Final report

The spatial distribution of health services in Zambia

Torsten Figueiredo Walter

June 2018

When citing this paper, please use the title and the following reference number:
S-41425-ZMB-1

DIRECTED BY

FUNDED BY

The Spatial Distribution of Health Services in Zambia

Integrating Administrative Databases to Improve Resource Allocation in the Zambian Health Sector

Torsten Figueiredo Walter[1]

06/2018

Introduction

There is an overall lack of health resources in Zambia. For example, the number of doctors, nurses, and midwives per thousand population is only 0.98[2] while the WHO recommends a minimum of 4.45 (Scheffler et al. 2018). Given the shortage of health resources, an efficient use of the available resources is key. This report discusses how available administrative data can help guide policy makers' resource allocation decisions.

The Zambian Ministry of Health (MoH) and its partners collect a wide variety of data on the distribution of health services. This includes information on the location, infrastructure, staffing, and service provision of health facilities. However, in general this information is recorded in separate datasets that are not linked with each other. We integrate the various administrative datasets with additional publicly available data on population and infrastructure into a single health database. Two exemplary uses of the resulting database for resource allocation are demonstrated:

1. Analysis of the distribution of health workers across primary care facilities
2. Analysis of the spatial distribution of laboratories relative to population

Below we describe the lessons learnt during the database construction process, summarize the results from the two exemplary analyses, and suggest further uses of the database.

Acknowledgement

Many thanks go to the government and NGO partners who supported this work through providing data and key contextual information, specifically EQUIP Zambia, the Directorate of Human Resources, Directorate of Public Health & Research, Directorate of Clinical Care, and the Directorate of Policy & Planning, particularly the Information & Communication Technology and Monitoring & Evaluation units. Additionally, thanks go to the International Growth Centre for financial support and Innovations for Poverty Action for support in project development.

---

[1] Department of Economics, London School of Economics. Email: t.b.walter@lse.ac.uk. I thank Dylan Knaggs for invaluable research assistance.

[2] Source: Global Health Workforce Statistics database (2016).

Data Sources

Administrative Data from MoH

The following administrative datasets were used to create the comprehensive health database:

- **EQUIP Census:** In 2017, EQUIP Zambia conducted a census of all health facilities in the country. This census provides the most complete listing of health facilities in Zambia. It contains information on facility locations as well as a range of variables on facility infrastructure and equipment.
- **MoH Census**: The Ministry of Health has periodically conducted censuses of facilities in Zambia, with the most recent available census conducted in 2012-2013. This dataset includes information on facility infrastructure, outreach/catchment size, and a breakdown of services offered at each facility.
- **HRIS**: The MoH Human Resource Information System (HRIS) provides information on staffing levels across all government health facilities by cadre and position. Data used in this report is from January 2018.
- **HMIS**: The MoH Health Management and Information System (HMIS) provides key information on service delivery and patient outcomes by facility. It is maintained by the MoH Monitoring & Evaluation Unit (MoH M&E) and updated monthly. Data extracted for this report is from January 2017-January 2018 and focuses on birth statistics.

Population Data

In addition to MoH datasets, we also integrate external data on population and infrastructure. These allow for the computation of catchment areas and populations. The below datasets are available to the public and require no special licenses or permissions to use.

Several data sources provide high-resolution population estimates. The main source used here is WorldPop, a project from the University of Southampton that uses remote sensing of satellite imagery and statistical techniques to estimate population within $100m^2$ grid cells in Africa, Latin America, and Asia[3]. While this data provides population estimates at a very fine level, it should be noted that there is suggestive evidence that it tends to underestimate populations in very rural areas.

Travel Time

Based on geographic and road data from Google and OpenStreetMap, the Malaria Atlas Project (MAP)[4] at Oxford University has produced a dataset dividing the world into grid cells of approximately $1km^2$ indicating the difficulty to travel through each grid cell. This data can be used to calculate the time it takes to travel between any two points, which is highly valuable for this work as it allows for the calculation of catchment areas based on how long the population must travel to access health care rather than simply its straight-line distance from facilities.

---

[3] See http://www.worldpop.org.uk/ for details. Another data source considered in this project is the Gridded Population of the World (http://sedac.ciesin.columbia.edu/data/collection/gpw-v4). In general, results are not sensitive to the choice of population data. Therefore, only results based on WorldPop are reported here. Results based on the Gridded Population of the World are available on request.

[4] See Weiss et al. (2018) and https://map.ox.ac.uk/research-project/accessibility_to_cities/ for details.

Merging Administrative Datasets

Process

Because there are no common facility identifiers across the different administrative datasets, it was necessary to merge them by facility name. To begin this process, the datasets were cleaned individually, and facility names were standardized so that they had the same format across datasets (e.g. initials were expanded so that *HP* became *Health Post*). The initial merge attempt for each pair of datasets was to find exact matches by both name and district, and then by name alone (with these matches manually reviewed for accuracy). A combination of manual review and string matching techniques was used to find the remaining matches. The appendix contains detailed notes on the cleaning and merging processes.

Results

The below table indicates the match rates across datasets[5]:

Reference Dataset		Merged Dataset			
		EQUIP Census	**MoH Census**	**HRIS[6]**	**HMIS**
	EQUIP Census	2,650 (100%)	1,729 (65%)	2,010 (74%)	2,306 (87%)
	MoH Census	1,729 (88%)	1,956 (100%)	1,501 (77%)	1,770 (90%)
	HRIS[6]	2,010 (83%)	1,501 (64%)	2,329 (100%)	1,897 (82%)
	HMIS	2,306 (93%)	1,770 (71%)	1,897 (76%)	2,483 (100%)

As the table indicates, match rates vary significantly across datasets. This is due to a combination of varying levels of data quality and completeness within the datasets. Below are notes on factors that impact match rates:

- Most unmatched facilities are small health posts, which do not seem to be captured in all datasets – intentionally or not.
- Different datasets have varying levels of coverage of private facilities. While HRIS does not include any facilities where MoH staff are not working, likely excluding private facilities, all other datasets cover private facilities to some extent.
- HRIS is reported at the individual level and is missing facility names for 12% of staff. It is possible that entire facilities are excluded from the dataset due to this missingness.
- The MoH Census excludes all facilities built more recently than 2013. This explains why match rates from it to other datasets are generally high, but match rates from other datasets to the MoH Census are low.

[5] Percentages indicate the share of the reference dataset matched to the merged dataset, so the 1,729 matches between the EQUIP Census and MoH Census represent 65% of the EQUIP Census and 88% of the MoH Census.

[6] Excludes administrative offices and staff with a missing facility name (approximately 10% of the dataset)

Constructing Catchment Areas

We construct catchment areas for all primary care facilities (health posts and health centres)[7]. We use three different approaches to estimate catchment area populations. First, we use catchment area headcounts performed by facilities and collected in the EQUIP Census. These headcounts are available for 1,704 (68%) of primary care facilities. Then we construct catchment areas and estimates of their population based on the aforementioned WorldPop satellite data in two distinct ways. The first relies on finding the straight-line distance between each facility and each population grid cell centre point. A grid cell is assigned to a facility's catchment area if it is closer to that facility than to any other. The second uses travel time rather than straight-line distance. A grid cell is assigned to a facility's catchment area if it takes less time to travel to that facility than to any other (as determined by the MAP data previously described). In both cases, the catchment population of a specific facility is determined by summing the population over all cells in the facility's catchment area.

Figure 1 demonstrates the similarities and differences between the two satellite-data-based approaches. While the straight-line approach yields fairly regular polygons (left), travel time-based catchment areas conform more closely with natural boundaries, and extend along major roads (right). Overall, 30% of Zambia's area and 31% of its population lie in different catchment areas when using one method instead of the other.

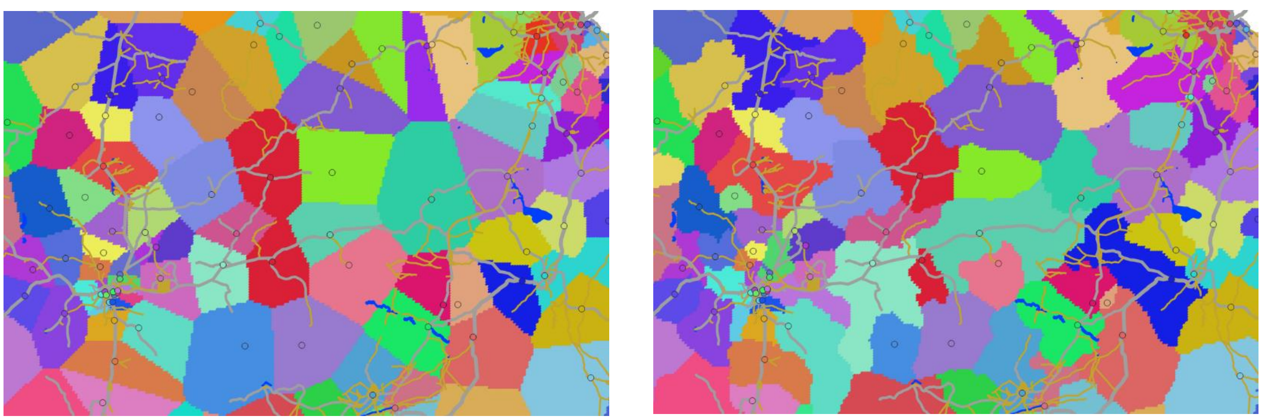

Figure 1: Inset map of straight-line (left) and travel-time (right) based catchment areas from WorldPop, as indicated by differently-shaded polygons. Dots indicate health facilities and lines indicate roads.

Figure 2 shows the correlation between satellite-based catchment population estimates and facility headcounts for all facilities where the latter are available. Independent from the estimation method, WorldPop-derived catchment populations are positively correlated with facility headcounts. Correlation coefficients are .37 and .35, respectively, and satellite-based estimates are generally lower than headcounts. Although satellite-based catchment population estimates do not line up perfectly with facility headcounts, they are helpful because they allow for an inclusion of facilities without headcounts into subsequent analyses. There are two reasons why these estimates may differ from headcounts. First, official facility catchment areas may not correspond to estimated catchment areas. Without a map of the official catchment areas, it is unfortunately not possible to assess this. Second, even if catchment

[7] The approach we develop can analogously be applied to higher-tier facilities.

areas were identical, estimates of their population would likely differ across the two sources (WorldPop and facility headcounts) due to their fundamentally different approaches to population estimation.

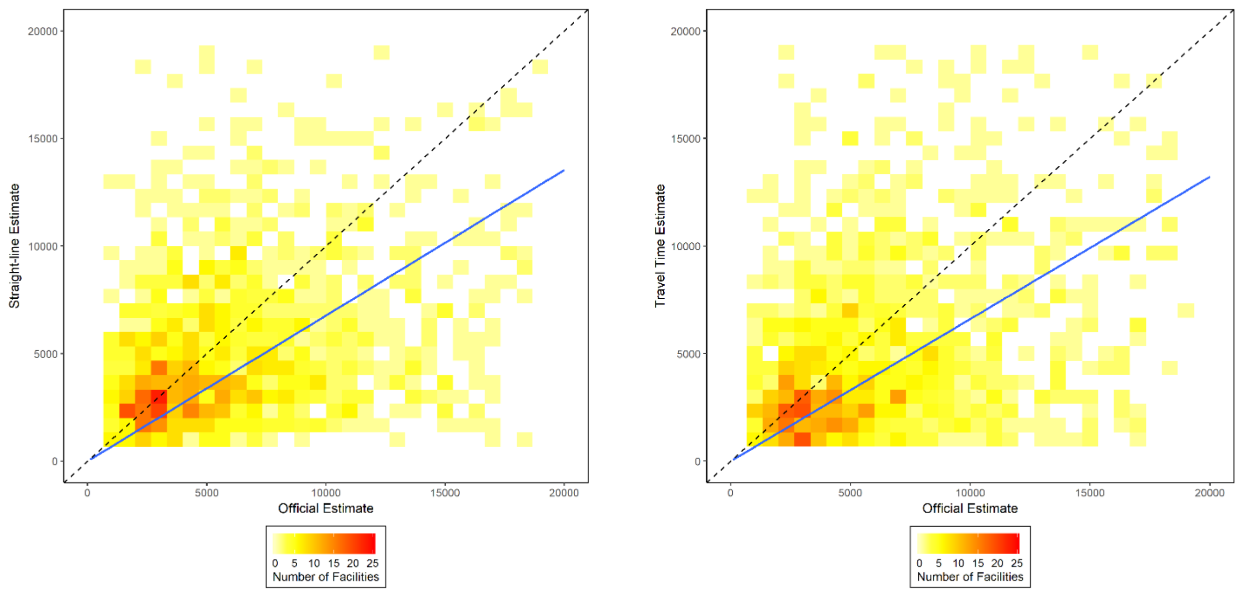

Figure 2: Comparison of straight-line (left) and travel-time (right) based catchment populations using WorldPop with catchment populations from facility headcounts, compiled by EQUIP. Black dashed line is 45°-line, and blue line linear regression line.

Applications

This section provides two examples of the use of the comprehensive health database. The examples study the distribution of health workers across primary care facilities and the access of the population to laboratory tests. These examples serve the purpose of illustration. Other potential uses of the comprehensive health database are discussed in the conclusion.

Distribution of health workers across primary care facilities

One basic use of the database is the analysis of facility staffing levels relative to catchment population. Figure 3 plots the distribution of population per medical staff across all primary care facilities for which catchment populations could be estimated using all three approaches described above. Medical staff is defined as nurses, midwives, clinical officers, environmental and community health workers, pharmaceutical staff, and doctors[8] and facilities are weighted by their catchment population. As the figure shows, there is a lot variation in relative staffing, independent from the estimation approach used. The relative staffing level of the primary care health facility of the average Zambian is such that there are 3,695 people per medical staff. However, the bottom 10% of the population live in catchment

[8] Results are qualitatively similar when using alternative definitions of medical staff.

areas of facilities with relative staffing levels of more than 8,133 people per medical staff while the top 10% of the population face staffing levels of less than 850 people per medical staff[9].

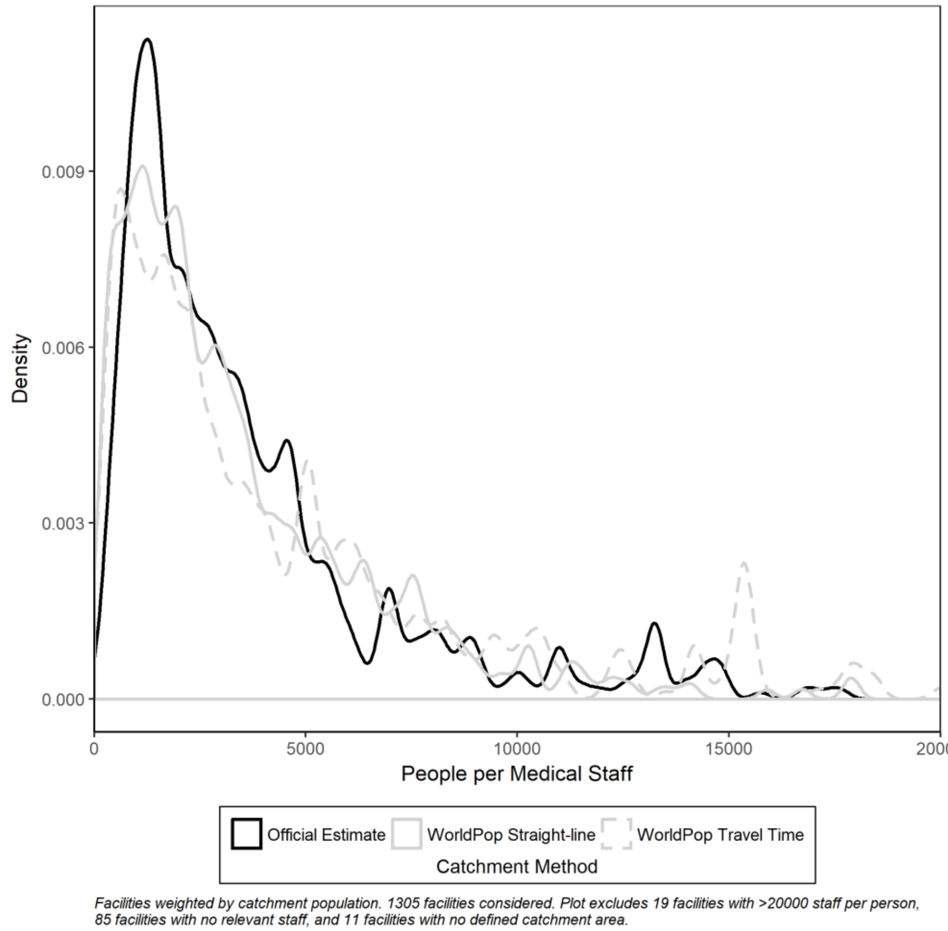

Figure 3: Distribution of population per medical staff across health facilities in Zambia

In order to verify the large dispersion in the relative supply of health workers across facilities, we consider an alternative indicator of demand, the number of births reported by each facility in HMIS. As Figure 4 illustrates, the average number of births per medical staff per month is strongly positively correlated with population (as measured by facility headcounts) per medical staff[10]. This corroborates the impression that facility staffing levels relative to demand are highly unequal.

Figure 5 maps the relative supply of health workers across facilities in Zambia based on official catchment population headcounts. The map shows the catchment areas of facilities with especially low

---

[9] Figures derived based on official estimates. WorldPop catchment populations are smaller and therefore provide smaller results. Straight-line estimates give a mean of 2,587 people per medical staff, with the bottom 10% having 430 people per medical staff and the top 10% at 5,759. Results based on travel time estimates are very similar.

[10] Analysis excludes facilities without regular birth data, defined as facilities without birth data in six months or fewer.

relative staffing levels in shades of red and those with high relative staffing levels in shades of green[11]. Catchment areas for facilities with missing data are in grey[12]. Lack and excess of staff are derived as follows. First, the staffing benchmark is computed by asking what the lowest population to medical staff ratio would be that could be achieved at all facilities if the objective were to equalize relative staffing across facilities (using official catchment populations for all facilities where this is available, this ratio is 2,054). Second, the difference between actual staffing and the benchmark is used to determine lack/excess[13].

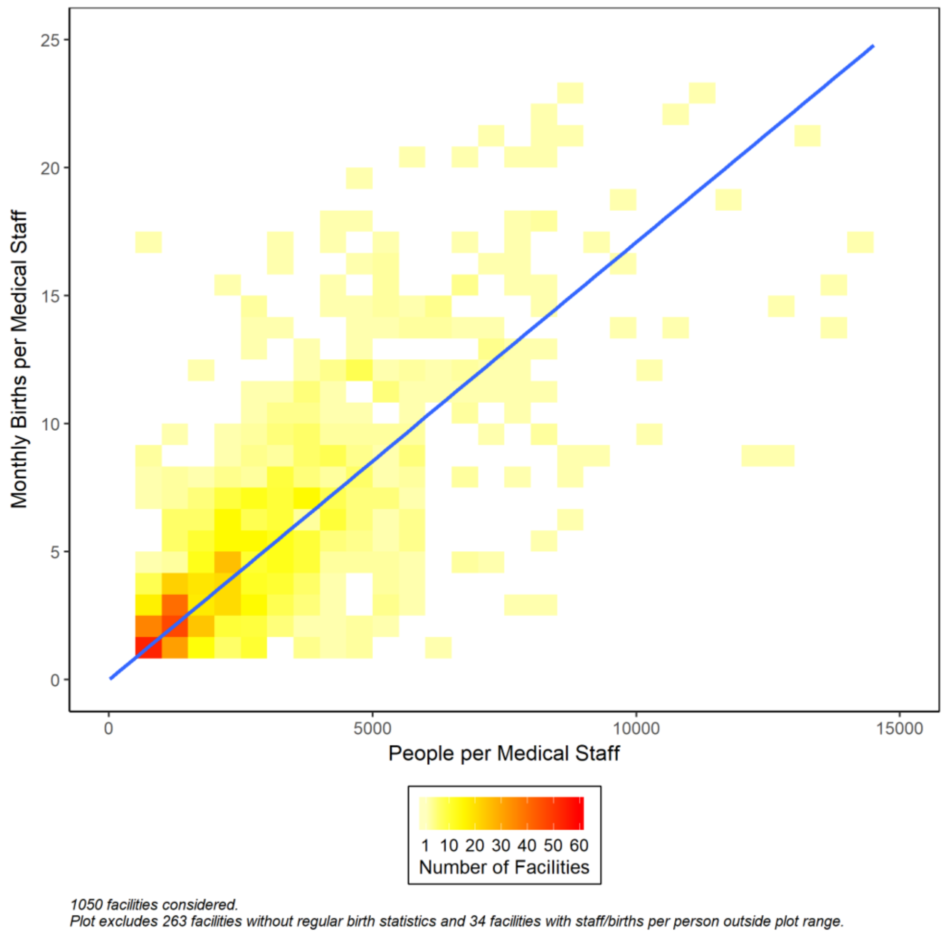

Figure 4: Distribution of births per medical staff across health facilities in Zambia

[11] The catchment areas used for illustration in the map are constructed using the previously described straight-line distance method.
[12] 928 facilities are missing data when using official catchment areas. 280 facilities are missing staff data because they could not be matched to HRIS. 655 facilities were not able to provide EQUIP with counts of their catchment population (some facilities fall into both categories). If WorldPop catchment populations were used, 339 additional facilities could be added to this analysis.
[13] Note that this is only one of many potential definitions of lack/excess of staff. Alternative definitions could easily be implemented instead.

The map indicates some broader regional trends, with a large number of comparatively well-staffed facilities in Lusaka and Copperbelt Provinces and more understaffed facilities in Northern, Luapula, and Muchinga Provinces. However, even within districts there are large differences in staffing levels, particularly in Eastern and Southern Provinces[14]. A decomposition of the variance in people per medical staff confirms this: the within-district variation is 4,595 while cross-district variation is 2,046[15].

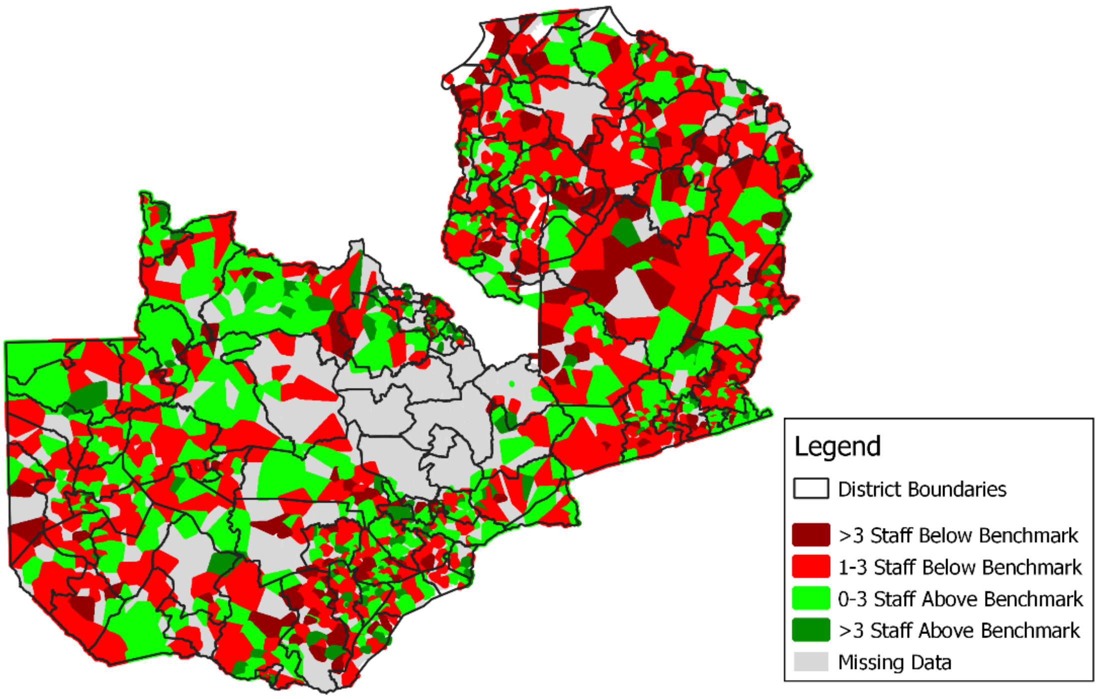

Figure 5: Map of health facility catchment areas with high (green) and low (red) staffing levels relative to catchment population

Taken together, the presented findings raise concerns regarding both equity in access to health services and efficiency in the provision of health services. Should medical staff be re-allocated across facilities to reach a more equitable distribution? Would such a re-allocation improve aggregate health outcomes? Why are some facilities understaffed while others are relatively well staffed? While these questions and other related ones remain unanswered here, we hope that the above analysis can provide a starting point for addressing these issues.

Access to Laboratories

Another example of the many potential uses of the comprehensive health database is the assessment of the spatial distribution of access to laboratories. We construct a laboratory's catchment area assuming each individual will travel to the closest primary care unit (as defined by travel time) to have a sample

[14] Maps of relative lack and excess supply of health workers based on WorldPop show broadly similar results. A version of the same figure based on WorldPop Straight-line is presented in the appendix. More than 70% of facilities above/below benchmark from official estimates are also above/below benchmark in WorldPop.

[15] Within-district variation describes how different the relative staffing levels of facilities in the same district are from each other. Cross-district variation describes how different the average relative staffing levels of each district are from each other. Figures are based on facility headcounts from the EQUIP census.

taken, and from the primary care unit the sample will be sent to the nearest laboratory. In both instances we use travel time to measure distance. In total there are 461 laboratories and different labs cover wide ranges of both area and population. Their catchment areas vary between 9km^2 at the 10th percentile and over 3,800 km^2 at the 90th percentile. Catchment population ranges from under 5,000 at the 10th percentile to nearly 75,000 at the 90th percentile [16].

Figure 6 maps all laboratories and their accessibility. Access to laboratories as measured by the sum of the travel time from place of residence to the nearest primary care facility and the travel time from this facility to the nearest lab is indicated in shades of grey. Note that this does not indicate how long it takes to receive laboratory results. It indicates transport time to the lab only and does not include any potential administrative or processing delays.

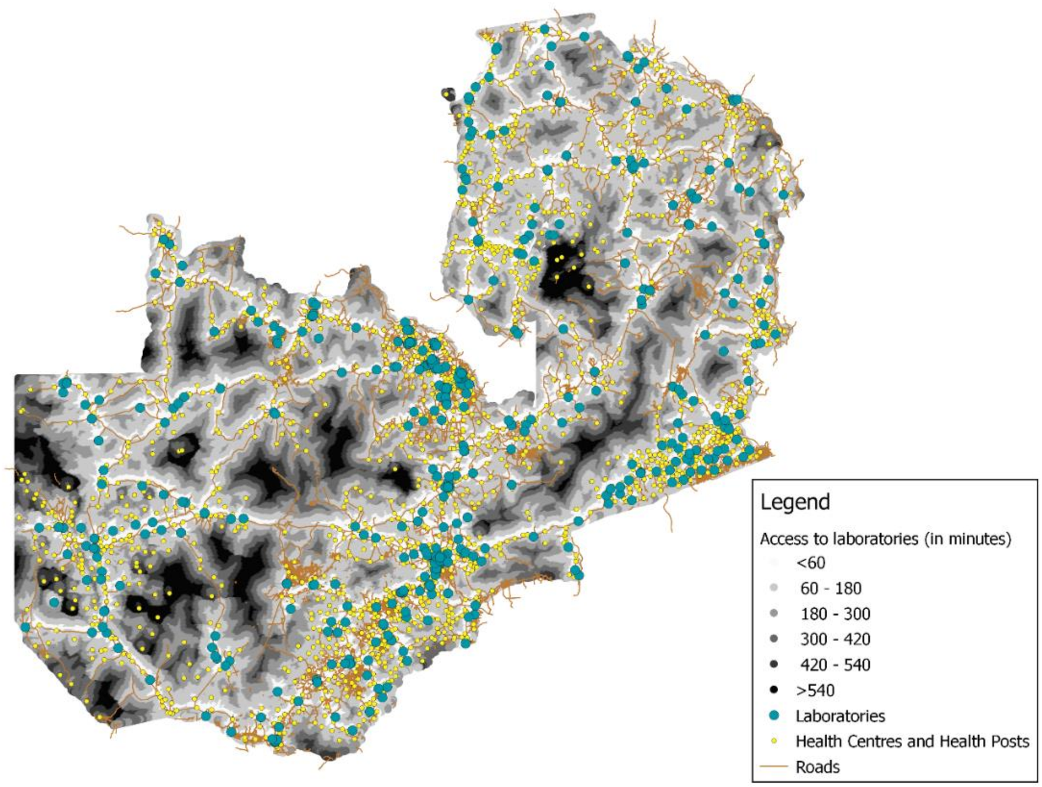

Figure 6: Map of access to laboratories

Figure 7 plots the cumulative distribution of access to laboratories across the population. It can be seen that there is substantial variation in access. The median person in Zambia has an access time of approximately 30 minutes to a lab, but for 10% of the population, this time is greater than 2.8 hours.

This second example illustrates how the comprehensive health database may be helpful when making decisions about the location of new health infrastructure, in this case laboratories. While population access to laboratories is likely to be an important factor for such decisions, it is clearly not the only one.

[16] Excludes private facilities with labs and facilities within 0.5km of another facility with a lab (usually this is a Hospital Affiliated Health Centre).

But other factors such as the spatial distribution of demand for specific types of laboratory analyses could also be assessed based on the comprehensive health database as long as the relevant indicators are contained.

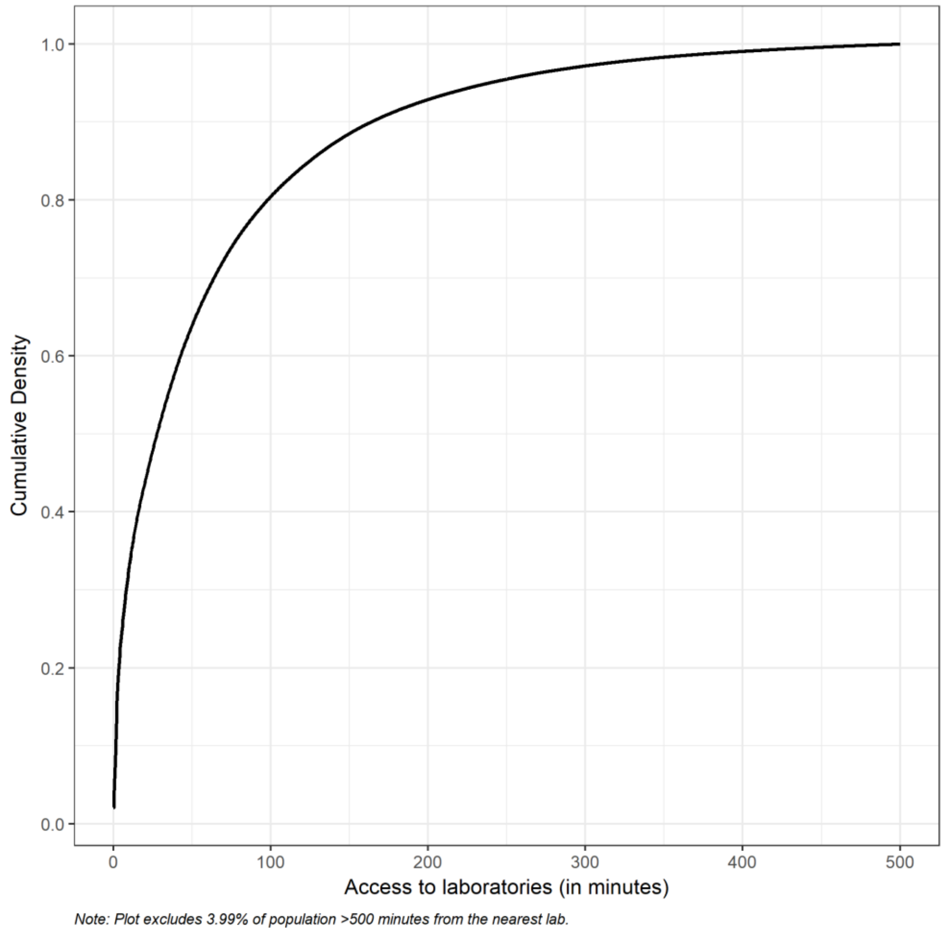

Figure 7: Cumulative distribution of access to laboratories across the population

Future Work and Recommendations

As noted, the above analyses are only two examples of the many applications that a comprehensive database can yield. Other examples include, but are not limited to:

- Location choice for the construction of new health facilities: analysis based on the comprehensive database could inform about the spatial distribution of access to and utilization of existing facilities and thus help identify areas of under-provision. In this context, the addition of facility construction dates to the database could be of interest so that the effects of past health facility construction on population health and demand at neighboring facilities can be better understood.
- Monitoring spatial disease patterns and coordinating countermeasures: the database allows for monthly monitoring of the occurrences of specific diseases across space. Disease patterns can then be related to observable facility and population characteristics and such linkages may provide guidance for countermeasures.

- Supply management: the database could help coordinate supply of facilities with medical equipment if relevant information such as health facility and storage center inventories were added to the system.

To fully develop the two exemplary applications and to create additional applications will require significant input from MoH and other local health experts, but the benefits seem large. In addition, the developed database should be updated regularly to maximize the usefulness of the developed database for policy and decision-making.

A number of recommendations based on the experience of building the database may be helpful with regards to a potential institutionalization of the database:

- Introduction of common facility identifiers across administrative datasets: While some datasets include a unique facility identifier that is internally meaningful, these identifiers do not relate to identifiers in other datasets. Developing a single identifier for every facility and applying it to all administrative datasets would allow for quick and error-free merges across datasets, and the ready extension of the database when new data is collected.
- Completion of datasets: Some facilities are missing from various datasets, even after accounting for their differences in target coverage. For example, a government facility may be contained in the EQUIP Census, but not in HRIS. In this case, the facility coordinates and catchment population are available, but there is no information on the staffing of the facility. Consequently, the facility cannot be considered when assessing the spatial distribution of medical staff. If key variables were collected for all facilities, a comprehensive analysis across the universe of facilities could be conducted.
- Central collection of catchment area boundaries and population counts: One key variable that is not available for all facilities is headcounts of the catchment population. These are in principle available from health facilities via the EQUIP Census, but many facilities do not have headcount data. Moreover, it is unclear which areas a facility is responsible for and thus, which areas its headcounts cover. Mapping the areas that facilities are responsible for and collecting headcounts from those areas would allow for improved resource allocation and planning.

References

Scheffler, Richard M., James Campbell, Giorgio Cometto, Akiko Maeda, Jenny Liu, Tim A. Bruckner, Daniel R. Arnold and Tim Evans (2018): Forecasting imbalances in the global health labor market and devising policy responses, Human Resources for Health 16:5.

Weiss, Daniel, A. Nelson, H.S. Gibson, W. Temperley, S. Peedell, A. Lieber, M. Hancher, E. Poyart, S. Belchior, N. Fullman, B. Mappin, U. Dalrymple, J. Rozier, T.C.D. Lucas, R.E. Howes, L.S. Tusting, S.Y. Kang, E. Cameron, D. Bisanzio, K.E. Battle, S. Bhatt, and P.W. Gething. A global map of travel time to cities to assess inequalities in accessibility in 2015. (2018). Nature. doi:10.1038/nature25181.

Appendix

Relative Staffing Based on WorldPop Catchment Populations

Figure 8 maps the relative supply of health workers across facilities in Zambia. This is identical to Figure 5, but it is based on WorldPop Straight-line catchment populations rather than the official headcounts collected by EQUIP. As discussed, results across figures are broadly similar. Because WorldPop is available nationwide, there are fewer areas with missing data (only EQUIP facilities that couldn't be matched to HRIS are excluded).

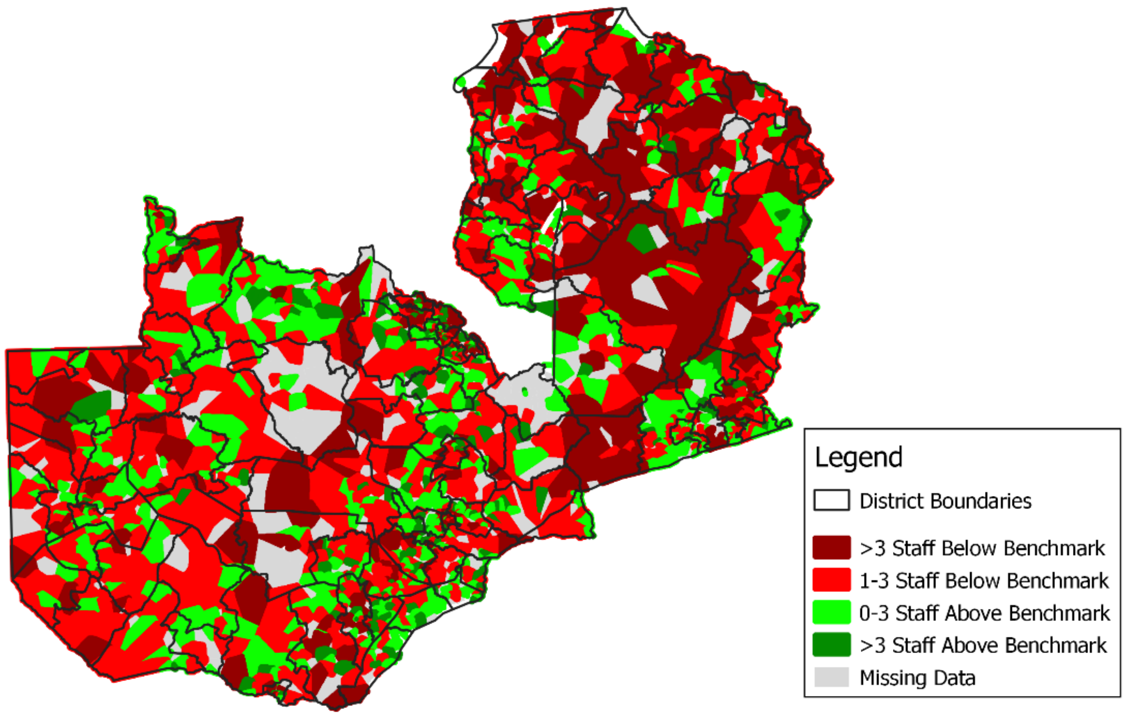

Figure 8: Map of health facility catchment areas with high (green) and low (red) staffing levels relative to catchment population, based on WorldPop Straight-line catchment populations

Notes on Data Cleaning

EQUIP Census

Missing Facilities

During the merge process, it was found that a small number of facilities included in HRIS, the MoH Census, and/or HMIS were not contained in the EQUIP Census[17]. Since none of the datasets apart from the EQUIP Census includes health facility coordinates and facility location information is essential to our exemplary analyses, we drew on the 2007 health facility census conducted by JICA to obtain coordinates

[17] As previously discussed, all datasets were missing some facilities included in others. However, none of the other datasets had analogous data that could be used to supplement them.

for as many of the facilities missing in the EQUIP Census as possible. Facility coordinates were added in the following way:

- The straight-line distance from each facility in EQUIP to the nearest facility in JICA was calculated, and pairs of nearest neighbor facilities between the censuses were identified.
- For each pair of nearest neighbors, the similarity between their names was calculated using a commonly-used string comparison algorithm.[18]
- Facilities from the JICA census were flagged if they had significantly different names than the nearest EQUIP facility or if they were not the nearest neighbor to any facility in the EQUIP Census. Each flagged facility was manually verified to ensure that it was not contained in the EQUIP Census. If verification was successful, facilities were added to the EQUIP dataset.

Through this process, 102 facilities were added to the EQUIP dataset. All references to EQUIP include these 102 facilities.

Duplicate Facilities

A small number of facilities in the same district were found to have identical names, suggesting that the facilities were duplicated in the dataset. Some of these were found to be the same facility, while others seem to be data entry errors in the facility name field, as the HMIS name was different. Duplicates were dropped, while the names of non-duplicate entries were changed to align with the HMIS name.

HRIS

Standardising Facility and District Names

For each MoH employee, HRIS contains the name of the district and the facility they work at. Facility and district names were standardized to ensure identical districts and facilities were spelled identically (e.g. Buleya HP and Buleya Health Post were both adjusted to Buleya Health Post). Cases where the identity of a facility was ambiguous (e.g. a facility name had a typo, and auto-correction procedures would not lead to a unique match with another facility name in the database) were reviewed manually.

Cleaning Individual Data

HRIS data includes two individual identifiers, personnel number and NRC number. While NRC records were nearly complete (only missing for one employee), personnel numbers were missing for nearly 2,000 employees. Approximately 500 NRCs appeared multiple times in the dataset. Based on manual review, identical NRCs generally represented the same employee, so duplicates were dropped. Where other variables (e.g. facility) varied within the same NRC, records that contained a personnel number were kept over those that did not because manual review suggested that entries with personnel numbers were of higher quality.

Notes on Data Merging

This section describes notable issues in the process of merging the different administrative datasets and cases where the process substantially deviated from that described in the main body of the report.

[18] Levenshtein string distance was used, which calculates the number of characters it is necessary to change between two strings until they are identical e.g. the string distance between “road” and “read” is .25 because 1 out of four characters must be changed for them to be identical. Facilities were flagged if over half of the characters in the shorter facility name would need to be changed to match the longer facility name.

- For pairs of datasets with a large number of unmatched facilities, approximate string matches were identified to support the manual review process. The match algorithm identified pairs of facilities with a significant number of consecutive 2-letter pairs in common (e.g. excluding the facility type, Buleya Health Post and Bulaya Health Post have three pairs in common – BU, UL, YA).
- As previously noted, HRIS was at the individual staff level, and contained different spellings for the same facilities. Most of these discrepancies were identified during the initial cleaning process, but others were identified only during manual review of the merges and it is possible that others were not identified at all. This could artificially deflate the match rate for HRIS if a match was found for only one facility spelling but not another. However, this should not have any significant impacts on the results presented in this report, as 97% of HRIS staff were matched to an EQUIP facility (with high rates for matches to the other datasets as well).

The International Growth Centre (IGC) aims to promote sustainable growth in developing countries by providing demand-led policy advice based on frontier research.

Find out more about our work on our website www.theigc.org

For media or communications enquiries, please contact mail@theigc.org

Subscribe to our newsletter and topic updates www.theigc.org/newsletter

Follow us on Twitter @the_igc

Contact us
International Growth Centre,
London School of Economic and Political Science,
Houghton Street,
London WC2A 2AE

International Growth Centre

DIRECTED BY

FUNDED BY

Designed by soapbox.co.uk