only included once, and variables that were extremely correlated (~0.99) with total population were excluded. In total, 472 variables across 3074 counties were included in the analysis. The full list of included and excluded variables is available in Appendix 1. Given the large number of variables, many of which were further correlated with each other; it was left to the machine learning approach to choose the most appropriate variables from the available list. Data was generally complete (ignoring Ogala County), except for the CPO3 dataset which lacked data for several hundred counties, which were geographically clustered in the Rocky Mountains. The accessibility score was highly skewed, as were most of the independent variables. A log transformation was applied to all variables, which reduced this skewness. For the accessibility measure, ln(SA) was used; for all other variables a ln(SA+1) transformation was used due to the prevalence of zeroes in some ACS data. Results were not sensitive to the use of the inverse hyperbolic sine function in place of the log transformation. The results of a predictive model using such data is reported in Appendix 2 and is broadly similar to that reported later in Table 2.

Our aim was to develop a simple, easy to interpret model that would systematically identify variables that were associated with county-level spatial accessibility. The first iteration of the model used an ordinary least squares approach, using a subset of the variables available that we knew from prior literature and our own judgement were likely to be significant. A random forest approach employed as part of sensitivity analyses, identified the most important variables as generally relating to Native American populations. We had not included any Native American variables in our baseline OLS model and decided a more systematic approach that cast a wide net in variable selection would be merited after all. We therefore sought to identify an appropriate machine learning approach. This allowed all variables to be to be incorporated into the analysis, rather than a subset, with penalisation used to select the most relevant variables.

An approach was used whereby Random Forest approaches (Breiman 2001) were used to impute data which was missing for the included variables (specifically for the cp03 dataset), and a Least Absolute Shrinkage and Selection Operator (lasso) approach subsequently used for variable selection; such an approach is called the “imputed lasso” (Lu and Petkova 2014). Lasso is a penalized regression approach that seeks to improve the prediction accuracy and interpretability of regression models by altering the model fitting process so that only a subset of the provided covariates are included in the final model (Tibshirani 1996). The imputed lasso allows for a more parsimonious and transparent model than a standalone random forest approach and overcomes potential biases, particularly around patterns in missing data (Lu and

Petkova 2014) – which may have otherwise been an issue for our analyses given the non-randomly distributed missing economic data in our data. For this reason, we used the imputed lasso approach, which had the added advantage of creating a relatively simple explanatory model to understand in broad terms of the relevant underlying processes and associations. The imputation and lasso were carried out using the *missForest* and *hdm* packages in R respectively.

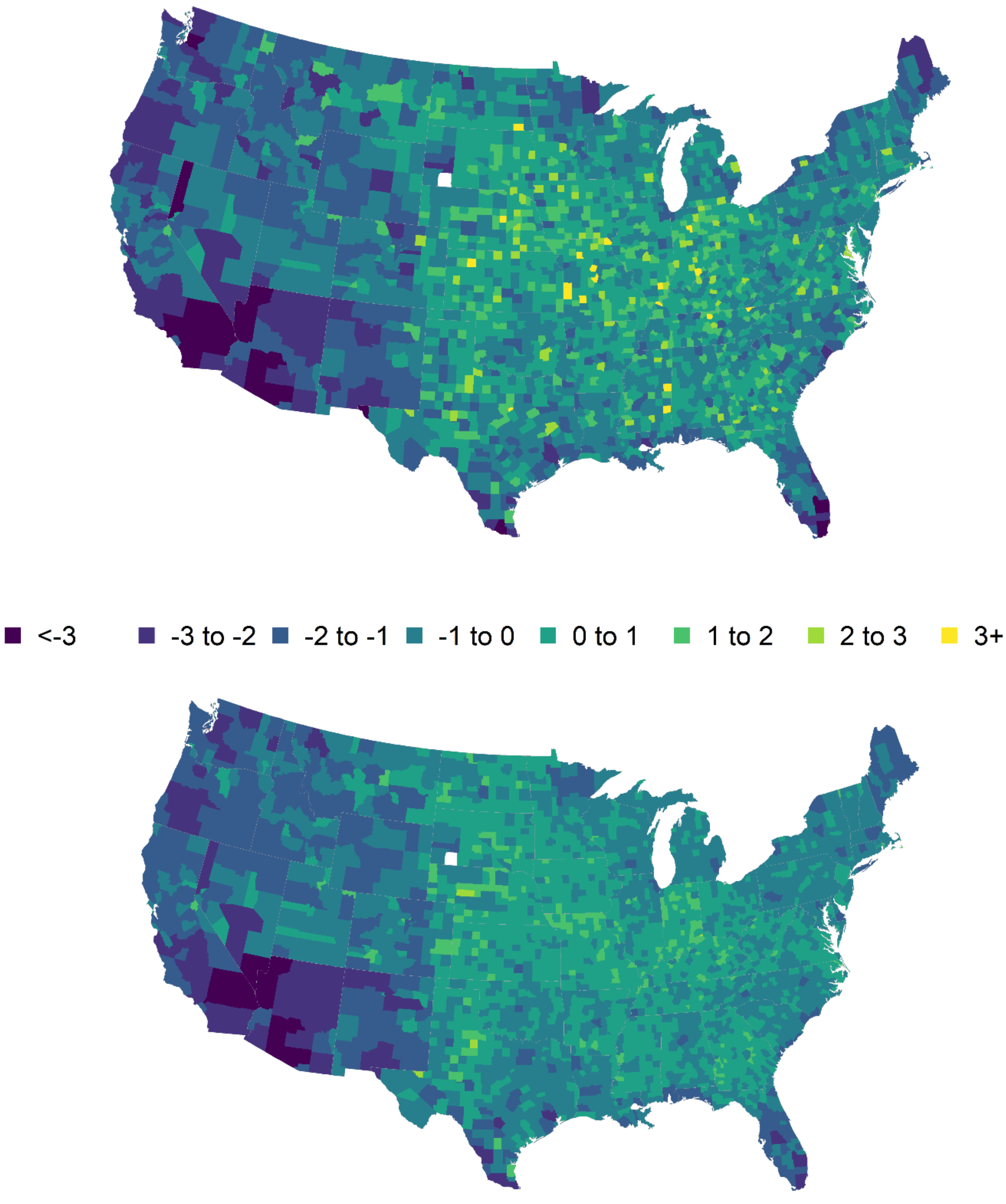
Potential predictors in lasso are typically centred and scaled – normalised with reference to the variable's mean and standard deviation, making them essentially z-scores. This was carried out for all variables, including the accessibility measure. The target variable was therefore finally defined as scale(ln(SA)), which indicated each county's accessibility compared to the national average. Data for 2,281 randomly chosen counties (75% of the data) were used for estimation (the 'training' data) and data for the remaining 25% of counties were used to validate the predictive models' out of sample performance (the 'test' data). Analyses were carried out using the post-lasso approach in the *hdm* R package (selecting the "double selection" method in *hdm*'s associated *rlassoEffect* function). This approach uses a data driven, non-arbitrary penalty function (Bach *et al.* 2018) and creates a more easily interpretable model both by further reducing the number of variables present and by allowing for the inference of post-selection confidence interval, providing further interpretability. Without the post-lasso approach, confidence intervals cannot be reliably estimated due to the introduced bias from variable exclusion (Bach, Chernozhukov and Spindler 2018). Using imputed lasso and post-lasso together should in principle therefore reduce the potential for bias in a number of ways, and to our knowledge, this is the first time such an “imputed post-lasso” has been used.

3. Results

3.1 Geographical accessibility

Figure 2 shows that accessibility to high-quality nursing home care is generally highest in a band from the eastern Rockies through the Midwest. Counties west of the Rockies generally have poorer access, particularly in the southwest and along the Pacific coast. Predicted accessibility derived from the lasso model is also shown.

Figure 2 – County level accessibility (top) and lasso predictions (bottom). All results are reported in standard deviations from the mean.


<-3
-3 to -2
-2 to -1
-1 to 0
0 to 1
1 to 2
2 to 3
3+

3.2 Predictive analyses

Table 1 provides an overview of the characteristics of counties according to quartile of high-quality nursing home accessibility, from Quartile 1 (low) to Quartile 4 (high). Counties with lower accessibilities tend to have higher populations; most likely because these counties tend to also have high absolute populations of elderly people, which is included in the calculation of the accessibility measure directly. It is interesting that median age and old-age dependency ratio rise with accessibility; it may be related to the fact that urban populations can be expected to be younger.

It is also initially counter-intuitive that income seems to be negatively associated with accessibility; this however may relate again to higher salaries in urban areas, which display a negative univariate relationship with spatial accessibility. It may also be that new homes do not choose to locate in disproportionately young areas, or where land values are most expensive (i.e. city centres). Mirroring the findings of the previously cited studies, counties that have the highest proportion of white people have the best access. This may at least partially be related to clustering of minority groups in cities.

Table 1- Median county levels of selected variables before transformation, for 4 equal quartiles split by accessibility (Q1 is lowest accessibility, Q4 is highest). "NA" is non applicable.

ACS Description	Data set	Code	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Total population- total- estimate	s0101	HC01_EST_VC01	80955	33508	16262	11752
Population density	*NA*	*popDensity*	32.6	22.8	13.3	11.1
*Income and benefits (in 2016 inflation-adjusted dollars) - total households - median household income (dollars) **	cp03	HC01_VC85	49356	44474	45471	45427
Summary indicators - sex ratio (males per 100 females)-total- estimate	s0101	HC01_EST_VC36	98.0	97.9	98.6	98.7
Race - one race - White-percent	dp05	HC03_VC49	85.5	89	92.3	94.2
Race - one race - Black or African American-percent	dp05	HC03_VC50	2.8	3.5	1.8	1.3
Race - one race - Asian-percent	dp05	HC03_VC56	1.1	0.6	0.4	0.4
Race - one race - American Indian and Alaska native-percent	dp05	HC03_VC51	0.6	0.3	0.3	0.2
Race - one race - Native Hawaiian and Other Pacific Islander-percent	dp05	HC03_VC64	0	0	0	0
Hispanic or Latino and race - total population - Hispanic or Latino (of any race)-percent	dp05	HC03_VC88	7.9	3.6	2.9	2.7
Sex and age - median age (years)-estimate	dp05	HC01_VC23	38.8	40.7	41.7	42.2
Sex and age - 65 years and over-estimate	dp05	HC01_VC29	13428	5915	3003	2129
Summary indicators - age dependency ratio - old-age dependency ratio-total-estimate	s0101	HC01_EST_VC38	25.4	28.3	30	30.9
Average household size-Total- Estimate	s1101	HC01_EST_VC03	2.55	2.50	2.48	2.46
Commute time – Mean travel time – minutes	cp03	HC01_VC36	22.2	23.6	24.1	24.2
Accessibility (after scaling)	*NA*	*NA*	-1.05	-0.28	0.29	1.02

The coefficients associated with the exploratory predictive model reported in Figure 2 are shown in Table 2. Where possible, these are grouped by general headings, such as the percentage of the civilian population over 16 employed in a given industry, sex/age and race.

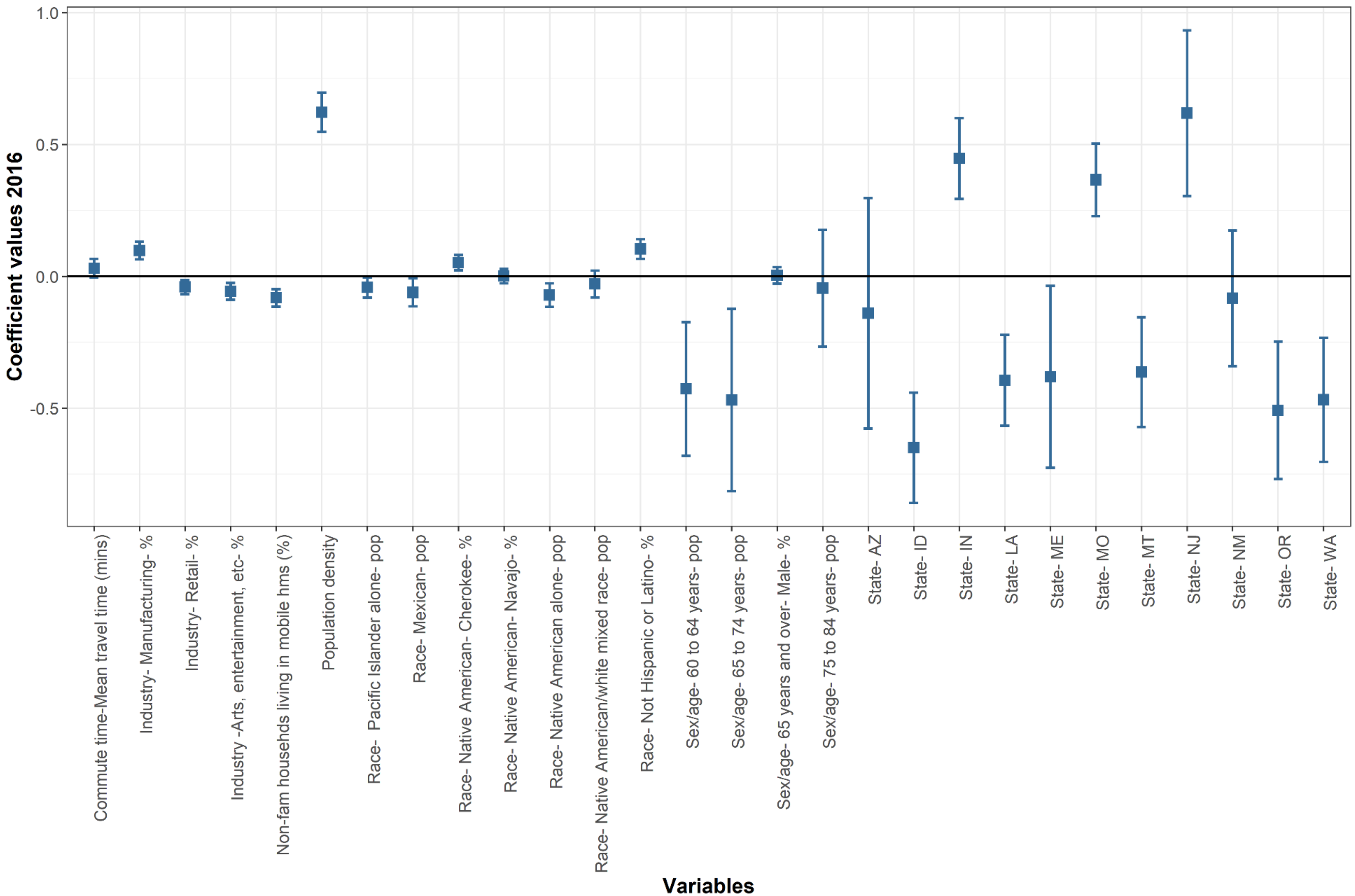
Because the variables were scaled, coefficients that are larger in absolute terms can be said to have a larger effect.

Table 2 – Variables selected by post-lasso algorithm (using 'rlassoEffects' in the 'hdm' R package) for models that estimate spatial accessibility for 2016. Excluded variables (those given a weight of zero) are not shown.

	Estimate	Std. Error	t value	Pr(>|t|)		Data set	Code
Commute time – Mean travel time – minutes	0.034	0.019	1.839	0.07	.	cp03	HC01_VC36
Households living in mobile/other structures – non-family – %	-0.089	0.017	-5.092	<0.001	***	s1101	HC05_EST_VC29
Population density	0.675	0.041	16.493	<0.001	***	*NA*	*popDensity*
Industry of civilian employment							
Arts, entertainment, accommodation and food – %	-0.061	0.016	-3.822	<0.001	***	cp03	HC01_VC60
Manufacturing – %	0.102	0.018	5.746	<0.001	***	cp03	HC01_VC52
Retail – %	-0.042	0.014	-3.1	0.002	**	cp03	HC01_VC54
Race							
American Indian and Alaska Native – Cherokee – %	0.054	0.014	3.942	<0.001	***	dp05	HC03_VC52
American Indian and Alaska Native – Navajo – %	0.001	0.008	0.172	0.86		dp05	HC03_VC54
American Indian and Alaska Native/white mixed race – pop	-0.03	0.024	-1.26	0.21		dp05	HC01_VC72
American Indian/Alaska Native alone – pop	-0.075	0.023	-3.32	0.001	***	dp05	HC01_VC96
Ethnic Mexican – pop	-0.063	0.027	-2.328	0.020	*	dp05	HC01_VC89
Native Hawaiian and Other Pacific Islander alone – pop	-0.044	0.021	-2.069	0.039	*	dp05	HC01_VC98
Not Hispanic or Latino – %	0.109	0.018	5.886	<0.001	***	dp05	HC03_VC93
Sex and/or age group							
60 to 64 years – pop	-0.437	0.11	-3.976	<0.001	***	dp05	HC01_VC17
65 to 74 years – pop	-0.486	0.142	-3.409	0.001	***	dp05	HC01_VC18
75 to 84 years – pop	-0.048	0.102	-0.468	0.64		dp05	HC01_VC19
65 years and over – % of population that is Male	0.005	0.014	0.338	0.74		dp05	HC03_VC38
State dummy variables							
Arizona	-0.171	0.119	-1.441	0.15		*State*	*StateAZ*
Idaho	-0.668	0.093	-7.162	<0.001	***	*State*	*StateID*
Indiana	0.458	0.104	4.401	<0.001	***	*State*	*StateIN*
Louisiana	-0.402	0.066	-6.119	<0.001	***	*State*	*StateLA*
Maine	-0.397	0.073	-5.458	<0.001	***	*State*	*StateME*
Missouri	0.375	0.098	3.811	<0.001	***	*State*	*StateMO*
Montana	-0.382	0.099	-3.848	<0.001	***	*State*	*StateMT*
New Jersey	0.652	0.141	4.605	<0.001	***	*State*	*StateNJ*
New Mexico	-0.097	0.115	-0.85	0.40		*State*	*StateNM*
Oregon	-0.542	0.09	-6.048	<0.001	***	*State*	*StateOR*
Washington	-0.49	0.113	-4.336	<0.001	***	*State*	*StateWA*

The lasso was successful in generating a relatively parsimonious model that substantially reduced the number of variables included. 28 of the 472 possible variables were selected, of which 11 were state dummies; other states were considered close enough to the average to not merit inclusion in their own right. The model appears to correspond reasonably well to the actual data, as seen in Figure 2. Its adjusted r-square for the test dataset was 0.613 (for the training set it was 0.597) and the corresponding mean absolute error for the test dataset was 0.451 (0.444 for the training set). Estimation of confidence intervals around the coefficients are shown in both Table 2 and Figure 4. Except for state dummy variables, confidence intervals are generally narrow.

Figure 3 – Variables' coefficient valuation estimates and associated 95% confidence intervals.



Population density can be seen to have a noticeable and large positive effect; perhaps a surprising finding given that it had a negative association with accessibility at a univariate level (shown in Table 1). Otherwise state dummies often have large effect sizes and demonstrate some geographic clustering – e.g. Oregon and Washington state exhibiting inferior access. Independent of state effects, counties with a large number of residents in mobile homes, counties with a relatively large elderly population or high numbers of ethnic Mexican, Native American or Pacific Islander residents appear to experience poorer access, while those in industrialised counties and those in more densely populated counties would appear to enjoy superior access, other things being controlled for. Median commute time is marginally positive, implying that the most accessible areas may be the hinterlands surrounding urban areas (as with Reddy (2020)).

While the total, male and female populations for above 65 were not included in the prediction model, other age-related variables – which clearly correlate with these figures – have been selected by the lasso model and show a large negative effect, as might be expected.

The Midwest's high level is accessibility is evident in the model, both in terms of the coefficients associated with the state dummy variables, as well as through other factors, such as their historic association with the manufacturing industry. For counties in Indiana with a high level of workers in manufacturing, for example, the effects are assumed to be additive.

The results for a corresponding analysis of accessibility in 2011, including a predictive analysis using the same imputed post-lasso approach (using ACS data from 2007-2011), were largely consistent with the 2016 findings reported in this paper. An overview of these results is shown in Appendix 3.

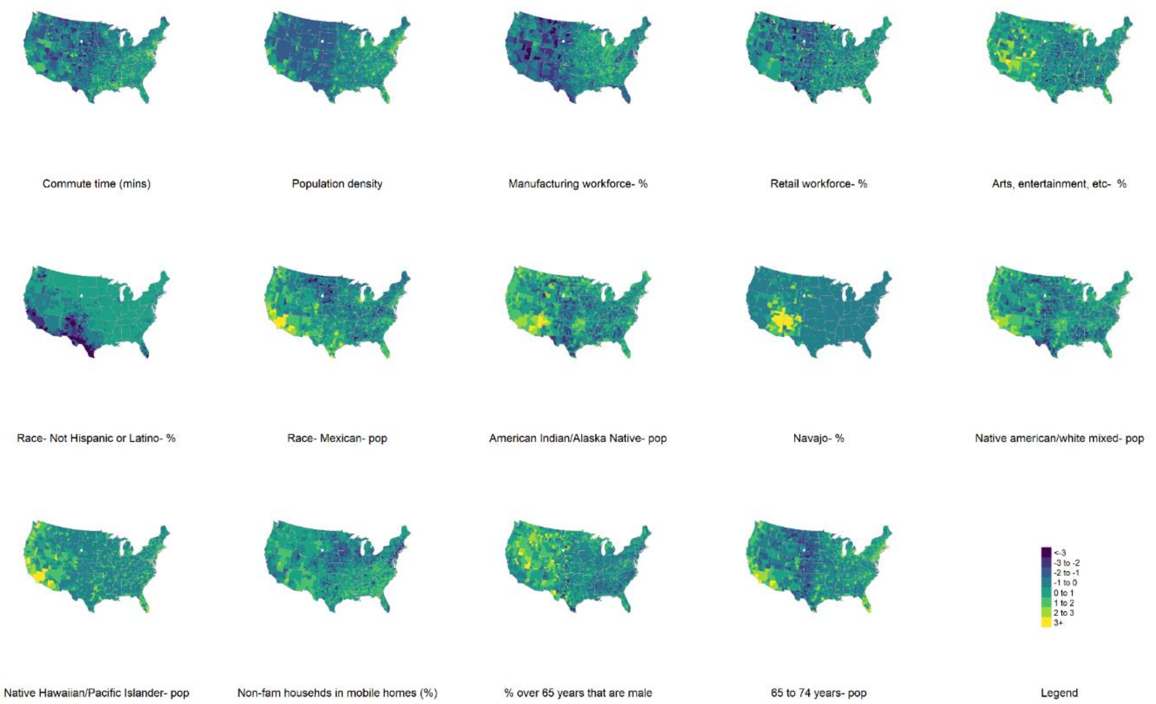
4. Discussion

The results make clear that geographic inequalities in access to high-quality NH care exist, and that these do not take the form of “unpatterned inequality” (Talen 1997). Rather, the variation in accessibility follows a clear pattern: population age distribution, ethnicity, population density and populations living in mobile homes appear to dominate these models, alongside state effects. The fact that NH locations were geocoded based upon NHC central datasets may contribute to the model’s ability to discern such socio-economic patterns reliably (McLafferty *et al.* 2012).

That population density is strongly positively associated with accessibility to high-quality NHs may reflect simple economics – where demand is high because the population is concentrated, so too will be supply of NHs in general (including good quality NHs). Nonetheless, since high-quality nursing homes tend to cluster in wealthy areas (Tamara Konetzka, Grabowski, Perraillon and Werner 2015), this picture may be more nuanced. It may also reflect previous findings that homes in urban counties have higher star ratings (Lutfiyya, Gessert and Lipsky 2013), and hence counties with a high population density might be expected to be closer to high quality homes. Table 1 showed that county population densities are in fact negatively associated with accessibility on a univariate level.

Regardless of methodology used, access is particularly poor for counties with high numbers of ethnic Mexicans, Native American and Pacific Island populations. These groups are clustered in the western half of the country (visible in Figure 4), where spatial accessibility is lowest anyway; for counties with high populations of such groups, accessibility is even worse, given they were found to be significant independent of state effects. The one unequivocally positive coefficient relating to race is the proportion of non-Hispanic residents. While race has a complicated relationship with income in the US historically, it is noteworthy that various measures of income, poverty and SNAP payments included, race *still* comes through as significantly correlated with access. The coefficient for Cherokee populations is positive, but does not counteract the negative coefficient of Native American status; as a result, areas with high Cherokee populations can be considered to have less bad accessibility than other areas with high proportions of Native Americans, but still worse than would otherwise be expected.

Figure 4 - 2016 spatial distribution of transformed data, by standard deviations from their respective means, for variables selected for both 2011 and 2016 predictive models. For more information on the 2011 model, see Appendix 3



While measures of average income were not selected by the model, it is notable that wealth-related indicators such as the numbers living in mobile homes was a significant determinant of access (especially since no variables included in the model directly related to wealth). In our dataset, non-family mobile home residency was highly negatively correlated (<-0.6) with mean family income and the proportion of women employed in the labour force, and similarly positively associated with the proportion of families below the poverty level and proportions of the population not in employment. It may therefore be included by the model as a composite proxy for several wealth and income related factors. While wealth and income are obviously related, privately-funded residents of NHs (which are strongly related to nursing home quality (Grabowski 2001, Park and Martin 2018)) are likely to rely on previously accumulated wealth rather than current earnings. That areas where there has been little accumulation of real estate equity and relatively high levels of poverty are less likely to present attractive areas for care homes to invest is unsurprising.

It may seem surprising that counties with high populations of African Americans who have consistently (on an individual level) been shown to experience barriers to access health services generally (Marmot 2005, Millman 1993) – including nursing homes (C. Reed and Andes 2001, Tamara Konetzka, Grabowski, Perraillon and Werner 2015, Yuan, Louis, Cabral, Schneider, Ryan and Kazis 2018) – generally do not appear to face *spatial* barriers. It may be that the physical distances involved between more affluent white and less affluent African American communities are much less than those experienced by more physically isolated groups such as Native Americans and Hispanic groups. Thus, it is worth remembering that physical proximity in terms of the measures used here is a necessary but not sufficient condition for access to good quality care in an environment close to one's friends and family. That whites are more likely to live in mobile homes (Johnson *et al.* 2018), is also worth noting in this context given its relationship with spatial access.

Spatial accessibility is not the only form of accessibility, however, and only part of McIntyre's (2009) "availability". It is possible that areas with high accessibility by our measure may have yet have low levels of accessibility depending on other factors (Ngui and Vanasse 2012), and the results further highlight the interconnectedness between availability, affordability and accessibility.

In terms of affordability, Medicare's low reimbursement rates (Grabowski 2001) inevitably influence the distribution of NHs. Increasing this rate in general – or in a targeted way focussing on areas with the lowest level of spatial accessibility – may go some way to addressing this, given the clear predictions found between economically deprived counties (and relatively poor ethno-cultural groups) and spatial accessibility. Expanding Medicare/Medicaid to allow payments for complementary/substitute services such as home care (which is not currently covered) may also improve equity. Though there are practical challenges in doing so in sparsely populated rural areas – alongside the cultural challenges previously described – both issues could potentially be ameliorated by paying for care to provided locally from 'within' these communities.

Racial groups are not homogenously distributed around the country and it is clear that certain groups (by coincidence or otherwise) happen to be clustered in its most poorly served regions. Even if spatial accessibility were improved in such areas, levels of overall accessibility may remain poor if practical cultural barriers (such as being able to talk to a doctor in your language, or a provider who 'gets' your culture's norms) are not addressed appropriately. For historic and cultural reasons, Native Americans in particular may face barriers to leaving their communities and homelands – but at the same time it appears that such sparsely populated and relatively poor areas cannot attract high-quality care homes given current market incentives.

4.1 Limitations

Lasso has been criticised for the fact that in situations where there are multiple highly correlated variables (as was the case in our dataset) which are competing for model inclusion, typically only one such variable will be chosen in the final model. Elastic nets are an alternative technique that tends to include all such correlated variables (at reduced relative weighting) rather than choosing between these, but naturally does so at the cost of including more variables in the final model. Given that we consciously intended to create a simplified and parsimonious model, "the advantage of choosing several correlated items together versus only one item from a group of correlated items is not clear" (Lu and Petkova 2014). Furthermore, it is not currently possible to derive meaningful confidence intervals for elastic net models (at least using packages available in R), which would have further undermined the interpretability of findings. Nonetheless, any future studies using the approach outlined in this paper should carefully consider the robustness of findings to the choice of variables for inclusion in LASSO, if they are to be used to inform practice. This may be particularly important where the variables act as proxies for an underlying unobserved variables, as was the case for mobile home residency here.

County-level data estimates being available only over a 5-year window means that there is no perfect way to match a given year of nursing home data with ACS data. Furthermore, Ogala County has a predominantly Native American population which has been excluded from our analyses.

This study used several simplifying assumptions, especially due to the computational intensiveness of several stages. The accessibility measure implicitly assumes equal underlying ‘need’ for nursing home care for all people over 65, regardless of income, cultural group, family situation and so on. It also assumes all high-quality nursing homes are effectively interchangeable, and so does not consider locational interdependence of services or agglomeration effects (White 1979).

The fact that county centroids were used to represent the county’s location may introduce bias if there are recurring national patterns about the groups who happen to live closer or further away from centroids. We used Haversine distance from county geographical centroid to care homes, rather than road distance, which arguably could have led to more accurate results in terms of time/barriers to access (Houston 2005) – although there are challenges with measuring these also (Delamater *et al.* 2012). The study also did not consider access to public transport facilities, or the likelihood of subgroups of local populations to rely on these – which may further impact on accessibility (Arcury *et al.* 2005, Syed *et al.* 2013) and equity more generally.

This accessibility measure used the total number of beds in each NH, rather than total number of *free* beds. This was because occupancy rates may be expected to fluctuate more over time, leading to unstable and less meaningful results; however, this means that any consistent patterns or variation in accessibility related to occupancy has not been captured in the model, which may lead to bias (Houston 2005).

4.2 Suggestions for further research

Given the large quantity of variables under investigation, the extent to which individual coefficients may vary regionally has not been investigated using geographically weighted regression, though it would be interesting to see if such variation exists.

It may be interesting to apply the analytical approaches employed in this paper against the star ratings for staffing levels, quality measures and inspections, further building upon the work of Yuan et al. (2018).

It may also be interesting to investigate the impacts of accessibility on both competition and care quality. Zhao (2016) previously showed that competition (measured by the Herfindahl-Hirschman index) and quality have a mixed relationship (though they found that easy-to-understand information is useful in driving up quality). Spatial accessibility may be an interesting alternative measure for competition, worthy of further investigation– and one which may lead to better patient information given reputational impacts of nearby NHs.

5. Conclusions and implications

This is the first paper to our knowledge that addresses county-level spatial accessibility to high-quality nursing home care. We provide a formal definition of accessibility and calculated this metric for each county in the contiguous USA. We thereafter investigated racial, socio-demographic and economic factors' associations with accessibility, using an innovative application of machine learning techniques. The paper illustrates that such a machine learning approach can be used to cast a wide net and select the most important such variables, while creating a parsimonious model that describes spatial accessibility. This approach could thereby help identify heretofore unknown areas for targeted follow-up analyses, such as better understanding whether specific barriers to access exist for Pacific Island populations.

Spatial accessibility was found to be particularly high in the Midwest and low in the southwest and along the Pacific coast. This analysis found there to be several issues – alongside geographic location – that are tied up with access to high-quality nursing home care, including: the size of the county's elderly, ethnic its population density, proportions living in mobile homes, patterns in local employment, and Hispanic, Native American and Pacific Islander populations. It is noteworthy that despite the inclusion of hundreds of variables, some of the best predictors of accessibility to NH care related to local populations of specific minority racial groups. The model's out of sample predictions were relatively accurate given that the independent variables used only socio-demographic data and excluded the seemingly more relevant Nursing Home Compare datasets.

Tests of equity of access determine whether there are "systematic differences in use and outcomes among groups in U.S. society" (Millman 1993) arising from barriers to care. This paper provides clear evidence that there are systematic differences between racial and economically disadvantaged groups in terms of their geographical access to nursing homes.

This may also go some way to explaining the differences in use of nursing homes between these groups (Davis 2005, Edward and Biddle 2017, Thomeer et al. 2014). We do not claim that this is causal, but believe that the clear associations found merits further study.

The results of our analyses are consistent with the inverse care law, that “the availability of good medical care tends to vary inversely with the need of the population served” (Hart 1971), which tends to arise where a free market decides where such facilities are to be located. Amelioration of this process requires increased government intervention and redistribution efforts, and it behoves decision makers to consider whether action is required to address the spatial inequities that this paper has demonstrated.

References

J. E. Allsworth, R. Toppa, N. C. Palin and K. L. Lapane (2005) Racial and ethnic disparities in the pharmacologic management of diabetes mellitus among long-term care facility residents. *Ethn Dis,* 205-212.

T. A. Arcury, J. S. Preisser, W. M. Gesler and J. M. Powers (2005) Access to transportation and health care utilization in a rural region. *The Journal of Rural Health,* 31-38.

T. Astell-Burt, R. Flowerdew, P. J. Boyle and J. F. Dillon (2011) Does geographic access to primary healthcare influence the detection of hepatitis C? *Social Science & Medicine,* 1472-1481.

P. Bach, V. Chernozhukov and M. Spindler (2018) Valid Simultaneous Inference in High-Dimensional Settings (with the hdm package for R). *arXiv preprint arXiv:1809.04951*.

D. Z. Bliss, O. Gurvich, K. Savik, L. E. Eberly, S. Harms, C. Mueller, J. F. Wyman, J. Garrard and B. Virnig (2015) Are there racial-ethnic disparities in time to pressure ulcer development and pressure ulcer treatment in older adults after nursing home admission? *Journal of aging and health,* 571-593.

J. R. Bowblis, H. Meng and K. Hyer (2013) The Urban-Rural Disparity in Nursing Home Quality Indicators: The Case of Facility-Acquired Contractures. *Health Services Research,* 47-69.

L. Breiman (2001) Random forests. *Machine learning,* 5-32.

S. C. Reed and S. Andes (2001) Supply and segregation of nursing home beds in Chicago communities. *Ethnicity & health,* 35-40.

P. Y. Cornell, D. C. Grabowski, E. C. Norton and M. Rahman (2019) Do report cards predict future quality? The case of skilled nursing facilities. *Journal of Health Economics*.

J. A. Davis (2005) Differences in the health care needs and service utilization of women in nursing homes: Comparison by race/ethnicity. *Journal of women & aging,* 57-71.

P. L. Delamater, J. P. Messina, A. M. Shortridge and S. C. Grady (2012) Measuring geographic access to health care: raster and network-based methods. *International journal of health geographics,* 15.

J. Edward and D. J. Biddle (2017) Using geographic information systems (GIS) to examine barriers to healthcare access for Hispanic and Latino immigrants in the US south. *Journal of racial and ethnic health disparities,* 297-307.

D. C. Grabowski (2001) Medicaid reimbursement and the quality of nursing home care. *Journal of health economics,* 549-569.

J. T. Hart (1971) The inverse care law. *The Lancet,* 405-412.

R. Haynes, G. Bentham, A. Lovett and S. Gale (1999) Effects of distances to hospital and GP surgery on hospital inpatient episodes, controlling for needs and provision. *Social science & medicine,* 425-433.

D. S. Houston (2005) Methods to test the spatial mismatch hypothesis. *Economic Geography,* 407-434.

D. Johnson, R. Thorpe, J. McGrath, W. Jackson and C. Jackson (2018) Black–White Differences in Housing Type and Sleep Duration as Well as Sleep Difficulties in the United States. *International journal of environmental research and public health,* 564.

A. Jones, R. Haynes, V. Sauerzapf, S. Crawford, H. Zhao and D. Forman (2008) Travel times to health care and survival from cancers in Northern England. *European journal of cancer,* 269-274.

A. E. Joseph and D. R. Phillips (1984) *Accessibility and utilization: geographical perspectives on health care delivery*: Sage.

S. Kalogirou and R. Foley (2006) Health, Place and Hanly: modelling accessibility to hospitals in Ireland. *Irish Geography,* 52-68.

M. Kelly, A. Morgan, S. Ellis, T. Younger, J. Huntley and C. Swann (2010) Evidence based public health: a review of the experience of the National Institute of Health and Clinical Excellence (NICE) of developing public health guidance in England. *Social science & medicine,* 1056-1062.

A. A. Khan (1992) An integrated approach to measuring potential spatial access to health care services. *Socio-economic planning sciences,* 275-287.

R. S. Kirby, E. Delmelle and J. M. Eberth (2017) Advances in spatial epidemiology and geographic information systems. *Annals of epidemiology,* 1-9.

F. Lu and E. Petkova (2014) A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Statistics in medicine,* 401-421.

M. N. Lutfiyya, C. E. Gessert and M. S. Lipsky (2013) Nursing home quality: A comparative analysis using CMS Nursing Home Compare data to examine differences between rural and nonrural facilities. *Journal of the American Medical Directors Association,* 593-598.

M. Marmot (2005) *Status syndrome: How your social standing directly affects your health*: A&C Black.

M. Marmot and R. Bell (2012) Fair society, healthy lives. *Public health,* S4-S10.

D. McIntyre, M. Thiede and S. Birch (2009) Access as a policy-relevant concept in low-and middle-income countries. *HealtH economIcs, polIcy and laW,* 179-193.

S. McLafferty, V. L. Freeman, R. E. Barrett, L. Luo and A. Shockley (2012) Spatial error in geocoding physician location data from the AMA Physician Masterfile: implications for spatial accessibility analysis. *Spatial and spatio-temporal epidemiology,* 31-38.

M. Millman (1993) *Access to health care in America*: National Academies Press.

V. Mor, J. Zinn, J. Angelelli, J. M. Teno and S. C. Miller (2004) Driven to tiers: socioeconomic and racial disparities in the quality of nursing home care. *The Milbank Quarterly,* 227-256.

A. N. Ngui and A. Vanasse (2012) Assessing spatial accessibility to mental health facilities in an urban environment. *Spatial and spatio-temporal Epidemiology,* 195-203.

Y. J. Park and E. G. Martin (2018) Geographic Disparities in Access to Nursing Home Services: Assessing Fiscal Stress and Quality of Care. *Health services research,* 2932-2951.

M. Petticrew, S. Cummins, C. Ferrell, A. Findlay, C. Higgins, C. Hoy, A. Kearns and L. Sparks (2005) Natural experiments: an underused tool for public health? *Public health,* 751-757.

M. Rahman and A. D. Foster (2015) Racial segregation and quality of care disparity in US nursing homes. *Journal of health economics,* 1-16.

M. Rahman, D. C. Grabowski, P. L. Gozalo, K. S. Thomas and V. Mor (2014) Are dual eligibles admitted to poorer quality skilled nursing facilities? *Health services research,* 798-817.

B. P. O. N. Reddy, Stephen

O'Neill, Ciaran (2020) Developing composite indices of geographical access and need for nursing home care in Ireland using multiple criteria decision analysis. *HRB Open*.

L. R. Shugarman and J. A. Brown (2006) Nursing home selection: How do consumers choose? Volume I: Findings from focus groups of consumers and information intermediaries. *Nursing*.

S. Song (1996) Some tests of alternative accessibility measures: A population density approach. *Land Economics,* 474-482.

S. T. Syed, B. S. Gerber and L. K. Sharp (2013) Traveling towards disease: transportation barriers to health care access. *Journal of community health,* 976-993.

E. Talen (1997) The social equity of urban service distribution: An exploration of park access in Pueblo, Colorado, and Macon, Georgia. *Urban geography,* 521-541.

E. Talen and L. Anselin (1998) Assessing spatial equity: an evaluation of measures of accessibility to public playgrounds. *Environment and planning A,* 595-613.

R. Tamara Konetzka, D. C. Grabowski, M. C. Perraillon and R. M. Werner (2015) Nursing home 5-star rating system exacerbates disparities in quality, by payer source. *Health Affairs,* 819-827.

R. Tibshirani (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* 267-288.

United States Department of Agriculture - Economic Research Service (2013) Rural-Urban Continuum Codes.

K. T. Unroe, M. A. Greiner, C. Colón-Emeric, E. D. Peterson and L. H. Curtis (2012) Associations between published quality ratings of skilled nursing facilities and outcomes of Medicare beneficiaries with heart failure. *Journal of the American Medical Directors Association,* 188. e181-188. e186.

US Census Bureau (2018) 2018 US gazetteer files. US Census Bureau, Geography Division Washington, DC, (Available from https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html.)

US Census Bureau American Community Survey. (Available from http://factfinder.census.gov.)

A. N. White (1979) Accessibility and public facility location. *Economic Geography,* 18-35.

D.-H. Yang, R. Goerge and R. Mullner (2006) Comparing GIS-based methods of measuring spatial accessibility to health services. *Journal of medical systems,* 23-32.

Y. Yuan, C. Louis, H. Cabral, J. C. Schneider, C. M. Ryan and L. E. Kazis (2018) Socioeconomic and geographic disparities in accessing nursing homes with high star ratings. *Journal of the American Medical Directors Association,* 852-859. e852.

X. Zhao (2016) Competition, information, and quality: Evidence from nursing homes. *Journal of health economics,* 136-152.