Global variation in travel time to the nearest hospital

Simon Yadgir

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Public Health

University of Washington

2019

Committee:

Greg Roth

Robert Reiner

Program Authorized to Offer Degree:

Department of Global Health

University of Washington

Abstract

Global variation in travel time to the nearest hospital

Simon Yadgir

Chair of the Supervisory Committee:

Greg Roth

Department of Global Health

The time it takes to travel to the nearest hospital is an important measure of access to healthcare services, particularly for acute illnesses that require immediate medical attention. Despite the public health significance of this measure, it has never been quantified in a comparable way at a high resolution for the globe. We used a combination of Ministry of Health and open source data on the locations of hospitals to estimate the travel time to the nearest hospital at the 5x5 kilometer level for inhabited areas of the world. We used machine learning algorithms and spatial simulation models to estimate, with uncertainty, the locations of hospitals in each country, and applied a recently published friction map to calculate travel time to the nearest hospital from each 5x5 kilometer raster. We found that 85% (84% to 86%) of all persons were living within 90 minutes of a hospital. We also found large variation in the proportion of individuals living within 90 minutes of a hospital across countries, with most high-income countries having more than 95% of individuals living within 90 minutes of a hospital. We also found a large difference in the mean travel time to the nearest hospital by urbanicity, being 13 (11 to 18) minutes among urban rasters and 82 (78 to 90) minutes among rural rasters.

Contents

Definitions 4
Introduction 5
Methods 5
Data 5
Analysis 7
Predicting hospital locations 7
Simulation 8
Calculating travel times 9
Cross-validation 10
Results 10
Global results 10
Cross validation results 13
Discussion 15
Limitations 16
Conclusion 17
References 17

Definitions

AUROC: Area under the receiver-operator

BRA: Brazil

$c_{90,i}$: The probability that raster i is within 90 minutes of a hospital

CCH: Cell(s) containing a hospital

CLL: Complementary log-log transformation

ETH: Ethiopia

GBM: Gradient boosted machine

Km: Kilometer

ML: Machine learning

MoH: Ministry of Health

p_{90}: The proportion of individuals living within 90 minutes of a hospital for a given location

PER: Peru

RF: Random forest

SSA: Sub-Sarahan Africa

UK: United Kingdom

USA: United States of America

Introduction

Acute hospital care is essential for prevention of mortality due to many common causes of death, including acute myocardial infarction. Despite the importance of acute hospital care, access to these life-saving services varies significantly both across and within countries. One of the major reasons for this disparity is the lack of physical access to these services.[1,2] Physical access to care is defined as a health system's ability to physically reach individuals in need of a specific intervention. For acute illness requiring immediate medical attention, physical access is primarily driven by the amount of time it takes for an individual to arrive at a care facility. Although travel time to hospital is only one component of healthcare access and quality,[3] it is particularly important for diseases that require services only available at a hospital. Additionally, acute diseases such as myocardial infarction often require immediate medical attention, and delays in care are associated with worse outcomes, including death.[4,5]

Previous attempts to measure time to reach acute care facilities have been focused within one or several countries,[1,3,6] and no such measure exists for most locations. Measuring time to acute care is important for identifying areas where access to care is low, as well as comparing performance of different health systems in terms of the availability of acute care. Country-level estimates of access hide drastic within-country differences.[7,8] For this reason, comparable and consistent geospatial estimates of access are necessary for public health policymakers. Identifying disparities in access to care at both the country and geospatial level is a crucial step in allocating spending of health resources.

Using a combination of government reported hospital data and open source data, we estimated the travel time to an acute hospital care for all locations at the 5x5 kilometer level. Population-weighted country-level measures of access were also generated, including the proportion of individuals in a location living within ninety minutes of a hospital. Simulation was used to generate 95% uncertainty intervals.

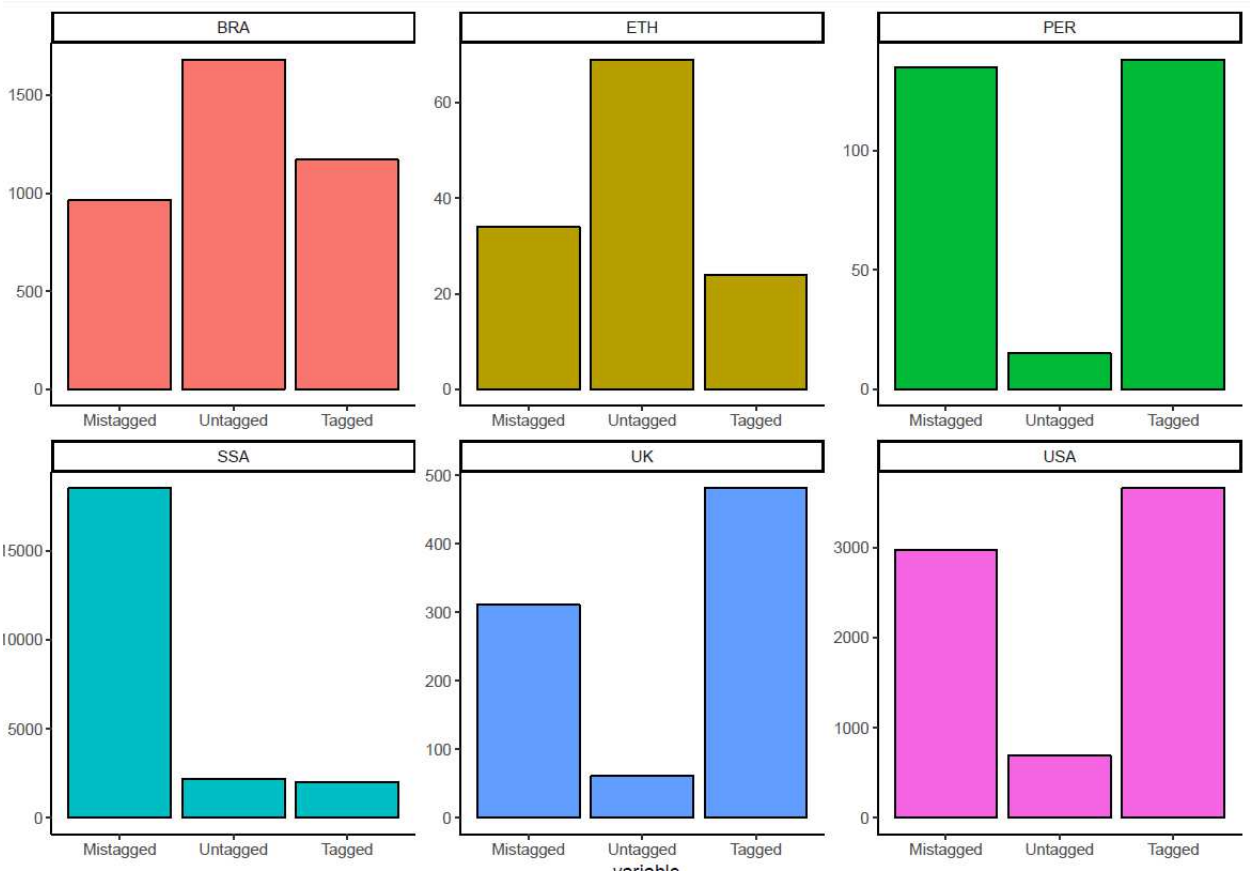
Methods

Data

We collected data reported by national governments describing the latitude and longitude of each hospital within a given country. This data was identified by a convenience search through Ministry of Health websites, and via the Global Burden of Disease collaborator network. Ministries of Health provided hospital locations from 2015 to 2017. In Sub-Saharan African countries, we used the public sector hospital locations reported by Ouma et al.[1] For the purpose of this project, we assumed these databases to reflect the true distribution of hospitals in these locations for the year 2017. We dropped hospitals that did not provide acute in-patient care for adults, such as outpatient-only facilities, psychiatric-only facilities, or pediatric-only facilities.

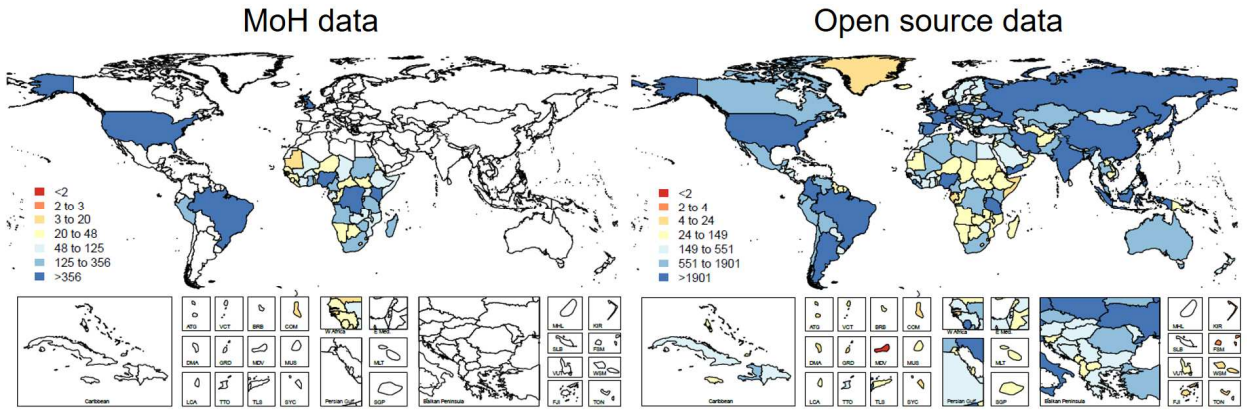
In addition to hospital data reported from the Ministries of Health, we used a collection of geolocated health sites from an open source database.[9] These data are collected through open source geotagging projects, and may or may not represent a true hospital. Additionally, use of open source software likely varies by location, so many hospitals are likely untagged. Figure 1 shows discrepancies between MoH data and open source data where both are available. Mistagged hospitals are facilities tagged in the open source data that are not present in the MoH data. Untagged hospitals are facilities tagged in the MoH data that are not present in the open source data.

Figure 1: Mistagged, untagged, and correctly tagged Open source data points



Though these data are biased, they are available for every country, unlike Ministry of Health reported hospital location data. As described below, we use the MoH data in combination with the open source data to predict the locations of hospital for locations where MoH data was not collected. Figure 2 shows the locations where MoH and open source data were available, as well as the number of facilities reported in each dataset.

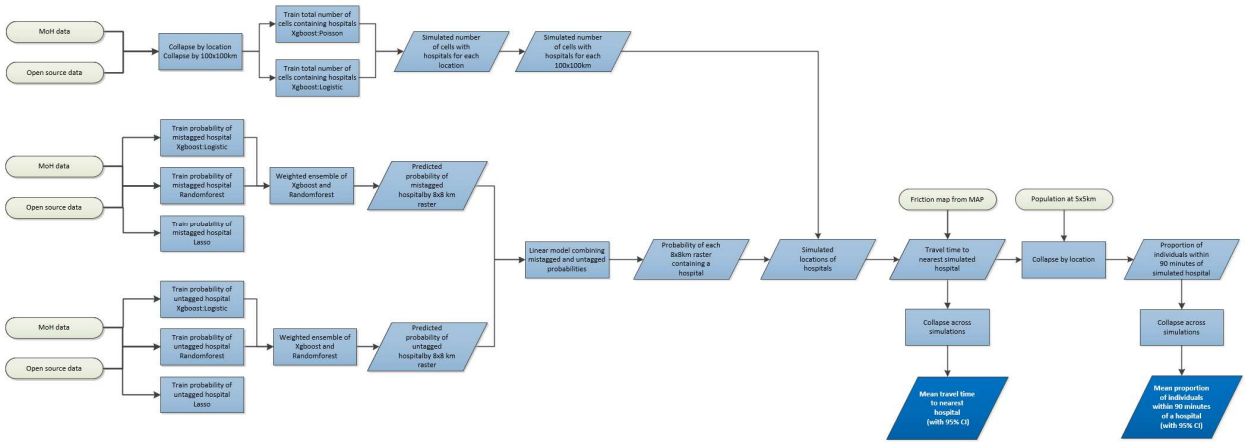
Figure 2: Number of health sites reported in MoH and Open source datasets



The Malaria Atlas Project's friction map was used to estimate travel time to the nearest facility. The methods of generating this map have been reported elsewhere. Briefly, the friction map represents the average time in minutes to travel one meter in a given 1x1 kilometer (km) raster on the globe, taking into account available modes of travel.[2] The friction map combines information on roads, terrain, slope, and other factors, and uses predictive algorithms based on the probable mode of transportation in order to calculate friction for each raster. This map provides the best estimation of time it takes to travel across any surface on the globe. All predictive models used to estimate hospital locations in countries without MoH data used geospatial covariates available in the Local Burden of Disease database at the Institute of Health Metrics and Evaluation.[10]

Analysis

Figure 3: Model diagram



Predicting hospital locations

We used a multi-step framework to estimate the mean travel time to a hospital from any inhabited area on the globe. For locations where MoH data was not available, we created a predictive model to estimate the probability that a given 8x8 km raster was a cell containing a hospital (CCH). This predictive model involved four stages, outlined below and in Figure 3. Each model was trained on locations where both MoH and open source data were available, and we used two machine learning (ML) algorithms: a random forest (RF) model using the randomForest package in R[11] and a gradient

boosted model (GBM) using the xgboost package in R.[12] Other algorithms (logistic regression, lasso regression) were also run, but dropped due to poor predictive performance (not shown).

First, the total number of 8x8 CCH within a location l containing a hospital was estimated. We used a RF, a GBM with a Poisson link, and a GBM with a logistic link. We simultaneously estimated the proportion of 8x8 km rasters that were CCH within each 300x300 km raster. l_r denotes the r^{th} 300x300 km raster in location l:

$$p_{l_r} = \frac{p_{l_r,pois} + p_{l_r,logist} + p_{l_r,rf}}{3}$$

Where p_{l_rpois}, $p_{l_r,logist}$, $p_{l_r,rf}$ are the predicted proportion of CCH from the Poisson GBM, logistic GBM, and RF models. This step informs within-country trends in hospital distribution. All spatial and location-level covariates were used in these models, as well as the number of 8x8 km rasters containing health facilities tagged in the open source data, and the number of 8x8 km rasters containing health facilities tagged in the open source data where a 'hospital' tag was included for at least one facility.

Next, we estimated the probability that each 8x8 km raster was a CCH among rasters containing a facility tagged in the open source data, accounting for mistagged hospitals. We separately estimated the probability that each 8x8 km raster was a CCH among rasters that did not contain a facility tagged in the open source data, accounting for untagged hospitals. The predictions for mistagged hospitals and untagged hospitals were averaged. RF predictions were given 2/3 weight, and GBM predictions were given 1/3 weight, based on the cross-validation performance described below. r_i denotes the i^{th} 8x8 km raster in 300x300 km raster r.

$$mistagged_{r_i} = \frac{2}{3} * mistagged_{r_i,rf} + \frac{1}{3} * mistagged_{r_i,gbm}$$

$$untagged_{r_i} = \frac{2}{3} * untagged_{r_i,rf} + \frac{1}{3} * untagged_{r_i,gbm}$$

Where $mistagged_{r_i}$ is the probability of being a mistagged CCH, and $untagged_{r_i}$ is the probability of being an untagged CCH. $mistagged_{r_i,rf}$ and $mistagged_{r_i,gbm}$ are the probabilities of cell i in raster r being mistagged CCH, predicted by the RF and GBM algorithms, respectively. $untagged_{r_i,rf}$ and $untagged_{r_i,gbm}$ are the probability being an untagged CCH predicted by the RF and GBM algorithms, respectively. Lastly, the probabilities that a cell was mistagged and untagged were combined with a simple linear regression model:

$$invlogit(p_{r_i}) = \beta_0 + \beta_1 CLL(1 - mistagged_{r_i}) + \beta_2 CLL(untagged_{r_i})$$

Where p_{r_i} is the probability that 8x8km raster i is a CCH, and CLL is the complementary log-log transformation. Combining the probabilities of being mistagged CCH and untagged CCH with this model allowed us to empirically weight each model.

Simulation

Simulation was used to incorporate uncertainty in our predictions of which rasters were CCH, and was done in three stages. The simulation process was repeated 40 times for each location. First, for

each location, the total number of 8x8 CCH was simulated, by averaging the predictions from each location level model:

$$n_{l,pois} \sim Pois(\lambda_{l,gbm})$$

$$n_{l,logist} \sim Binom(p_{l,gbm}, cells_l)$$

$$n_{l,rf} \sim Pois(\lambda_{l,rf})$$

$$n_l = \frac{n_{l,pois} + n_{l,logist} + n_{l,rf}}{3}$$

Where $\lambda_{l,gbm}$, $p_{l,gbm}$, $\lambda_{l,rf}$ are the fit parameters from the location-level Poisson GBM, logistic GBM, and RF models, respectively, for location l. n_l represents the total number of 8x8 km CCH in location l for a given simulation.

Second, we simulated the number of 8x8 km CCH for each 300x300 km raster. To ensure that the 300x300 km simulations were consistent with the simulations at the location level, we simulated from a modified multinomial distribution

$$\boldsymbol{p_l} = (p_{l_1}, p_{l_2}, \dots p_{l_R})$$

$$\boldsymbol{k_l} \sim Multinom^*(\boldsymbol{p_l}, n_l)$$

Where R is the number of 300x300 km rasters in location l, $\boldsymbol{p_l}$ is the vector of estimated proportions of 8x8 km CCH for each 300x300 km raster in location l, and $\boldsymbol{k_l}$ is the vector of simulated numbers of 8x8 km CCH within each 300x300km for location l for a given simulation. The multinomial distribution was modified to allow us to weight probabilities by the number of 8x8 km rasters contained in each 300x300 km raster, which differed due to irregular location shapes. This step allowed us to simulate the geographic distribution of 8x8 km CCH within a location using expected trends based on the machine learning models.

Last, we simulated whether or not each 8x8 km raster was a CCH:

$$\boldsymbol{p_r} = (p_{r_1}, p_{r_2}, \dots p_{r_I})$$

$$\boldsymbol{q_r} \sim Multinom(\boldsymbol{p_r}, k_r)$$

Where I is the number of 8x8 km rasters in 300x300 km raster r, $\boldsymbol{p_r}$ is the vector of predicted probabilities of being a CCH for each 8x8 km raster within 300x300km raster r for each simulation, and $\boldsymbol{q_r}$ is the vector of simulated of binary values whether or not each 8x8 km raster was a CCH.

This simulation schema ensured that the number of hospitals for each location, and that the distribution of those hospitals within each location, was a realistic number based on the MoH data observed. Simulation was done 40 times for prediction, and 100 times for out-of-sample cross-validation.

Calculating travel times

The shortest travel time to a hospital was calculated for each raster by combining the simulated hospitals with the friction map for each location. In order to preserve spatial granularity of the friction map, this step was done at the 1x1 km level. We used Dijkstra's algorithm to identify the shortest path

from the center of each 1x1 km cell to the center of the nearest simulated CCH, and accumulated the friction of that path to calculate the travel time for each raster. Distance and travel time calculations were done with the gdistance package in R.[13] Travel times were aggregated to 5x5 km rasters in order to match the granularity of the available urbanicity covariate, using the raster package in R. The proportion of individuals living within 90 minutes of a hospital, p_{90}, for each location was calculated by the following equation:

$$f(x) = \begin{cases} 0 & if\ x > 90, \\ 1 & otherwise \end{cases}$$

$$c_{90,i} = f(time_i)$$

$$p_{90} = \sum \left(\frac{c_{90,i} * pop_i}{\sum pop_i} \right)$$

Where $time_i$ is the estimated travel time for raster i, $c_{90,i}$ is whether or not raster i is within 90 minutes to a hospital, p_{90} is the proportion of individuals living within 90 minutes of a hospital, and pop_i is the population for raster i.

The mean, 2.5th and the 97.5th quantiles of the calculated travel times and p_{90} were reported for locations where no MoH data was available, otherwise only a mean was reported. Simulation was done by location, thus we assumed that individuals did not travel outside of the location for acute hospital care.

Cross-validation

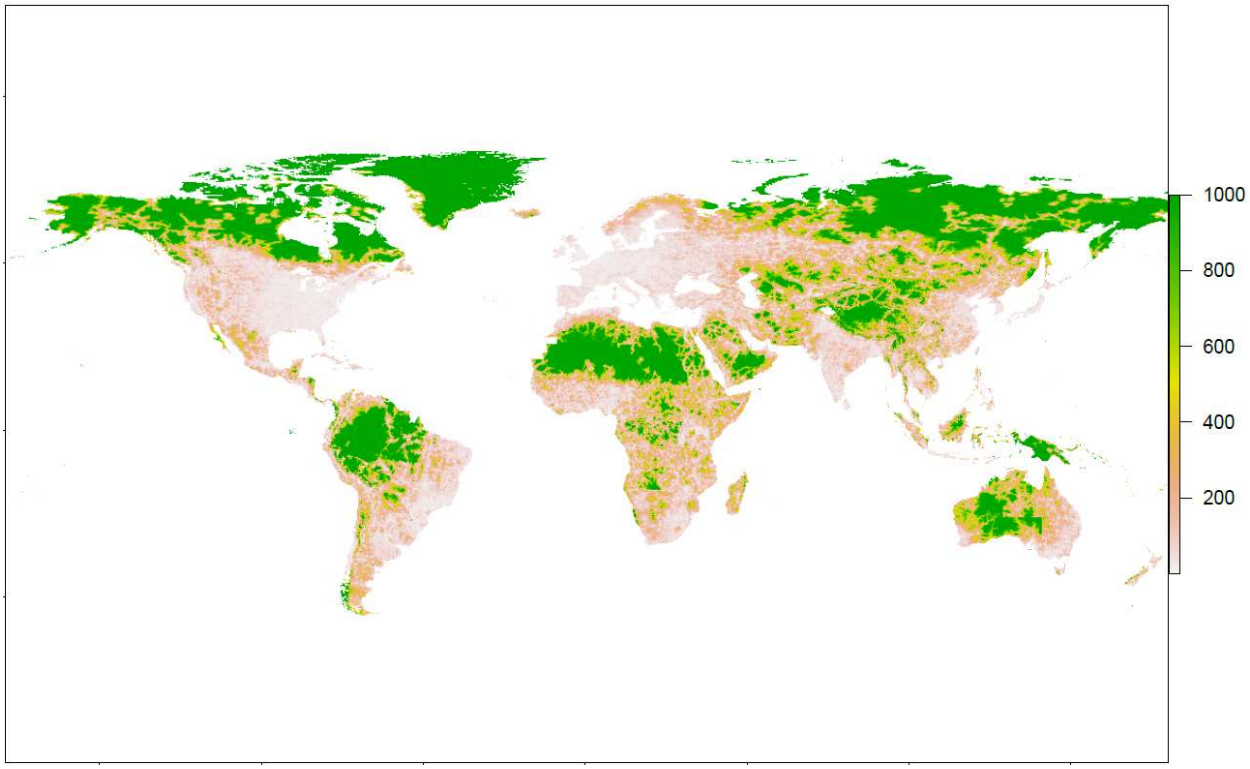
All models were cross-validated using a location-level knockout schema. For each location where both MoH and open source data was available, that location was held out and the entire pipeline was run on all other locations. Predictions of travel time and p_{90} were then made for the held-out location. These predictions were compared to the calculated travel time and p_{90} based on the MoH data, which we take to represent truth. This out-of-sample cross-validation allows us to approximate model performance for locations where no MoH data is available. Cross-validation metrics are discussed in the Results section.

Results

Global results

We found large variation in travel time to the nearest facility globally, with low travel times in Europe and High-income North America, and high travel times in sparsely populated areas of the world, as well as parts of Asia, South America, and Africa. Figure 4 maps the mean travel time for each 5x5 km raster.

Figure 4 Mean travel time (minutes)



Our results showed that 85% of the global population lived within 90 minutes of a hospital, representing 1.05 billion people living farther than 90 minutes from a hospital. Figure 5 shows the proportion of simulated draws where each raster was within 90 minutes of a hospital. We found significant variation in p_{90} across countries, ranging from 0.06 in the Cook Islands to 1 in Singapore. Among high-income locations, 82% of countries had a p_{90} greater than 95%. Countries in central Sub-Saharan Africa and the Middle East/Central Asia were most likely to have p_{90} less than 50%. Figure 6 maps p_{90} by country.

Figure 5 Proportion of simulations where each raster was within 90 minutes to a hospital

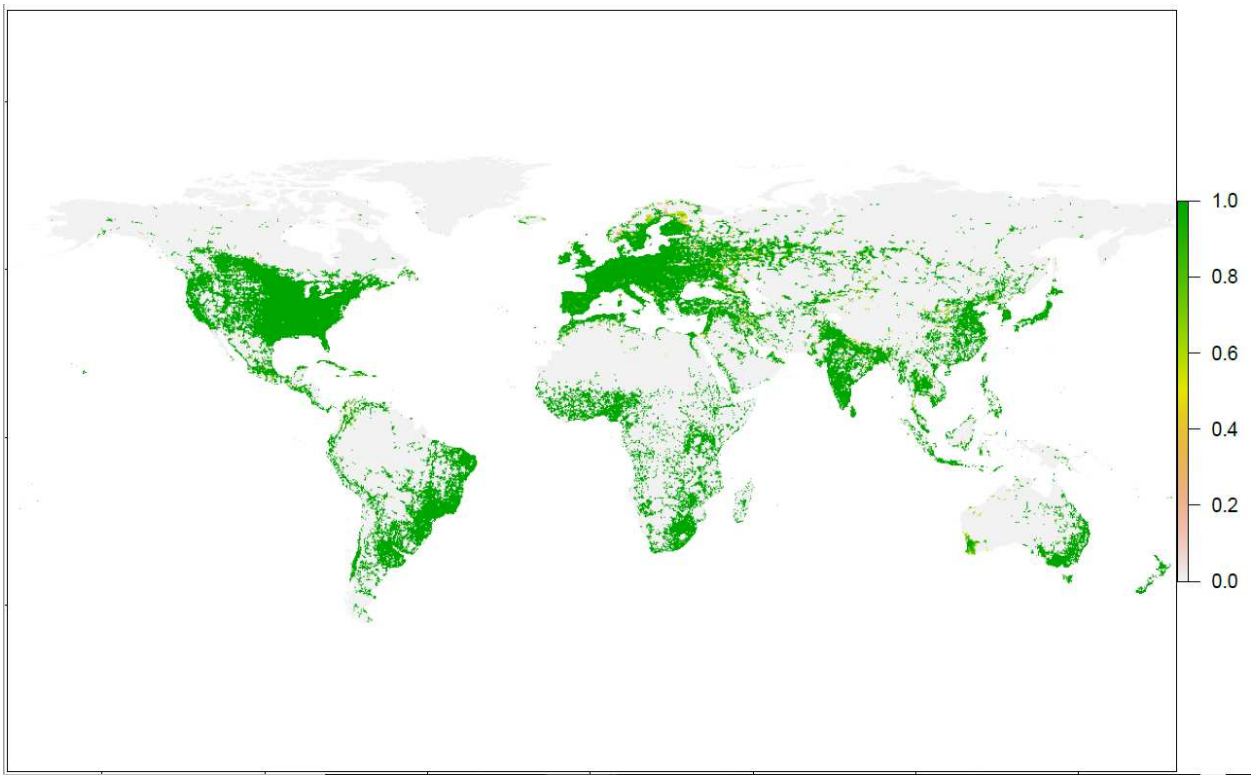
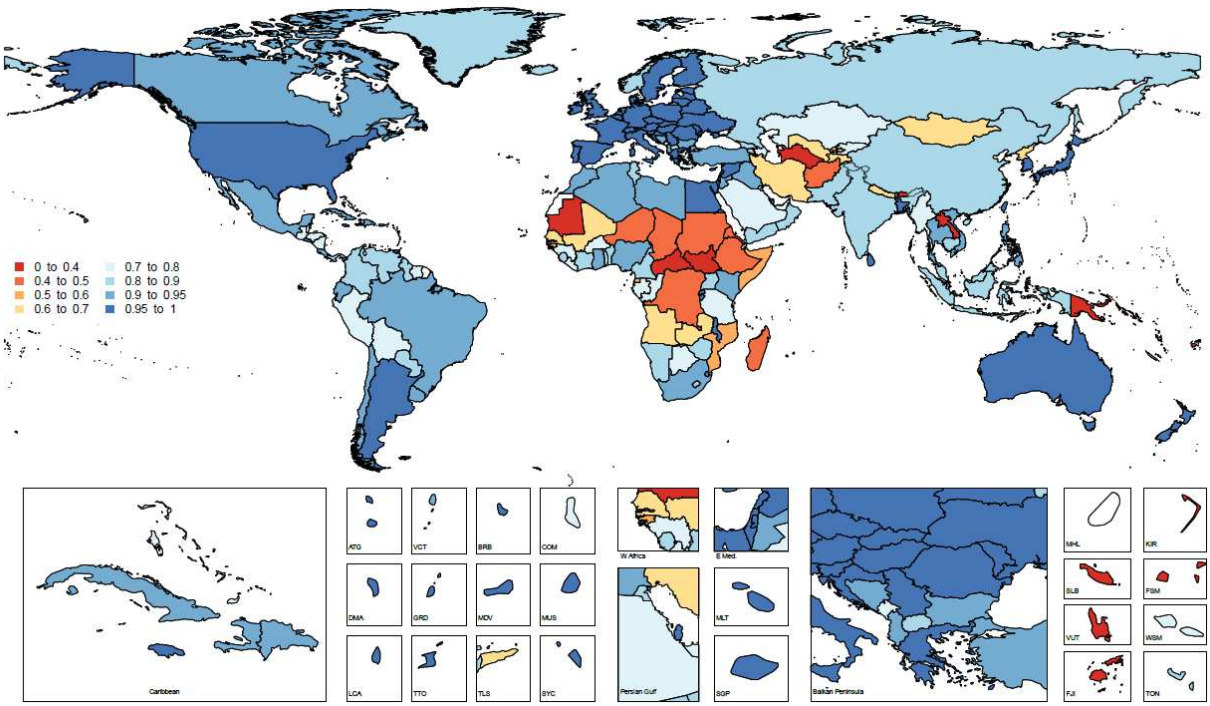


Figure 6 Proportion of individuals living within 90 minutes to a hospital



We estimated the mean population-weighted travel time to be 13 minutes for all urban locations and 82 minutes for all rural locations. Overall, 98% of individuals living in urban rasters lived within 90 minutes to a hospital, compared to 76% of individuals living in rural rasters.

Cross validation results

Our country-level cross-validation schema gave an overall mean absolute error in out-of-sample predicted travel time of 172 minutes. This high error was due to areas where both the true and predicted travel times were very high (>1000 minutes), but the difference was still large. After removing these rows, the mean absolute error was 35 minutes. This identified the inability of our model to precisely predict very high travel times, though our model could identify whether or not travel time was greater than 1000 minutes. We found the 95% coverage, defined as the proportion of 5x5km rasters that had the true travel time between the 95% credible interval of our estimate, was 85%. The mean absolute error and the 95% coverage are shown by dataset and urbanicity in Figure 7 and Figure 8, respectively. We found higher mean error in rural locations. We found worse 95% coverage in high-income locations. 90% of the absolute differences between the out of sample predicted travel time and the true travel time were less than 98 minutes, after removing rows where both the true and predicted travel times were greater than 1000 minutes.

Figure 7 Absolute out of sample error in travel time

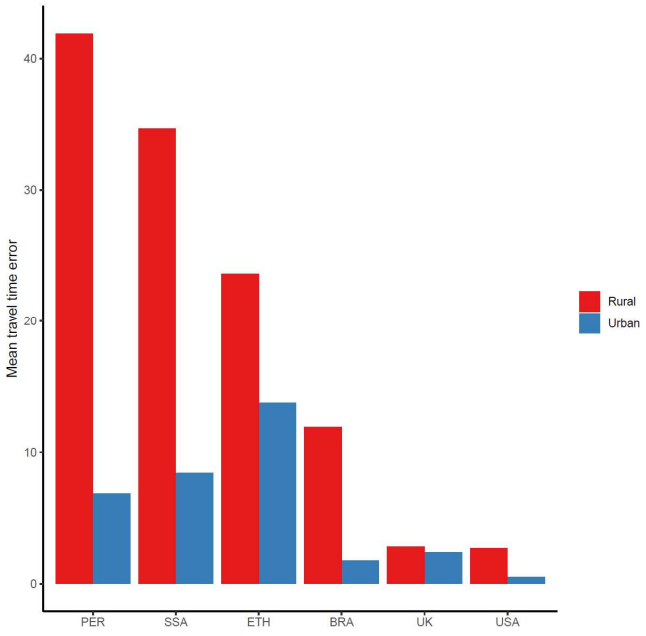
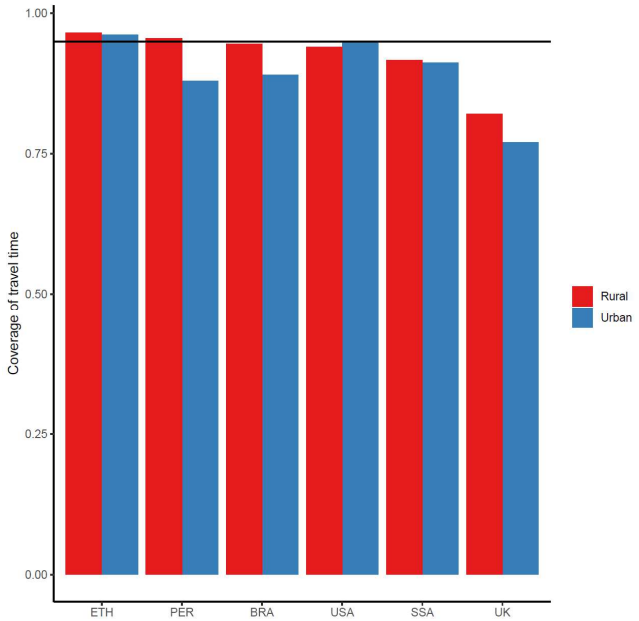
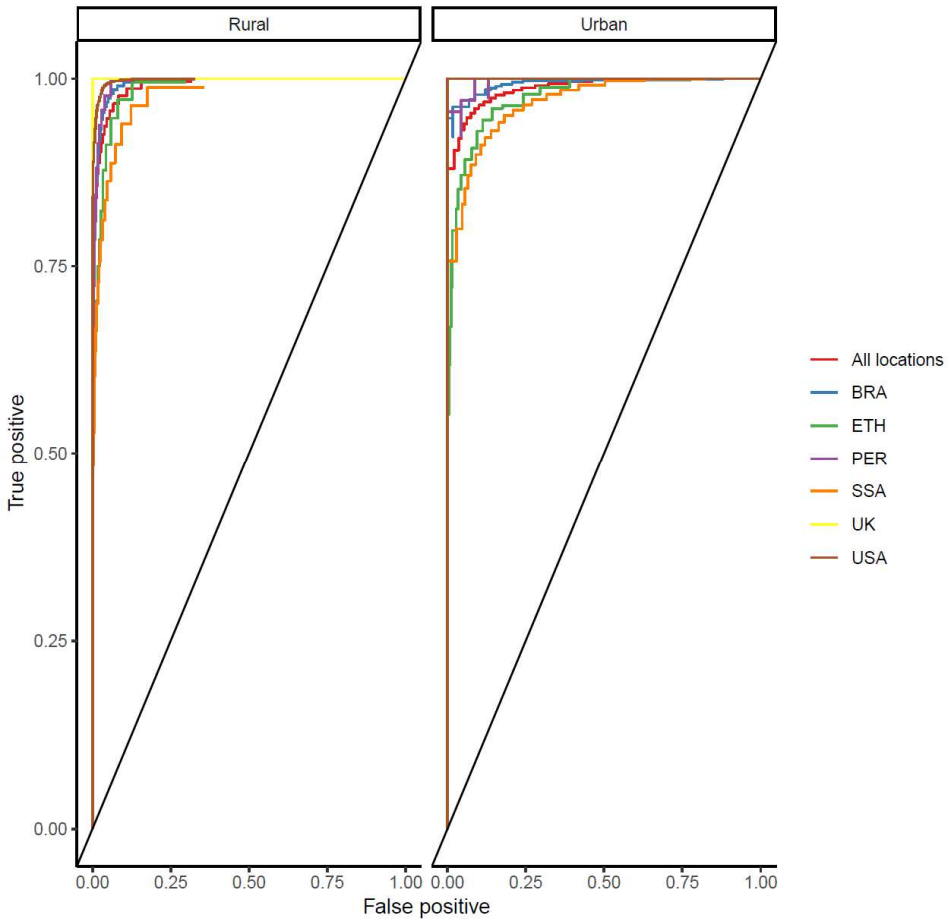


Figure 8 95% out of sample coverage of travel time predictions



We compared the out of sample predictions for $c_{90,i}$ to whether or not a 5x5km raster was truly within 90 minutes of a hospital. Using 0.5 as a cutoff, we had 0.88 specificity, and 0.97 sensitivity, representing an extremely high out of sample predictive success. Figure 9 shows the receiver-operator curves for the out of sample predictions of $c_{90,i}$. The area under the receiver-operator curve (AUROC) was 0.985 across all urban locations, and 0.984 across all rural locations.

Figure 9 Receiever-operator curves for c_(90,i), by dataset



Discussion

In estimating travel time to nearest facility, we provide a global geospatial measure of access to care for the globe. We found large variation in the proportion of individuals living within ninety minutes of a hospital (p_{90}) by location. In many low income countries, we found p_{90} to be lower than 50%, suggesting that improved access to existing health facilities or new health facilities are required to care for individuals with illnesses that require treatment at a hospital. This lack of access has implications for meeting universal health coverage goals.

We highlight high p_{90} in much of the world, with many countries containing more than 95% of it's population within ninety minutes of a hospital. Despite this high coverage, many of these health systems are likely failing to prevent burden of disease. As individual facilities vary greatly in the scope, capacity, and quality of care available, strengthening these areas for locations with high p_{90} could be a priority in addressing lack of physical access.[3] Addressing issues of socio-economic inequality is likely a strong driver of access to care.[14]

We also found large within-country variation in predicted travel time to the nearest hospital. There was large variation in population-weighted mean travel time by urbanicity. Although travel time is

just one component of physical access to care, these results identify major gaps in access in parts of Africa and Central. This disparity is exacerbated by lower access to specific interventions, hospital capacitiy, quality of care, and economic access in rural areas.[15,16] Future studies should explore how these crucial components of care vary across space.

Limitations

This analysis estimated travel times to the nearest hospital based on estimated hospital locations, and MAP's friction map, which provided us the ability to calculate travel times. However, there are many other factors that influence physical access to acute care, including recognizing the need for medical care, access to personal or public transportation, and the ability to contact and receive emergency out-of-hospital care. Additionally, the acute care services available varies drastically across facilities. This paper does not address variation in these aspects of physical access. It also does not address socio-economic barriers to care, or variation in quality of care. Travel time to acute facilities does provide an "upper bound" on access to care, as being able to reach a hospital is a minimum requirement for receiving many health services.

There are many limitations and assumptions associated with the data used in this study. We assumed that government reported hospital locations were complete, which is unlikely to be true, particularly in lower income locations. Validating the MoH datasets would require a concerted census of hospitals. We included reported hospital censuses from a range of years, thus we assumed that no new hospitals have been built or existing hospitals had been closed. Additionally, our calculation of travel time is dependent on the available friction map, which does not capture variation in travel times due to seasonal trends in road quality and will not take into account emergency helicopter services. The friction map has no uncertainty associated with it, for this reason, we are unable to estimate uncertainty in locations where MoH data is available. A concerted effort to collect and standardize data about hospital locations and services is needed in order to improve our certainty about access globally.

Our model for estimating locations of hospitals and travel times also has limitations. Our rigorous out-of-sample validation demonstrates that our model is limited in its ability to capture all untagged hospitals. Additionally, our MoH training dataset included no data from the Middle East and Asia, and it is possible that geographic patterns in hospital locations and use of open source software is different in these locations than in locations where we had MoH data. Due to computational limitations, we were limited to using only forty simulations. This number is unlikely to appropriately estimate 95% credible intervals. Calculating travel times by location allowed us to significantly reduce the computational burden of estimating this for the globe. However, it forced us to assume that individuals would not cross borders in order to receive care. In geographically large countries, we were further required to assume that individuals would not cross subnational unit borders to receive care. These assumptions likely cause us to overestimate travel time near location borders. Finally, because we estimated whether or not a 8x8 km raster was a CCH and not the precise coordinates of each hospital, our estimates of distance to nearest hospital are imprecise by up to $2 * \frac{8}{\sqrt{2}} km = 11km$, which may represent a clinically significant delay in time to receiving acute care.

Conclusion

In estimating travel time to nearest facility, we provide a global geospatial measure of access to care for the globe, allowing policy-makers to begin understanding physical limitations in access to care at a high resolution. We identified large variation in access both within and between countries, identifying the need for improvement in hospital care coverage in low-income and rural areas.

Although travel time provides "upper bound" on physical access, it does not necessarily reflect true access to care. Future work to estimate coverage of specific interventions, quality of services, and equal levels of access across socio-economic groups is necessary to identify where improvements health systems are most needed.

References

1. Ouma, P. O. *et al.* Access to emergency hospital care provided by the public sector in sub-Saharan Africa in 2015: a geocoded inventory and spatial analysis. *Lancet Glob. Health* **6**, e342–e350 (2018).
2. Weiss, D. J. *et al.* A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **553**, 333–336 (2018).
3. Christie, S. & Fone, D. Equity of access to tertiary hospitals in Wales: a travel time analysis. *J. Public Health* **25**, 344–350 (2003).
4. de Souza, V. C. & Strachan, D. P. Relationship between travel time to the nearest hospital and survival from ruptured abdominal aortic aneurysms: record linkage study. *J. Public Health* **27**, 165–170 (2005).
5. Ravelli, A. C. J. *et al.* Travel time from home to hospital and adverse perinatal outcomes in women at term in the Netherlands. *BJOG Int. J. Obstet. Gynaecol.* **118**, 457–465 (2011).
6. Cheng, G. *et al.* Spatial difference analysis for accessibility to high level hospitals based on travel time in Shenzhen, China. *Habitat Int.* **53**, 485–494 (2016).
7. Buhlungu, S., Daniel, J. & Southall, R. *State of the Nation: South Africa 2007*. (HSRC Press, 2007).
8. Perry, B. & Gesler, W. Physical access to primary health care in Andean Bolivia. *Soc. Sci. Med.* **50**, 1177–1188 (2000).
9. healthsites.io. Available at: https://healthsites.io/#about. (Accessed: 8th June 2019)

10. Local Burden of Disease. *Institute for Health Metrics and Evaluation* (2016). Available at: http://www.healthdata.org/lbd. (Accessed: 8th June 2019)
11. Liaw, A. & Wiener, M. Classification and Regression by randomForest. **2**, 5 (2002).
12. Chen, T. xgboost: Extreme Gradient Boosting version 0.82.1 from CRAN. (2019). Available at: https://rdrr.io/cran/xgboost/. (Accessed: 8th June 2019)
13. Etten, J. van. *gdistance: Distances and Routes on Geographical Grids*. (2018).
14. Alter, D. A., Naylor, C. D., Austin, P. & Tu, J. V. Effects of Socioeconomic Status on Access to Invasive Cardiac Procedures and on Mortality after Acute Myocardial Infarction. *N. Engl. J. Med.* **341**, 1359–1367 (1999).
15. Joynt, K. E., Harris, Y., Orav, E. J. & Jha, A. K. Quality of Care and Patient Outcomes in Critical Access Rural Hospitals. *JAMA* **306**, 45–52 (2011).
16. Baldwin, L.-M. *et al.* Quality of Care for Acute Myocardial Infarction in Rural and Urban US Hospitals. *J. Rural Health* **20**, 99–108 (2004).