

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

1

2 **Abstract**

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

13 *Keywords*

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

15

16 **1. Introduction**

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

20 Biddle 2017, Marmot and Bell 2012), and subsequently how these influence access (Joseph
21 and Phillips 1984, Millman 1993) which in turn affects healthcare utilisation (Haynes *et al.*
22 1999, Jones *et al.* 2008) and hence health outcomes (Astell-Burt *et al.* 2011, Kirby *et al.* 2017,
23 Yang *et al.* 2006). In examining such relationships, several authors have employed a simplified
24 dichotomous approach ('financial vs other' (Joseph and Phillips 1984, Millman 1993) or
25 'spatial vs aspatial' (Khan 1992)). While parsimonious, these fail to capture the complex inter-
26 relationship between determinants. These relationships, alongside the impracticalities of
27 performing experimental approaches in such circumstances (Petticrew *et al.* 2005), can make
28 it difficult to isolate the underlying causal mechanisms by which they influence health. As a
29 result, advances typically rest upon the gradual accumulation of further evidence from natural
30 experiments and subsequently require careful interpretation (Kelly *et al.* 2010). As large
31 datasets for analysis become increasingly available, it can be more and more difficult to identify
32 which variables should be investigated as predictive (if not causative) of health outcomes.
33 Machine learning allows a principled and systematic approach to such variable selection.

34 Spatial accessibility is of particular importance for nursing homes as location is the most
35 frequently cited factor in the choice of home by residents (Shugarman and Brown 2006).
36 McIntyre *et al.* (2009) group determinants into three broad categories of accessibility:
37 availability (whether appropriate services are available where and when they are needed),
38 affordability (the ability to pay and consideration of the opportunity costs of doing so) and
39 acceptability (cultural perspectives/conditions that empower patients to use services and to 'fit'

40 with provider attitudes). These categories can interact with and through each other, better
41 capturing complex endogenous relationships and permitting a fuller picture of the connections
42 between disadvantage, access and outcomes to be constructed.

43 Previous studies have shown that overall care quality has measurable impacts on nursing home
44 (NH) residents' health (Cornell *et al.* 2019, Unroe *et al.* 2012). Nursing homes located in areas
45 with high levels of poverty (Park and Martin 2018) or in rural areas (Bowblis *et al.* 2013, Yuan
46 *et al.* 2018) appear to provide statistically worse care, as do NHs that have a high composition
47 of Medicaid-funded patients (Mor *et al.* 2004), due to increased fiscal stress (Park and Martin
48 2018). Patients simultaneously eligible for both Medicare and Medicaid are discharged from
49 hospital to NHs with poorer care quality than those eligible for Medicare alone (Rahman *et al.*
50 2014). Minorities receive statistically worse care (Allsworth *et al.* 2005, Bliss *et al.* 2015), in
51 homes which are often highly segregated by race (Mor, Zinn, Angelelli, Teno and Miller 2004).
52 In fact, residents seem to seek out homes with a majority of their own race, even at the cost of
53 proximity and care quality (Rahman and Foster 2015). Residents of socioeconomically
54 disadvantaged counties (Yuan, Louis, Cabral, Schneider, Ryan and Kazis 2018) and poor
55 neighbourhoods (Tamara Konetzka *et al.* 2015) have greater difficulties in accessing high-
56 quality nursing homes.

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

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

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

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

81

82 **2. Methods**

83

84 *2.1 Data*

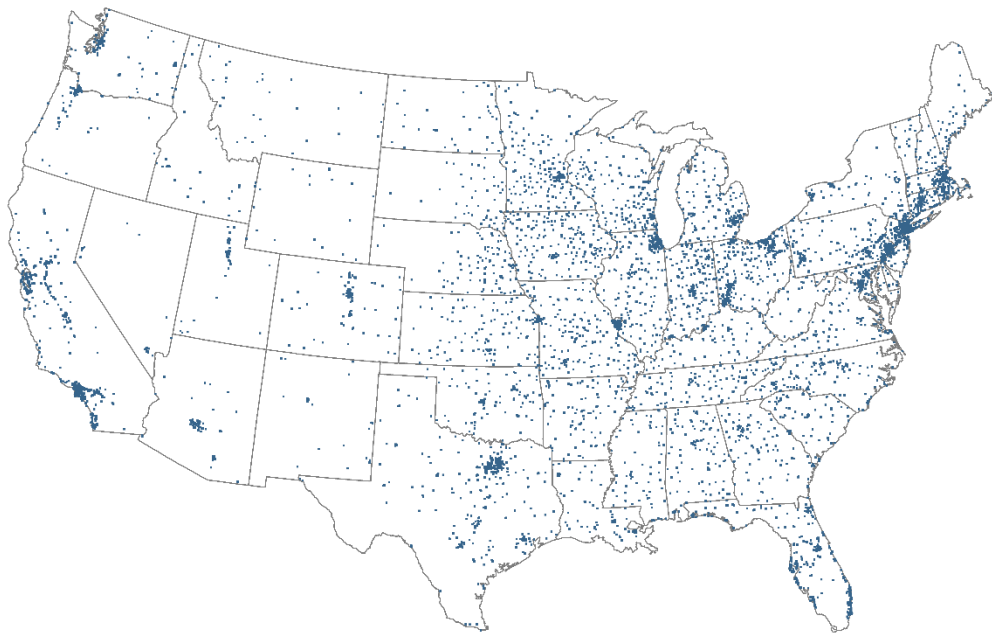
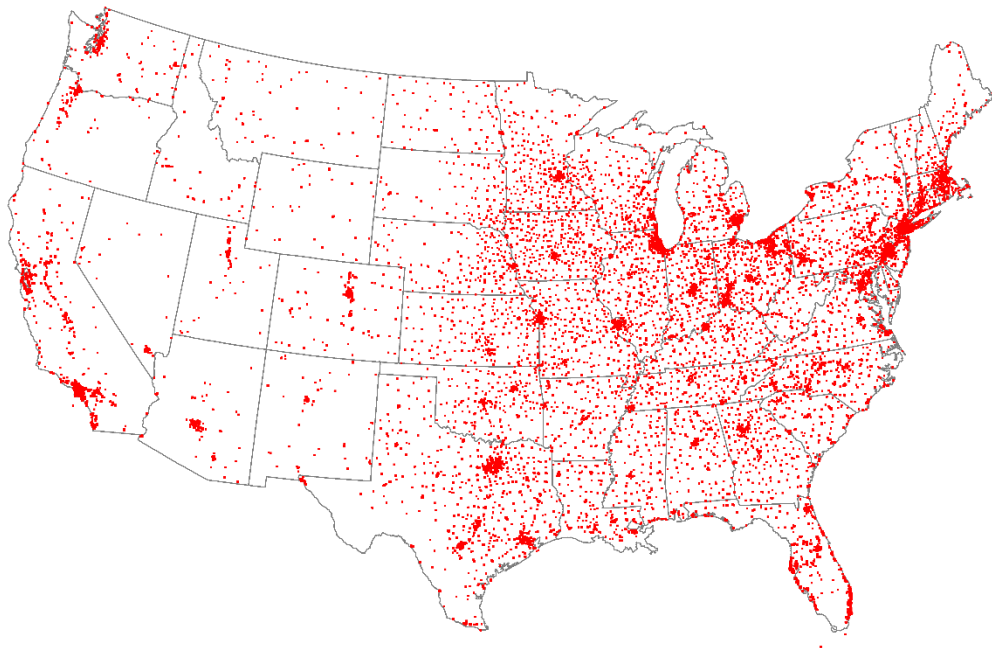
85 The US Centers for Medicare and Medicaid Services (CMS) publish “Nursing Home
86 Compare” (NHC) star ratings for all Medicare/Medicaid certified nursing homes in the US,
87 alongside all underlying data used to construct these ratings. NHs are given a score of one (low)
88 to five (high) stars based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

98 15,215 NHs were classified as high-quality in 2016. Figure 1 shows the locations of all geocoded
99 homes and, separately, those that were classified as high quality.

100

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

102



103

104 County-level estimates of socio-demographic data were available from the American
105 Community Survey (ACS) (US Census Bureau) over a five-year window (2012-2016). The
106 datasets included were “Comparative economic characteristics” (cp03), “ACS demographic
107 and housing estimates” (dp05), “Age and sex” (s0101) and “Households and families” (s1101).
108 Aside from the ACS and NHC data, a small number of further variables were incorporated or
109 derived for use in analyses: the counties’ population density, which we calculated based upon
110 each county’s total population and county land area (obtained from the US Census gazetteer
111 files (US Census Bureau 2018)); dummy variables representing which state the county is in;
112 and counties’ Rural-Urban Continuum Codes (RUCC), as published by the US Department of
113 Agriculture (2013). These describes how metropolitan a county is as one of nine categories; a
114 binary urban/rural variable was also included that we derived based upon it (as in Yuan et al.
115 (2018)).

116

117 2.2 *Accessibility measures*

118 For each county i , we calculated the population-weighted system-wide spatial accessibility to
119 high-quality nursing home care based upon a “gravity potential” model (Talen and Anselin
120 1998). Such gravity models have been shown to be the most sensitive techniques for explaining
121 population access to services (Song 1996). This accessibility measure took into account nearby
122 NHs located across county boundaries, while ensuring that far away home NHs are given

123 negligible weight. The model was similar to that used by Kalogirou and Foley (2006), whereby
124 the spatial accessibility of each county i is calculated by

$$125 \quad SA_i = \sum_j [(n_j / (p_i)) (1 / \max(d_{ij}^2, 1))]]$$

126 Where:

- 127 • n_j is the number of beds in each ‘high-quality’ nursing home j ;
- 128 • p_i is the population over 65 in county i (taken from the ACS “DP05” dataset); and
- 129 • d_{ij} is the distance in kilometres between the centroid of county i to nursing home j .

130 Distances less than 1km were set to this level to avoid attaching disproportionate weight
131 to NHs close to the county centroid (resulting from calculating $1/d_{ij}^2$).

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

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

138

139 2.3 *Descriptive analyses*

140 An exploratory model was developed to identify the associations between measures of
141 disadvantage and access to good quality care. Variables relating to the total/male/female

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

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

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

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

187 Potential predictors in lasso are typically centred and scaled – normalised with reference to the
188 variable’s mean and standard deviation, making them essentially z-scores. This was carried out
189 for all variables, including the accessibility measure. The target variable was therefore finally
190 defined as $\text{scale}(\ln(\text{SA}))$, which indicated each county’s accessibility compared to the national
191 average. Data for 2,281 randomly chosen counties (75% of the data) were used for estimation
192 (the ‘training’ data) and data for the remaining 25% of counties were used to validate the
193 predictive models’ out of sample performance (the ‘test’ data). Analyses were carried out using
194 the post-lasso approach in the *hdm* R package (selecting the “double selection” method in
195 *hdm*’s associated *rlassoEffect* function). This approach uses a data driven, non-arbitrary penalty
196 function (Bach *et al.* 2018) and creates a more easily interpretable model both by further
197 reducing the number of variables present and by allowing for the inference of post-selection
198 confidence interval, providing further interpretability. Without the post-lasso approach,
199 confidence intervals cannot be reliably estimated due to the introduced bias from variable
200 exclusion (Bach, Chernozhukov and Spindler 2018). Using imputed lasso and post-lasso

201 together should in principle therefore reduce the potential for bias in a number of ways, and to
202 our knowledge, this is the first time such an “imputed post-lasso” has been used.

203

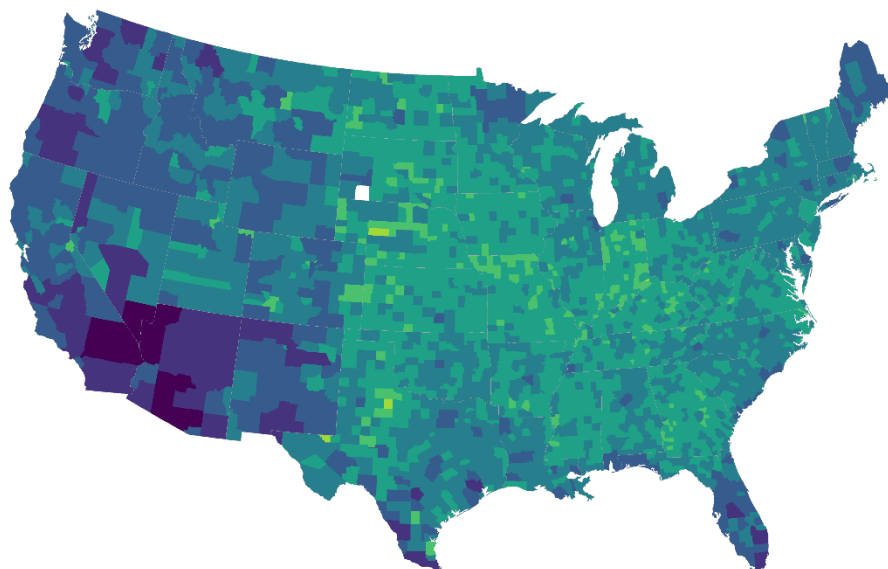
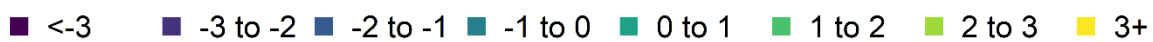
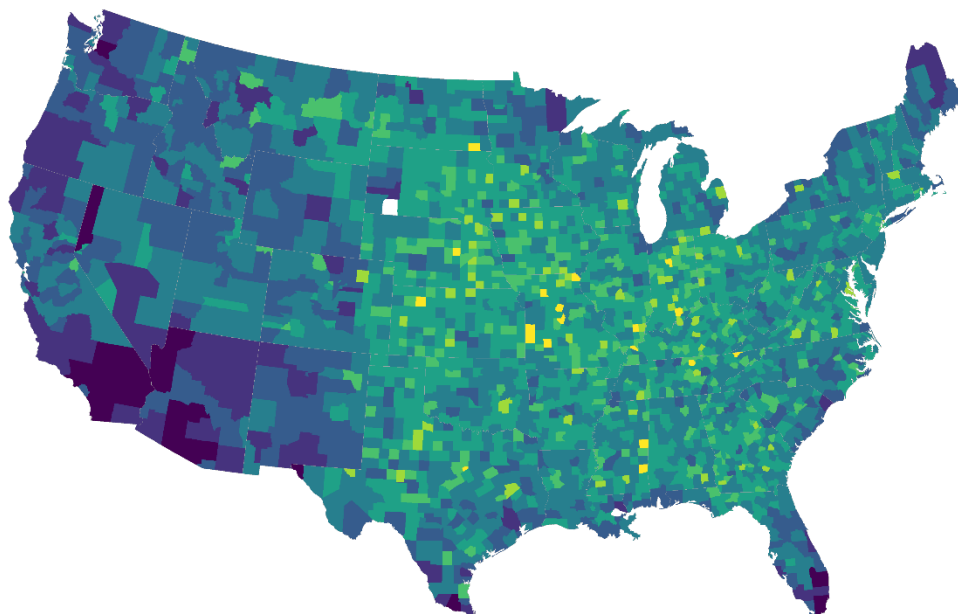
204 **3. Results**

205

206 *3.1 Geographical accessibility*

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

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



216 3.2 *Predictive analyses*

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

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

231

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

234

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

235

236 The coefficients associated with the exploratory predictive model reported in Figure 2 are

237 shown in Table 2. Where possible, these are grouped by general headings, such as the

238 percentage of the civilian population over 16 employed in a given industry, sex/age and race.

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

241
242

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

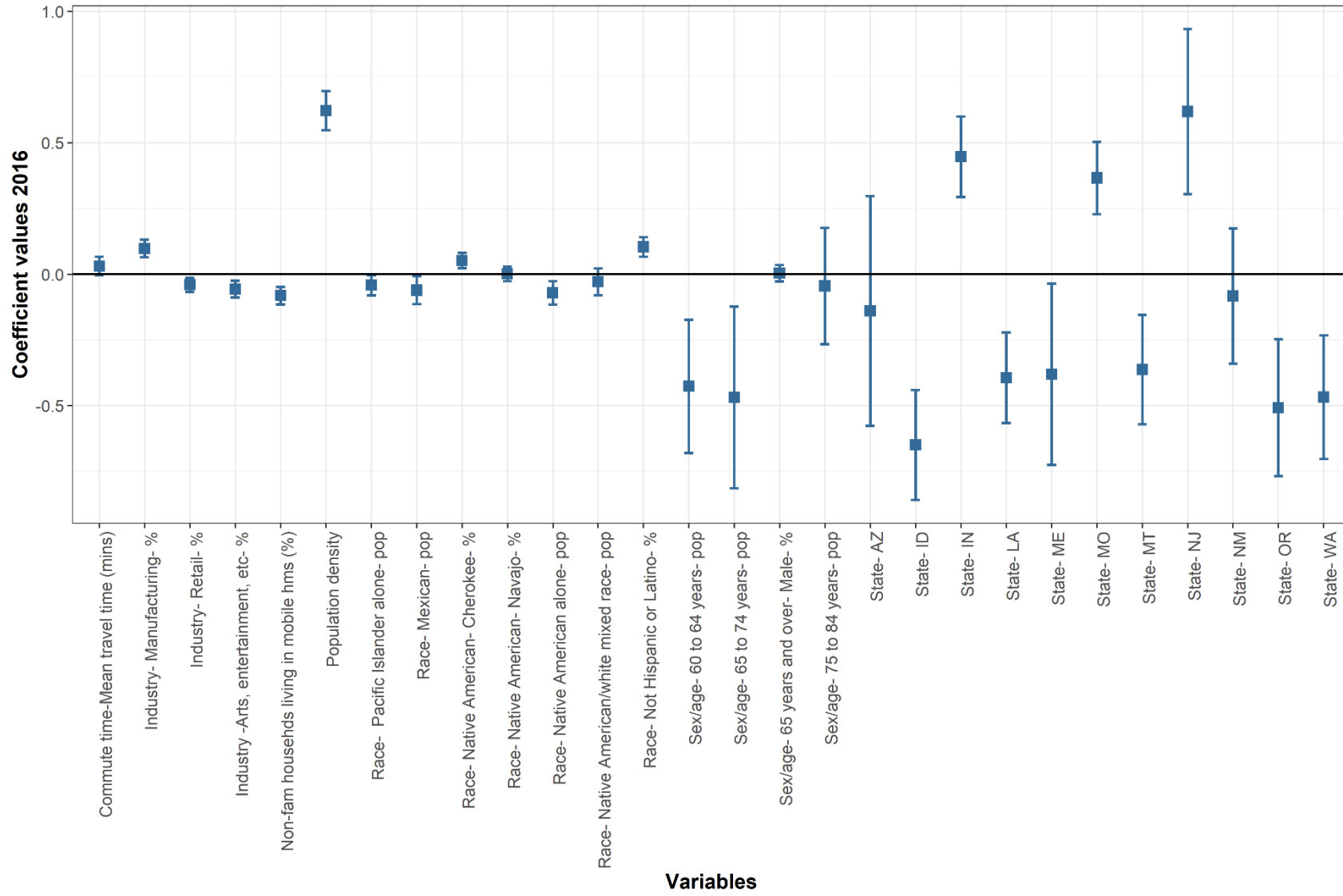
243

244 The lasso was successful in generating a relatively parsimonious model that substantially

245 reduced the number of variables included. 28 of the 472 possible variables were selected, of

246 which 11 were state dummies; other states were considered close enough to the average to not
247 merit inclusion in their own right. The model appears to correspond reasonably well to the
248 actual data, as seen in Figure 2. Its adjusted r-square for the test dataset was 0.613 (for the
249 training set it was 0.597) and the corresponding mean absolute error for the test dataset was
250 0.451 (0.444 for the training set). Estimation of confidence intervals around the coefficients
251 are shown in both Table 2 and Figure 4. Except for state dummy variables, confidence intervals
252 are generally narrow.

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



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

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

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

273 The results for a corresponding analysis of accessibility in 2011, including a predictive analysis
274 using the same imputed post-lasso approach (using ACS data from 2007-2011), were largely

275 consistent with the 2016 findings reported in this paper. An overview of these results is shown
276 in Appendix 3.

277

278 **4. Discussion**

279

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

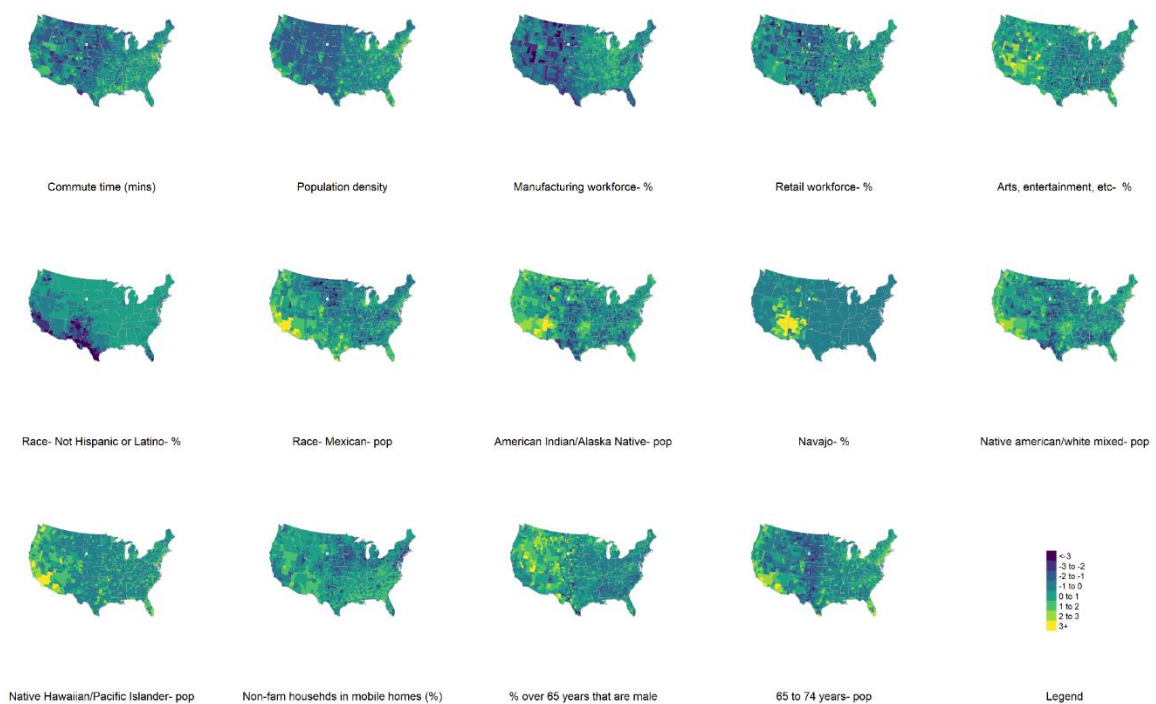
287 That population density is strongly positively associated with accessibility to high-quality NHs
288 may reflect simple economics – where demand is high because the population is concentrated,
289 so too will be supply of NHs in general (including good quality NHs). Nonetheless, since high-
290 quality nursing homes tend to cluster in wealthy areas (Tamara Konetzka, Grabowski,
291 Perrailon and Werner 2015), this picture may be more nuanced. It may also reflect previous
292 findings that homes in urban counties have higher star ratings (Lutfiyya, Gessert and Lipsky
293 2013), and hence counties with a high population density might be expected to be closer to

294 high quality homes. Table 1 showed that county population densities are in fact negatively
295 associated with accessibility on a univariate level.

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

308
309

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



310

311 While measures of average income were not selected by the model, it is notable that wealth-
312 related indicators such as the numbers living in mobile homes was a significant determinant of
313 access (especially since no variables included in the model directly related to wealth). In our
314 dataset, non-family mobile home residency was highly negatively correlated (<-0.6) with mean
315 family income and the proportion of women employed in the labour force, and similarly
316 positively associated with the proportion of families below the poverty level and proportions
317 of the population not in employment. It may therefore be included by the model as a composite
318 proxy for several wealth and income related factors. While wealth and income are obviously
319 related, privately-funded residents of NHs (which are strongly related to nursing home quality
320 (Grabowski 2001, Park and Martin 2018)) are likely to rely on previously accumulated wealth

321 rather than current earnings. That areas where there has been little accumulation of real estate
322 equity and relatively high levels of poverty are less likely to present attractive areas for care
323 homes to invest is unsurprising.

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

336 Spatial accessibility is not the only form of accessibility, however, and only part of McIntyre’s
337 (2009) “availability”. It is possible that areas with high accessibility by our measure may have
338 yet have low levels of accessibility depending on other factors (Ngui and Vanasse 2012), and
339 the results further highlight the interconnectedness between availability, affordability and
340 accessibility.

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

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

359

360 4.1 *Limitations*

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

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

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

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

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

399

400 4.2 *Suggestions for further research*

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

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

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

413

414 **5. Conclusions and implications**

415

416 This is the first paper to our knowledge that addresses county-level spatial accessibility to high-
417 quality nursing home care. We provide a formal definition of accessibility and calculated this
418 metric for each county in the contiguous USA. We thereafter investigated racial, socio-

419 demographic and economic factors' associations with accessibility, using an innovative
420 application of machine learning techniques. The paper illustrates that such a machine learning
421 approach can be used to cast a wide net and select the most important such variables, while
422 creating a parsimonious model that describes spatial accessibility. This approach could thereby
423 help identify heretofore unknown areas for targeted follow-up analyses, such as better
424 understanding whether specific barriers to access exist for Pacific Island populations.

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

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

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

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

448 **References**

- 449 J. E. Allsworth, R. Toppa, N. C. Palin and K. L. Lapane (2005) Racial and ethnic disparities in the
450 pharmacologic management of diabetes mellitus among long-term care facility residents. *Ethn*
451 *Dis*, 205-212.
- 452 T. A. Arcury, J. S. Preisser, W. M. Gesler and J. M. Powers (2005) Access to transportation and health
453 care utilization in a rural region. *The Journal of Rural Health*, 31-38.
- 454 T. Astell-Burt, R. Flowerdew, P. J. Boyle and J. F. Dillon (2011) Does geographic access to primary
455 healthcare influence the detection of hepatitis C? *Social Science & Medicine*, 1472-1481.
- 456 P. Bach, V. Chernozhukov and M. Spindler (2018) Valid Simultaneous Inference in High-Dimensional
457 Settings (with the hdm package for R). *arXiv preprint arXiv:1809.04951*.
- 458 D. Z. Bliss, O. Gurvich, K. Savik, L. E. Eberly, S. Harms, C. Mueller, J. F. Wyman, J. Garrard and B. Virnig
459 (2015) Are there racial-ethnic disparities in time to pressure ulcer development and pressure
460 ulcer treatment in older adults after nursing home admission? *Journal of aging and health*, 571-
461 593.
- 462 J. R. Bowblis, H. Meng and K. Hyer (2013) The Urban-Rural Disparity in Nursing Home Quality Indicators:
463 The Case of Facility-Acquired Contractures. *Health Services Research*, 47-69.
- 464 L. Breiman (2001) Random forests. *Machine learning*, 5-32.
- 465 S. C. Reed and S. Andes (2001) Supply and segregation of nursing home beds in Chicago communities.
466 *Ethnicity & health*, 35-40.
- 467 P. Y. Cornell, D. C. Grabowski, E. C. Norton and M. Rahman (2019) Do report cards predict future quality?
468 The case of skilled nursing facilities. *Journal of Health Economics*.
- 469 J. A. Davis (2005) Differences in the health care needs and service utilization of women in nursing
470 homes: Comparison by race/ethnicity. *Journal of women & aging*, 57-71.
- 471 P. L. Delamater, J. P. Messina, A. M. Shortridge and S. C. Grady (2012) Measuring geographic access to
472 health care: raster and network-based methods. *International journal of health geographics*,
473 15.
- 474 J. Edward and D. J. Biddle (2017) Using geographic information systems (GIS) to examine barriers to
475 healthcare access for Hispanic and Latino immigrants in the US south. *Journal of racial and*
476 *ethnic health disparities*, 297-307.
- 477 D. C. Grabowski (2001) Medicaid reimbursement and the quality of nursing home care. *Journal of health*
478 *economics*, 549-569.
- 479 J. T. Hart (1971) The inverse care law. *The Lancet*, 405-412.
- 480 R. Haynes, G. Bentham, A. Lovett and S. Gale (1999) Effects of distances to hospital and GP surgery on
481 hospital inpatient episodes, controlling for needs and provision. *Social science & medicine*, 425-
482 433.
- 483 D. S. Houston (2005) Methods to test the spatial mismatch hypothesis. *Economic Geography*, 407-434.
- 484 D. Johnson, R. Thorpe, J. McGrath, W. Jackson and C. Jackson (2018) Black-White Differences in Housing
485 Type and Sleep Duration as Well as Sleep Difficulties in the United States. *International journal*
486 *of environmental research and public health*, 564.
- 487 A. Jones, R. Haynes, V. Sauerzapf, S. Crawford, H. Zhao and D. Forman (2008) Travel times to health
488 care and survival from cancers in Northern England. *European journal of cancer*, 269-274.
- 489 A. E. Joseph and D. R. Phillips (1984) *Accessibility and utilization: geographical perspectives on health*
490 *care delivery*: Sage.
- 491 S. Kalogirou and R. Foley (2006) Health, Place and Hanly: modelling accessibility to hospitals in Ireland.
492 *Irish Geography*, 52-68.
- 493 M. Kelly, A. Morgan, S. Ellis, T. Younger, J. Huntley and C. Swann (2010) Evidence based public health:
494 a review of the experience of the National Institute of Health and Clinical Excellence (NICE) of
495 developing public health guidance in England. *Social science & medicine*, 1056-1062.
- 496 A. A. Khan (1992) An integrated approach to measuring potential spatial access to health care services.
497 *Socio-economic planning sciences*, 275-287.

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

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

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

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

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

507 D. McIntyre, M. Thiede and S. Birch (2009) Access as a policy-relevant concept in low-and middle-
508 income countries. *Health economics, policy and law*, 179-193.

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

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

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

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

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

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

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

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

525 B. P. O. N. Reddy, Stephen

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

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

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

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

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

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

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

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

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

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

547 US Census Bureau (2018) 2018 US gazetteer files. US Census Bureau, Geography Division Washington,
548 DC, (Available from [https://www.census.gov/geographies/reference-files/time-](https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html)
549 [series/geo/gazetteer-files.html](https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html).)
550 US Census Bureau American Community Survey. (Available from <http://factfinder.census.gov>.)
551 A. N. White (1979) Accessibility and public facility location. *Economic Geography*, 18-35.
552 D.-H. Yang, R. Goerge and R. Mullner (2006) Comparing GIS-based methods of measuring spatial
553 accessibility to health services. *Journal of medical systems*, 23-32.
554 Y. Yuan, C. Louis, H. Cabral, J. C. Schneider, C. M. Ryan and L. E. Kazis (2018) Socioeconomic and
555 geographic disparities in accessing nursing homes with high star ratings. *Journal of the*
556 *American Medical Directors Association*, 852-859. e852.
557 X. Zhao (2016) Competition, information, and quality: Evidence from nursing homes. *Journal of health*
558 *economics*, 136-152.
559