

WHAT OPEN GEOSPATIAL DATA CAN AND CANNOT TELL US ABOUT GEOGRAPHIC ACCESS TO RURAL PUBLIC PRIMARY CARE: A VILLAGE-LEVEL CASE STUDY FROM SINDH, PAKISTAN

Arsalan Kaleem

BS Student, Department of Computer Science, The Shaikh Ayaz University, Shikarpur, Pakistan

Email: arsalanabbasi.here@gmail.com

ORCID: 0009-0002-9054-6965

DOI: <https://doi.org/10.5281/zenodo.22828857>

Keywords

spatial accessibility; health facilities; OpenStreetMap; data completeness; geocoding; Pakistan; Sindh

Article History

Received: 01 June 2026

Accepted: 23 August 2026

Published: 08 September 2026

Copyright @Author

Corresponding Author: *

Arsalan Kaleem

Abstract

Open geospatial data is now used as a matter of routine in accessibility analysis for low- and middle-income countries, but how complete that data actually is for rural health infrastructure has never been checked against an authoritative registry. This study measures the deficit, and what the deficit does to the results, for the province of Sindh, Pakistan. Travel time from 5,159 villages to the nearest health facility was computed over the OpenStreetMap road network using multi-source Dijkstra routing, with independent routing verification agreeing to within three percent. Two facility layers were compared while every other input to the model was held constant: open data alone ($n=2,725$), and open data combined with 616 facilities recovered from the register of the province's primary-care operator and geocoded by name within district ($n=3,308$). Only 5.2 percent of the register facilities fall within 250 metres of any facility in the open dataset, so the two layers are describing different populations of buildings rather than the same buildings at different levels of completeness. Open-data-only results overestimated the population-weighted median travel time by a factor of 2.01, and the population more than thirty minutes away from care by a factor of ten, 6.95 million people against 0.68 million. Facility counts per district rose by as much as thirty-four-fold, and only in rural districts, while the metropolitan districts gained nothing at all. District ranking priorities differed between the two facility layers, and so did the district carrying the maximum population-weighted underservicing burden. Satellite imagery validation of a stratified random sample confirmed a consistent facility structure at 84 percent of geocoded points. In this setting open geospatial data leaves out the vast majority of rural public health facilities, and analyses built on top of it risk confusing the populations that are actually underserved with the ones that are merely misreported.

1. INTRODUCTION

Whether a person seeks care and receives it depends in part on whether they can physically

reach a facility, and in the rural areas of low- and middle-income countries distance, road quality and whatever transport happens to exist all press

on that decision at once. Measurement of access has moved over the past decade from straight-line distance to network-based travel time, which at least accounts for the road a person actually has to travel along. Weiss et al. (2020) produced global maps of travel time to healthcare at one-kilometre resolution and reported that 91.1 percent of the world's population can reach a hospital or clinic within one hour given motorised transport, while only 56.7 percent can do so on foot. That work, like most of what is done at continental or global scale, assembled its facility layer out of OpenStreetMap, Google Maps and published research datasets.

The apparatus for doing this kind of analysis is mature and free. Boeing (2017) introduced OSMnx, which allows any researcher to download, model and route over the road network of any region on Earth in a few lines of code. Put that together with open facility inventories such as healthsites.io (Saameli et al., 2018) and open population surfaces, and the practical barrier to conducting a village-scale accessibility study has largely gone. What has not kept pace is any validation of the inputs.

The completeness of OpenStreetMap varies sharply with development status and urbanisation. Herfort et al. (2023) modelled building-footprint completeness across 13,189 urban agglomerations and found that while 1,848 urban centres exceed 80 percent completeness, completeness stays below 20 percent for 9,163 cities holding 48 percent of the global urban population, with South Asia carrying some of the greatest residual uncertainty. Their central warning applies directly here: where spatial bias in completeness goes unaccounted for, researchers end up drawing general conclusions that are only valid for the areas that happen to be well mapped. Maina et al. (2019) and Macharia et al. (2025) have argued in parallel that geolocated, comprehensive and openly licensed health facility databases remain absent for much of sub-Saharan Africa, and that the facility lists which do exist are frequently incomplete, lack coordinates, or lean toward higher-level facilities. Pakistan has had comparatively little attention in this literature, and

the spatial work that does exist has tended to sit at district level or to concern itself with urban tertiary facilities. Malik and Shaikh (2019) analysed the spatial distribution of five public-sector tertiary teaching hospitals in Lahore using OpenStreetMap road data and network analysis, finding them concentrated in the north-eastern and east-central parts of the city. Ansari et al. (2015) examined access to comprehensive emergency obstetric and newborn care across three rural districts of Sindh through facility exit surveys and household interviews, and documented substantial out-of-pocket expenditure and travel burden. Kamal et al. (2025) mapped geographic disparities in healthcare access barriers among married women nationally using survey data and spatial scan statistics. None of them asks whether the facility layer sitting in open data represents the infrastructure that is actually there.

That omission matters because the assumption is load-bearing. If open sources capture rural public primary care poorly, then every quantity downstream of it in an accessibility study, the travel times, the count of underserved people, the ranking of districts by need, inherits the shortfall. This paper therefore treats the routing model as an instrument for measuring the gap rather than as the contribution itself. Two questions are addressed:

- (i) How completely does open geospatial data represent rural public primary care infrastructure in Sindh, when measured against the published register of the province's rural primary-care operator?
- (ii) What does any shortfall do to accessibility estimates and to the district prioritisation those estimates would support?

One boundary should be stated at the outset. The register used as the comparator covers rural public primary care only: basic health units, rural health centres, dispensaries and mother-and-child health centres. Taluka and district headquarter hospitals sit with the provincial health department, private facilities are recorded nowhere centrally, and neither was obtained. What follows is a study of geographic access to rural public primary care. It is not a study of healthcare access in the wider sense,

and the shortfall it reports is a shortfall in how open data represents that one tier. The reason any of this matters is that accessibility analysis is increasingly used to decide where health infrastructure should be placed and which populations are most in need. If the data underneath is incomplete in a spatially structured way, the analysis will point at the wrong places and do so confidently. Establishing whether that is the case, and by how much, has to come before open data can be used responsibly for health planning in Pakistan.

2. Review of Literature

2.1 Measurement of spatial accessibility

Spatial accessibility to health services is conventionally split into potential and revealed access, and potential access is split again into distance-based, gravity-based and floating-catchment approaches. Travel-time measures dominate large-scale work because they can be interpreted by a reader and computed at scale. Weiss et al. (2020) are the clearest example, generating friction surfaces from road, railway, land-cover and topographic layers and computing least-cost travel time to the nearest facility for every pixel globally. Ye et al. (2024), constructing national one-kilometre hospital accessibility maps for China, state the principal limitation of pure travel-time measures plainly: they capture the geographic dimension of access and say nothing about facility capacity or about the population competing for it, which is why gravity models and two-step floating catchment area methods came afterwards. National applications of the same approach exist for Australia (Barbieri & Jorm, 2019) and for outbreak preparedness in low-income settings (Hulland et al., 2019). Both take the facility layer as given.

For this study the simpler travel-time formulation is the appropriate one, because the object of inquiry is the facility layer and not the accessibility metric. Introducing capacity weighting would confound the comparison between scenarios without addressing the question that was posed.

2.2 Completeness and bias in volunteered geographic information

OpenStreetMap is volunteered geographic information, and its coverage reflects the distribution of the people who volunteer. Herfort et al. (2023) demonstrated that although data inequalities have receded in part through humanitarian mapping, a complicated pattern of spatial bias persists across Human Development Index groups, settlement sizes and world regions. Yeboah et al. (2021) assessed OpenStreetMap quality at successive stages of a participatory mapping process across seven informal settlements in Africa and Asia, finding remote-mapping completeness running from 84 percent in the most favourable morphology down to effectively zero in the most difficult, which says that completeness is a property of the interaction between built form and mapping method rather than of a region on its own. Friesen et al. (2025), comparing walking access to healthcare between formal and informal areas across nineteen sub-Saharan African cities, note in their own limitations that incomplete OpenStreetMap data could bias what they report. That is the usual position. The risk is acknowledged and then left unmeasured.

Falchetta et al. (2020) and Maina et al. (2019) have documented the specific problem of health facility databases: national lists frequently exist but lack coordinates, are not openly licensed, or are biased toward hospitals over primary facilities. Macharia et al. (2025) renewed the call for geolocated, comprehensive, openly licensed facility datasets, noting that where authoritative locations are unavailable researchers fall back on indirect geocoding against digital gazetteers such as GeoNames, Google Maps and OpenStreetMap, which is precisely the procedure adopted in the present study, and one whose error characteristics are seldom reported by anyone who uses it.

2.3 Accessibility research in Pakistan

Pakistani work on healthcare accessibility has largely been done at coarse spatial resolution or in urban settings. Malik and Shaikh (2019) applied network analysis to tertiary hospitals in Lahore,

taking facility locations from Google Earth and the road network from OpenStreetMap. Village-level analysis of the kind attempted here has been done elsewhere in the subcontinent: Singh and Guite (2023) examined public health facility accessibility across the villages of Muktsar district in Indian Punjab, found the distribution significantly clustered, and used facility-to-population ratios to identify underserved and over-served areas, though within a single district rather than a province. Ansari et al. (2015) and Kumar et al. (2019) approached rural Sindh through facility surveys rather than spatial modelling, the latter assessing emergency obstetric and newborn care signal functions at sixty-three public facilities across twelve districts. Baig et al. (2025) analysed subnational malaria burden across 1,088 primary healthcare facilities managed by the People's Primary Healthcare Initiative in twenty-three districts of rural Sindh, calculating incidence against each facility's catchment population, an analysis which could not have been carried out without georeferenced facility locations, so those coordinates plainly exist inside the operator's own systems even though nothing of the sort is published.

Village-scale, province-wide accessibility analysis for Sindh using open data appears not to have

been attempted, and no study located in the course of this work tests whether the open facility layer for Pakistan is adequate in the first place. That gap is what the present paper is written into.

3. Materials and Methods

3.1 Study area

Sindh is Pakistan's second most populous province, covering approximately 140,900 square kilometres across twenty-seven third-level administrative districts as delineated by the Database of Global Administrative Areas. The urban and rural halves of the province stand apart very sharply: the five districts comprising metropolitan Karachi occupy roughly one percent of the land area while containing a large share of the population. In the south-east, Tharparkar district is arid desert with documented recurrent emergencies in child mortality and malnutrition, which makes it a useful test of whether an accessibility model picks up genuine deprivation or not.

3.2 Data sources

Table 1 lists the layers used. The vintage of the OpenStreetMap extract is stated explicitly, since a study concerned with data completeness has to say which snapshot it measured.

Table 1: Data sources, vintages and licences

Layer	Source	Vintage	Licence
Villages (n = 5,159)	OSM-derived settlement points	—	ODbL
Road network	OpenStreetMap via Overpass API	Queried Jul–Aug 2026	ODbL
Facilities (open)	healthsites.io and OSM points of interest	12 July 2026 extract	ODbL
Facilities (register)	PPHI Sindh district facility lists	Published PDFs	Operator
Boundaries	GADM version 4.1, level 3	—	Academic use
Gazetteer	GeoNames Pakistan	—	CC-BY 4.0
Population	WorldPop-derived village estimates	—	CC-BY 4.0

3.3 Network routing

Travel time was computed using OSMnx 2.1.0 (Boeing, 2017) and NetworkX 3.6.1. For each district the drivable road network was downloaded from OpenStreetMap, edge speeds were assigned by highway class, and edge traversal times derived from those. The largest weakly connected component was retained so that unreachable fragments could not distort the results. Villages and facilities were snapped to their nearest graph nodes.

Travel time from every graph node to its nearest facility was obtained by a single multi-source Dijkstra traversal seeded simultaneously from all facility nodes, which returns the same distances as repeated single-source searches at a fraction of the computational cost. Province-scale routing in a single request exceeds Overpass API limits, so execution proceeded district by district with a

fifteen-kilometre facility halo extending beyond each district boundary, which allows villages near an administrative edge to reach facilities across that edge. Mirror rotation and per-district retries made the procedure resumable and fault-tolerant. Road speeds were set for rural Pakistani conditions rather than taken from library defaults, which assume free-flowing traffic on maintained surfaces and are optimistic for unpaved and single-lane rural roads (Table 2). The values are judgement, not measurement. No empirical speed survey exists for rural Sindh, and none was conducted here. They are stated in full so that a reader can substitute their own. Because the same speeds are applied to both scenarios, the comparison between them does not depend on the values being right, though the absolute travel times do.

Table 2. Assumed travel speeds by OpenStreetMap highway classification

Highway class	km/h	Highway class	km/h
Motorway	90	Unclassified	25
Trunk	70	Residential	20
Primary	55	Service	20
Secondary	45	Track	15
Tertiary	35	Path	8

3.4 Validation of the routing model

Model output was cross-checked against independent routing for the village of Shah Bandar in Thatta district, a remote deltaic settlement. The model returned 103.8 minutes to the District Headquarter Hospital at Thatta and independent routing returned 101 minutes, a discrepancy below three percent. This is a single case and cannot establish accuracy across the province. It shows that the graph construction and speed assumptions produce a plausible result on one long rural journey, and no more than that. It says nothing at all about the facility layer, which is the subject of everything that follows.

3.5 Recovery of the operator facility register

The People's Primary Healthcare Initiative (PPHI) Sindh manages primary healthcare across the province's rural districts and publishes district-wise facility lists as portable document format files. These contain facility name, district and facility type, but no coordinates. Six regional documents were parsed programmatically, yielding 1,357 facilities across twenty-seven districts, comprising basic health units, rural health centres, government dispensaries, mother-and-child health centres and hospitals.

PPHI facilities are named for the settlement in which they stand, for example "BHU Setharja" or "GD Khanpur". The settlement component was extracted by stripping facility-type prefixes and

matched against a combined gazetteer of GeoNames populated places for Sindh, expanded to include all Latin-script alternate spellings, together with the village layer. Matching used the token sort ratio of the RapidFuzz library, with an automatic acceptance threshold of 88 and a manual review band of 78 to 87.

Three design decisions in this procedure turned out to be consequential, and they are reported here because they transfer to other settings.

First, matching must be constrained spatially. An initial unconstrained implementation searched the whole province and placed only 25.7 percent of facilities in the correct district, with a median displacement of 78 kilometres, while returning high name-similarity scores throughout. Rural Sindh settlements are commonly named after landowners or clan heads and those names recur across the province, so without a spatial constraint the matcher simply returns whichever identical name it encounters first. In the final implementation, gazetteer places were assigned to districts by spatial join and each facility was matched only within the candidate set of its own district.

Second, the candidate buffer has to scale with polygon size. A fixed twelve-kilometre buffer, used to admit places near district boundaries, over-

reached in small districts and caused their matches to fail containment tests for reasons that were purely geometric, which ranked districts by area instead of by data quality. A buffer proportional to each district's equivalent radius, capped at twelve kilometres, raised the containment rate from 78.5 to 90.9 percent.

Third, the choice of string similarity metric is not a neutral one. The token set ratio, a common default, scores subset matches as perfect: the settlement "Khalifa Huryah" against the gazetteer entry "Khalifa" returns 100. Manual inspection revealed that 30.3 percent of accepted matches were of this form, including cases where two distinct facilities collapsed onto a single settlement. Substituting the token sort ratio, which penalises unmatched tokens, removed the failure mode at the cost of a lower recovery rate.

The final procedure produced 507 name matches at or above threshold, of which 477, or 94.1 percent, fell within the unbuffered polygon of their expected district. A further 266 borderline matches were screened against name-similarity and district-containment criteria and verified by the author, of which 139 were accepted. The resulting layer contains 616 facilities, 45.4 percent of the published register.

Table 3: Recovery of the operator facility register, step by step. Of 1,357 published facilities, 616 were located with coordinates verified to fall inside their stated district

Step	Operation	Retained	Lost
1	Facilities parsed from six regional PDF documents	1,357	—
2	Settlement name extractable after prefix stripping	1,356	1
3	Name matched within own district at score 88 or above	507	849
4	Match falls inside the expected district polygon	477	30
5	Borderline matches, score 78 to 87, screened by hand	+139	127 of 266 rejected
	Recovered facility layer	616	741

3.6 Scenario design

Two facility layers were routed. Scenario A comprises open data alone (2,725 facilities). Scenario B adds the geocoded register after spatial deduplication at 250 metres. The threshold is wider than the tens of metres that would suit two building locations, because register facilities sit at settlement centroids rather than at buildings: a

basic health unit mapped in open data 200 metres from the centre of the village it is named after is the same facility counted twice. Of the 616 recovered facilities, 32 fell within that distance of an existing point and were dropped, leaving 584 additions. One further point fell outside the provincial boundary when clipped. Scenario B therefore contains $2,725 + 584 - 1 = 3,308$

facilities. The threshold was not varied; a narrower one would admit a few more near-duplicates and a wider one would discard genuine facilities in closely spaced settlements. The village layer, road network construction, speed assumptions and halo distance were held constant, so any difference in output is attributable to the facility layer and to nothing else.

3.7 Accessibility scoring and population weighting

An accessibility score on a zero-to-one-hundred scale was computed by linear interpolation between fixed time anchors, with travel time t in minutes mapped through the points (0, 100), (15, 100), (30, 85), (60, 55), (120, 15) and (240, 0), and clamped at either end. The sixty-minute anchor follows the interval conventionally used for emergency obstetric and newborn care, below which a referral is considered reachable. The others are not clinically derived. Thirty minutes is a common working benchmark for routine primary care and fifteen and one hundred twenty were chosen to give the curve resolution at both ends. Section 4.6 reports how far the district ranking moves when these anchors are changed. A population-weighted underservice burden was then calculated for each district as the sum across villages of the score deficit multiplied by village population, which ranks districts by the number of people poorly served rather than by the number of villages.

3.8 Independence of the Correction

Because the correction is itself derived from mapping data, its independence from the bias it addresses was tested by correlating per-district geocoding success against prior mapping density using Spearman's rank correlation. Metropolitan Karachi districts were excluded from the primary test because PPHI facilities there are named after residential colonies rather than settlements, so name-based geocoding fails there for reasons of naming convention rather than of data quality.

3.9 Satellite Validation

A stratified random sample of fifty geocoded facilities, allocated proportionally across districts with a fixed random seed, was assessed against high-resolution satellite imagery in QGIS. Each location was scored for the presence of a structure consistent with a small rural health facility, typically a walled compound of twenty to sixty metres containing one or two low buildings with road access, within five hundred metres. Verdicts of yes, no or unclear were recorded, and proportions reported with Wilson score confidence intervals.

4. Results

4.1 The coverage gap

Open sources contain 2,725 health facilities for Sindh, of which 2,246, some 82 percent, lie within the five metropolitan Karachi districts occupying approximately one percent of the provincial land area. Figure 1 shows the spatial consequence.

Health facility coverage in open geospatial data, Sindh

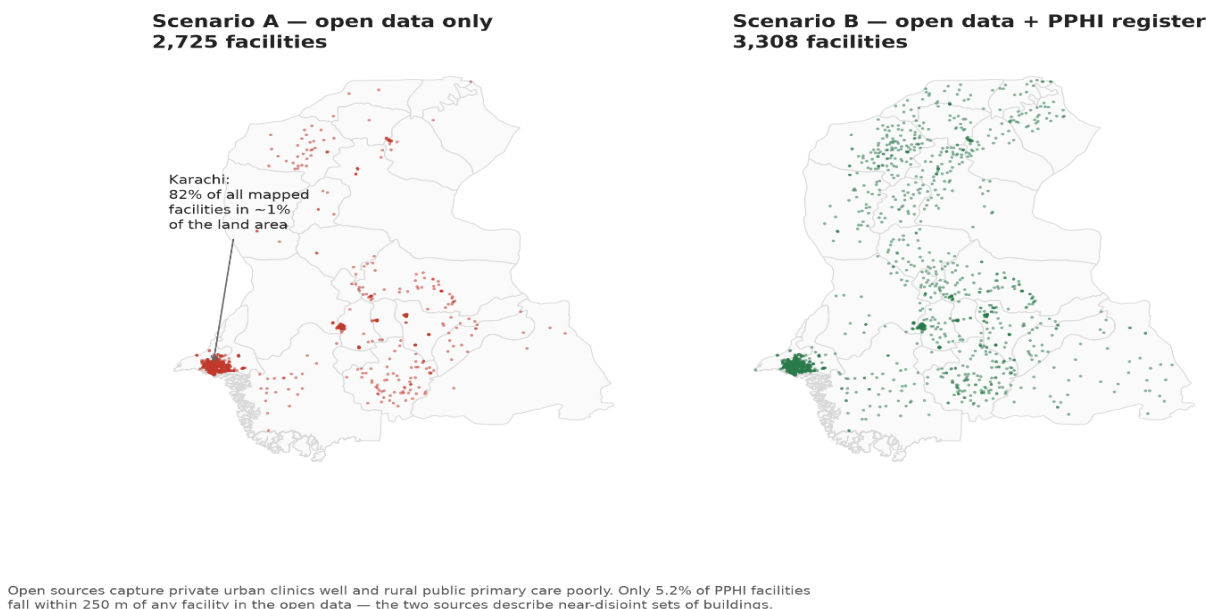


Figure 1. Health facility coverage in open geospatial data compared with the register-augmented layer.
Open sources capture private urban clinics densely and rural public primary care sparsely.

Of the 616 geocoded register facilities, only 32, which is 5.2 percent, fall within 250 metres of any facility present in the open data. That proportion is calculated against the facilities that could be located, not against the full register of 1,357. The 741 that could not be geocoded are, if anything,

less likely to appear in open sources than those that could. The two sources are therefore describing substantially different sets of buildings, not the same infrastructure at differing levels of completeness. That is a clearer statement of the coverage gap than any percentage of a total.

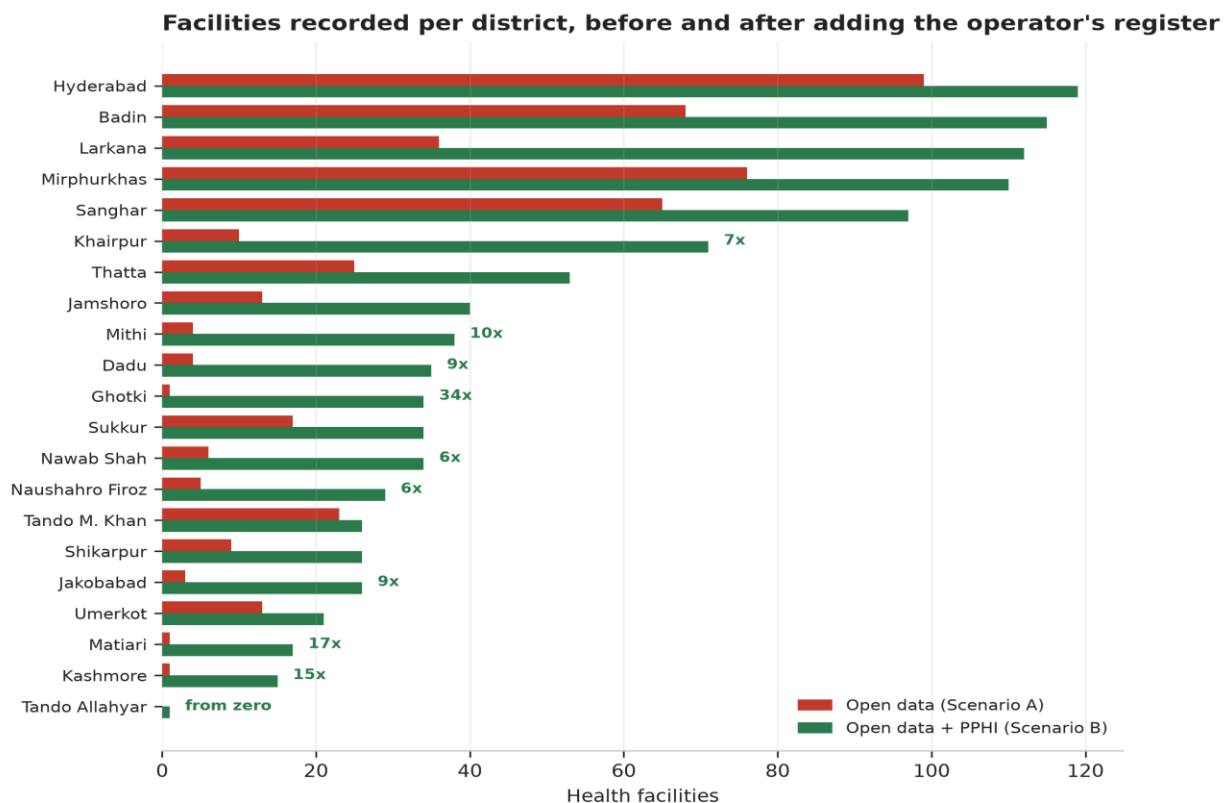
Table 4 and Figure 2 present the district-level position. The multiplier column expresses the Scenario B facility count as a ratio of the Scenario A count.

District	Open data	Scenario B	Multiplier	Register total
Ghotki	1	34	34.0	50
Kashmore	1	15	15.0	44
Matiari	1	17	17.0	36
Tharparkar (Mithi)	4	38	9.5	51
Dadu	4	35	8.8	50
Jacobabad	3	26	8.7	37
Khairpur	10	71	7.1	151
Naushahro Firoz	5	29	5.8	48

District	Open data	Scenario B	Multiplier	Register total
Tando Allahyar	0	1	—	47
Karachi East	800	801	1.0	8
Malir	226	231	1.0	35

Table 4. Facility counts by district under each scenario, selected districts. Metropolitan districts

gain nothing; rural districts gain by up to thirty-four-fold.



Karachi metropolitan districts omitted: they gain nothing and compress the axis.

Figure 2. Facilities recorded per district before and after adding the operator register. Karachi metropolitan districts are omitted because they gain nothing and compress the axis.

The pattern between urban and rural is a step function rather than a gradient. Every metropolitan district shows a multiplier of 1.0.

Tando Allahyar, a district with a population exceeding one million, contains no health facility whatsoever in the open data.

4.2 Consequences for estimated travel time

Table 5 summarises the province-level effect.

Measure	Scenario A	Scenario B	Factor
Median travel time (min)	19.1	9.2	2.08

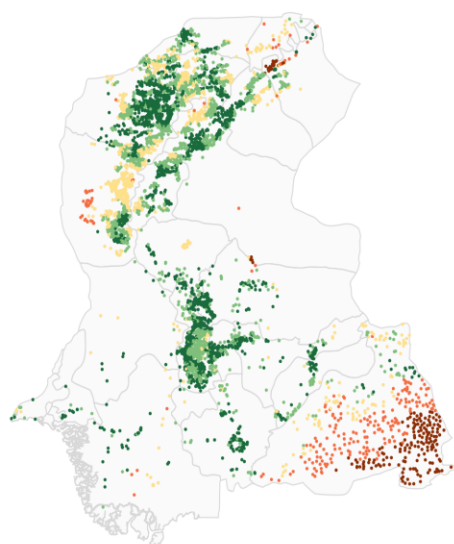
Measure	Scenario A	Scenario B	Factor
Population-weighted median (min)	14.3	7.1	2.01
90th percentile (min)	52.0	23.2	2.24
Maximum (min)	263.8	84.9	3.11
Villages over 30 min (%)	26.7	5.5	4.81
Villages over 60 min (%)	8.4	0.6	14.90

Open-data-only estimation overstates the population-weighted median travel time by a factor of 2.01. The estimate is robust to how degenerate cases are treated: excluding villages recording zero travel time under Scenario B gives a factor of 1.96, and excluding those below one

minute gives 1.94. It is also insensitive to the geocoding threshold, since an earlier and less conservative facility layer containing 768 geocoded facilities produced a factor of 2.16, so a twenty percent reduction in recovered facilities moved the estimate by four percent.

Estimated travel time to the nearest health facility, 5,159 villages in Sindh

Scenario A — open data only
median 19.1 min



Scenario B — open data + PPHI register
median 9.2 min

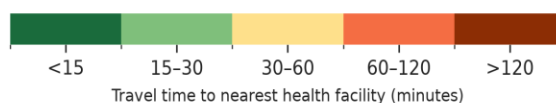
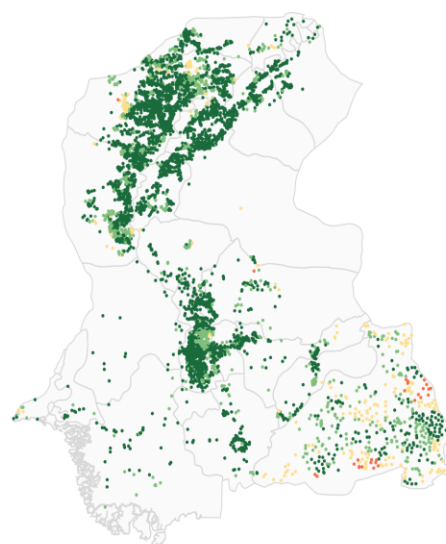
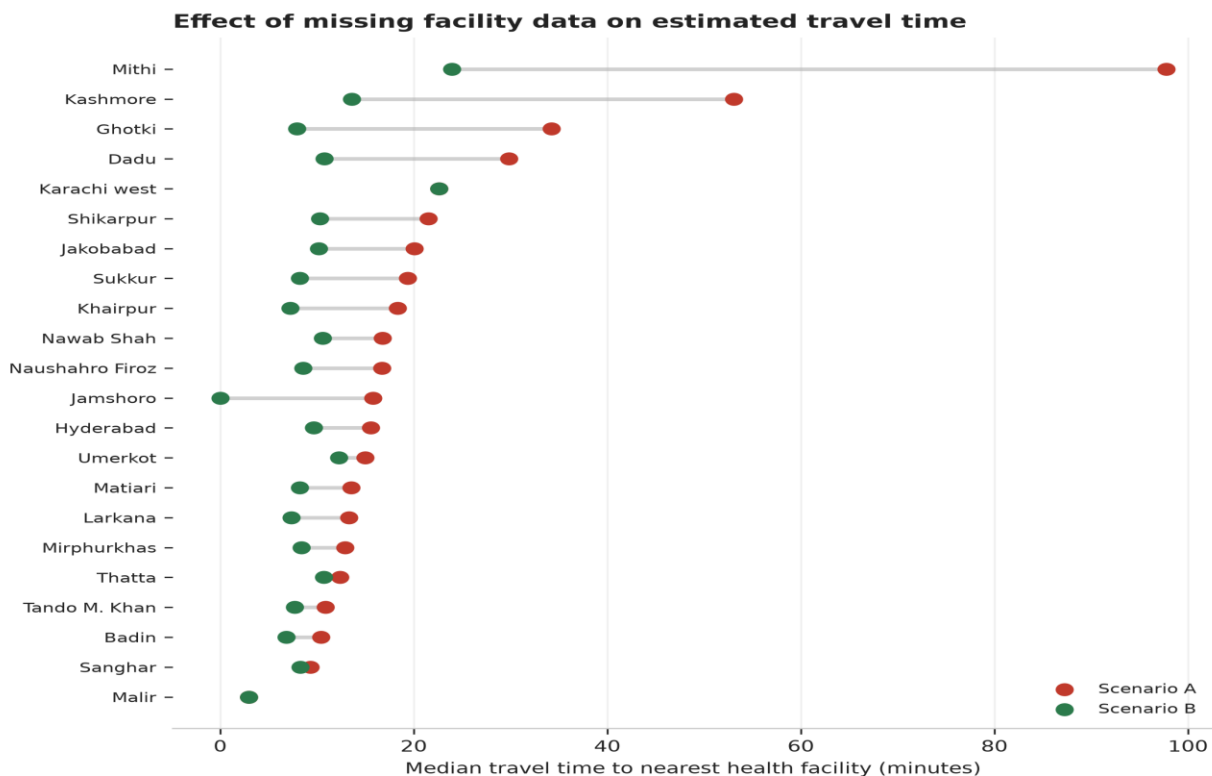


Figure 3. Estimated travel time to the nearest health facility for 5,159 villages under each scenario, on a shared classification.



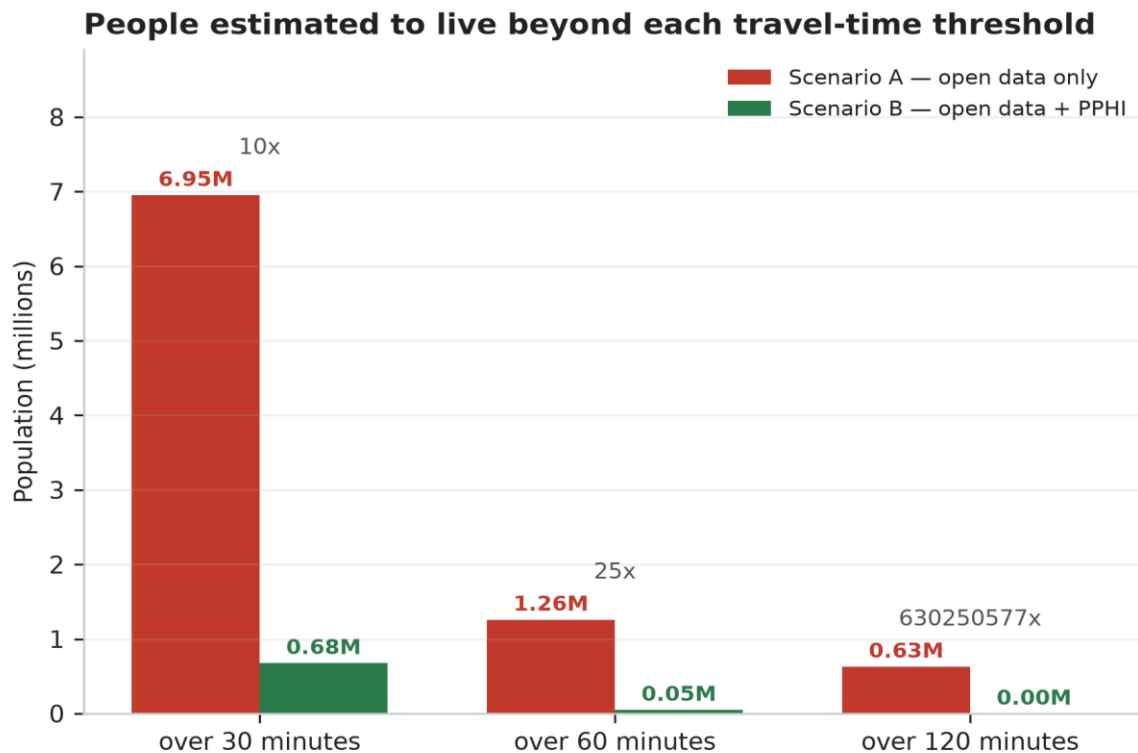
Districts with fewer than 10 villages excluded. Larger gaps indicate districts where open-data estimates were most misleading.

Figure 4. District median travel time under each scenario. Longer connecting segments indicate districts where open-data estimates were most misleading.

4.3 Consequences for population estimates

Translating travel time into affected population produces the largest discrepancies, because the

villages worst affected by the missing facilities are populous ones (Figure 5).



Total population covered: 49.3 million across 5,157 villages. Open-data-only estimates overstate the population beyond 30 minutes by a factor of 10.

Figure 5. Population estimated to live beyond each travel-time threshold under each scenario, across 49.3 million people in the 5,157 of 5,159 villages for which population estimates were available.

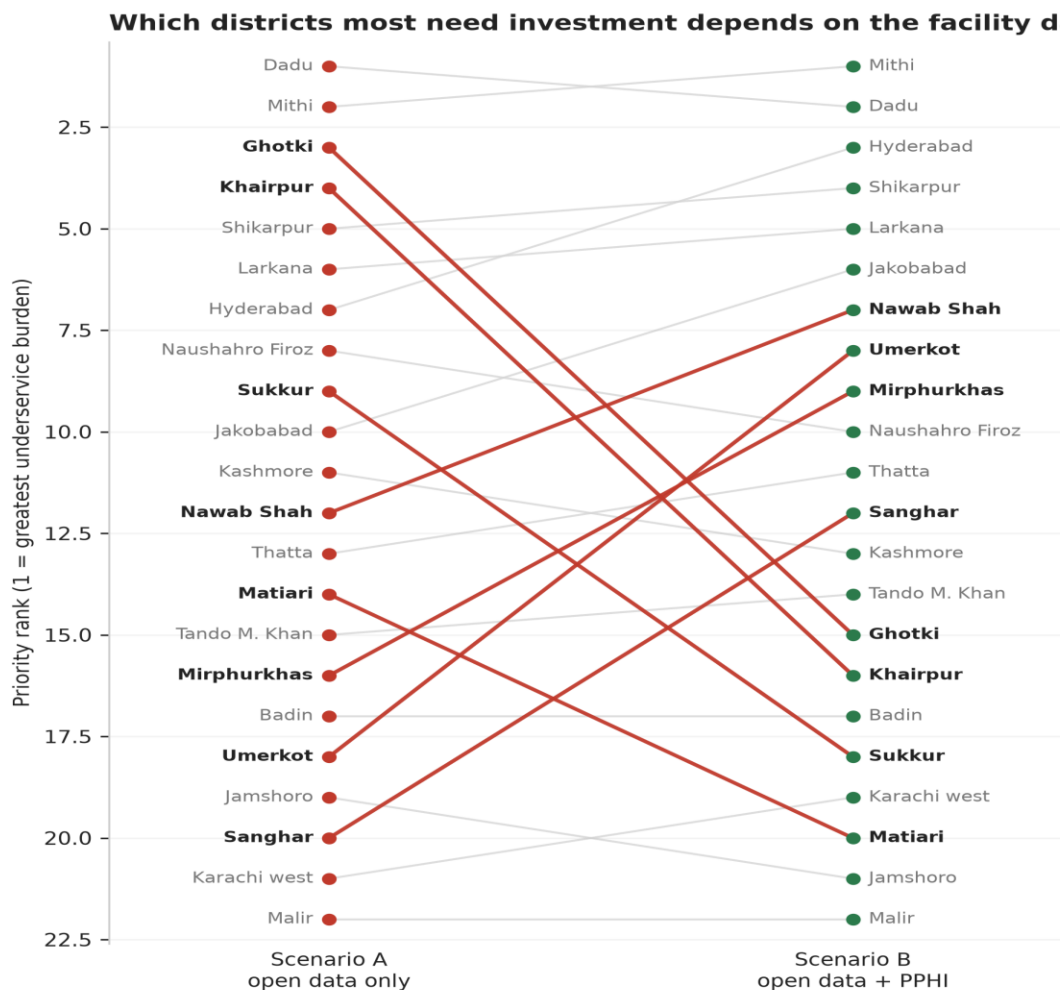
Under Scenario A, 6.95 million people appear to live more than thirty minutes from a health facility. Under Scenario B the figure is 0.68 million, a tenfold difference. At sixty minutes the corresponding figures are 1.26 million and 0.05 million, a twenty-five-fold difference. At one hundred twenty minutes, Scenario A identifies 0.63 million people while Scenario B identifies effectively none.

4.4 Consequences for district prioritisation

The results above concern magnitude. The ordering is the more consequential finding. Spearman rank correlation between district travel-

time rankings under the two scenarios is 0.657 ($p < 0.001$), with a maximum shift of fourteen places. When districts are ranked instead by population-weighted underservice burden, which is the quantity a health authority would actually act upon, the identity of the highest-priority district changes.

Under Scenario A, Dadu carries the greatest burden at 860,159, with Tharparkar second at 494,113. Under Scenario B, Tharparkar carries the greatest burden at 128,093, more than double Dadu's 58,726. Khairpur falls from fourth to sixteenth and Ghotki shifts twelve places. Figure 6 shows the reordering.



Ranked by population-weighted underservice burden. Districts shifting 5+ places highlighted (8 of 22). Open-data-only analysis misidentifies which districts are worst served, not merely by how much.

Figure 6. District priority ranking by population-weighted underservice burden under each scenario. Districts shifting five or more places are highlighted.

An analysis conducted on open data alone would therefore have directed attention to Dadu. With the operator's facility network accounted for, Tharparkar is unambiguously the province's priority, a district whose recurrent child mortality and malnutrition emergencies are independently documented.

4.5 Independence of the correction

Among the twenty-one rural districts, geocoding success was not significantly associated with prior

OpenStreetMap village mapping density (Spearman $\rho = +0.41$, $p = 0.064$). Association with open-data facility count was negligible and negative ($\rho = -0.12$, $p = 0.59$), which indicates that geocoding did not succeed preferentially where the open facility layer was already dense. The correction is therefore independent of the specific bias it addresses, though at $n = 21$ the village-density association is borderline and is reported as such rather than as an absence of effect.

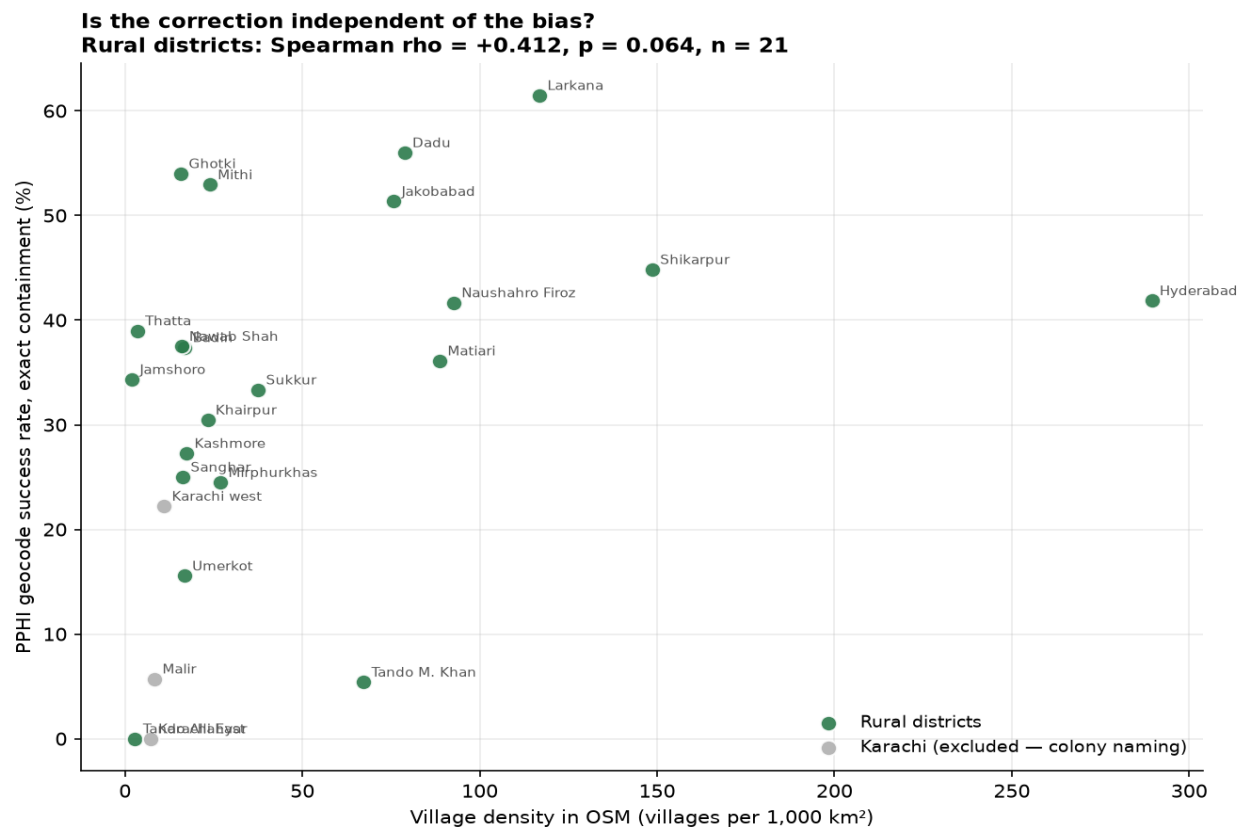


Figure 7. Geocoding success against prior OpenStreetMap village mapping density. Karachi districts are shown greyed and excluded from the primary test.

4.6 Validation and sensitivity

In the stratified random sample of fifty geocoded facilities assessed against satellite imagery, forty-two were scored as having a plausible facility structure within five hundred metres, five were scored negative and three unclear. Among clear verdicts this gives 89.4 percent (95% CI 77.4 to 95.4); treating unclear cases as negative gives 84.0 percent (95% CI 71.5 to 91.7). The conservative figure is the one reported as the headline.

Rank stability under alternative scoring bands was tested by varying the clinical time anchors. Under a stricter thirty-minute-centred scheme, district ranks correlated at $\rho = 0.976$ with the primary specification. Under a more lenient scheme anchored at one hundred twenty minutes, correlation fell to 0.716, which is what should be expected once Scenario B compresses nearly all travel times below thirty minutes and leaves a

scoring band that first penalises at two hours with almost nothing left to discriminate between. The resolution of a scoring scheme has to be matched to the quality of the facility data underneath it.

5. Discussion

5.1 What open data can and cannot support

The results allow a reasonably precise statement of what open geospatial data can be used for in this setting. It cannot support absolute travel-time estimates, which come out roughly twice too high, nor estimates of the population beyond a service threshold, which are out by an order of magnitude, nor district prioritisation, since the rankings are not preserved and the top-ranked district changes identity altogether. What it can support is method development, within-district relative comparison where the facility layer is locally complete, and the identification of gross

outliers, Tharparkar registering as extreme under both scenarios even though the degree differs by a factor of four.

A practical recommendation follows from that. Accessibility studies in comparable settings should either validate the facility layer against an authoritative register, or report results as bounds between an open-data-only scenario and a register-augmented one instead of as point estimates. The bounding approach costs almost nothing once the register is in hand and it communicates the uncertainty honestly.

5.2 Why the bias takes this form

The step function between urban and rural has a straightforward explanation. Volunteered geographic information follows its contributors, who are disproportionately urban and connected. A private clinic in Karachi is commercially visible and gets added to open databases by its operator or its customers, while a basic health unit in Tharparkar has neither a commercial incentive nor a local mapping community to do it. Humanitarian mapping is episodic and event-driven, generating coverage around disasters rather than systematic facility registers. This is consistent with the global pattern documented by Herfort et al. (2023) and with the informal-settlement findings of Yeboah et al. (2021), and it implies that the shortfall reported here is likely to generalise to other rural regions of South Asia.

5.3 Transferable methodological findings

Three findings arising from the geocoding procedure have application beyond this study. The first concerns script. Twenty-three percent of OpenStreetMap place names in Sindh are recorded in Urdu script. Taluka headquarter towns including Dokri, Ratodero, Naudero and Setharja returned no matches when searched in Latin script against a national extract of 29,746 places. Of 691 verified geocodes in an intermediate implementation, GeoNames supplied 637 and OpenStreetMap place names supplied 54. For Latin-script record linkage in Pakistan, OpenStreetMap place names are close to

unusable, and a Latin-script gazetteer carrying alternate-name coverage is a requirement.

Naming convention is the second. Rural Sindhi settlements are named after persons and clans, and those names recur widely across the province. Unconstrained fuzzy matching produced a 25.7 percent correct-district rate with 78 kilometres median displacement while returning high similarity scores throughout. A high string-similarity score is not evidence of a correct geocode in a setting like this one, and constraining the search spatially is a matter of correctness rather than of efficiency.

The third is metric selection. The token set ratio, widely used as a default, scores subset matches as perfect and produced a 30.3 percent subset-match rate in which distinguishing surnames were silently discarded. This failure mode is invisible in aggregate score distributions and was caught only by inspecting individual cases by hand against satellite imagery.

There is a further observation, about the limits of name-based linkage itself. In Tando Allahyar district the operator's facility names denote settlements by person and clan, Dodo Rind, Masoo Bozdar, Bachal Pittafi, while the gazetteer records the same territory under settlement-type toponyms such as Saidpur, Padhar and Matli. The two naming conventions describe the same landscape without ever intersecting, and no adjustment of the threshold can bridge them. Not one of the district's forty-seven facilities was recoverable.

5.4 The limits of what any remote method can establish

What this study does not establish should be stated plainly. Physical reachability, which the routing computes, is a different thing from physical presence, which satellite imagery can partially confirm, and both are different again from operational status, whether a facility is staffed, stocked with medicines, supplied with electricity and open when a patient arrives. Imagery confirms a roof; it cannot confirm a doctor. The problem of staff posted to a facility but

not attending it is invisible from orbit by construction.

Crossing from physical presence to operational status requires ground data, which in this context sits in provincial health information systems and in no open source at all. That crossing is the single highest-value addition that could be made to work of this kind, and no amount of improvement to remote methods will achieve it.

5.5 Implications

For researchers, the implication is that facility-layer completeness should be treated as a measured quantity rather than as an assumption, and that results should be bounded where it cannot be measured. For health authorities and facility operators, publishing facility coordinates alongside facility names would cost very little and would substantially change what external analysis is able to establish. The operator's own published research demonstrates that validated coordinates exist internally (Baig et al., 2025), so their absence from public documents is a choice and not a constraint. For the OpenStreetMap community, rural Pakistani health infrastructure is a documented and quantified mapping gap.

6. Limitations

The village layer is itself derived from OpenStreetMap and is a convenience sample rather than a census of settlements. Density ranges from 2.0 to 289.4 villages per thousand square kilometres across districts, a 148-fold spread. Individual village travel times remain valid, but unweighted province and cross-district aggregate partly reflect where contributors happened to be active, which is why all headline figures are population-weighted.

Geocoded facilities are placed at settlement centroids rather than at building locations. Typical positional error is nought to two kilometres, corresponding to nought to five minutes at rural driving speeds, which is small relative to travel times of twenty to one hundred minutes and biased toward underestimating travel time. Spot inspection during satellite validation suggested offsets frequently in the range of one to three

hundred metres, so the stated bound is a conservative one.

A total of 372 villages, 7.2 percent, record zero travel time under Scenario B because a facility geocoded to a settlement centroid snap to the same graph node as the village it is named after. This is defensible in that a village containing a basic health unit does have negligible travel time, but it is an artefact of the method and results are reported both with and without these cases.

Scenario B remains an undercount. Recovery rates were lowest in Tando Allahyar (2 percent), Tando Muhammad Khan (8 percent) and Umerkot (22 percent), so travel times for these districts remain upper bounds and are not presented as corrected. Seventeen villages, 0.3 percent, recorded marginally longer times under Scenario B (median plus 0.3 minutes, maximum plus 7.7 minutes), attributable to non-deterministic OpenStreetMap graph retrieval between runs rather than to the facility layer.

Two facility-type codes appearing in the operator's documents could not be resolved from any published source and were treated as primary-tier; they are flagged in the released dataset. The register covers primary care only, so secondary and tertiary coverage is incomplete by construction, since district and taluka headquarter hospitals fall under the provincial health department and were not obtained.

Finally, satellite validation was conducted by a single rater. Inter-rater reliability was not established, and the criterion applied, a plausible structure within five hundred metres, is permissive in dense settlements. The validation establishes that the coordinates fall in places where a facility could plausibly stand, not that the specific facility was identified.

7. Conclusion and Recommendations

Open geospatial data omits the majority of rural public primary care facilities in Sindh. Of the register facilities that could be located, only 5.2 percent appear in open sources, and the omission is spatially structured, with metropolitan districts unaffected while rural districts gain up to thirty-four-fold once the register is incorporated.

Accessibility estimates built on open data alone overstate population-weighted travel time by a factor of 2.01, overstate the population beyond thirty minutes from care by a factor of ten, and reorder district priorities to the point where the district identified as most in need changes identity. The routing methods themselves are standard and behaved plausibly where they could be cross-checked, matching independent routing to within three percent on the single long rural journey tested against independent routing. The deficiency lies wholly in the inputs. That distinction indicates where the effort should be directed, which is toward facility data rather than toward more sophisticated accessibility metrics.

Three recommendations follow. First, accessibility studies in comparable settings should validate the facility layer against an authoritative register where one exists, and otherwise report bounded estimates rather than point estimates. Second, health facility operators and provincial authorities should publish facility coordinates alongside facility names, since the marginal cost is negligible and the analytical gain is substantial. Third, researchers conducting name-based record linkage in South Asian contexts should constrain matching spatially, verify containment, and avoid similarity metrics that score subset matches as perfect, each of which produced a materially wrong result during this study before it was corrected.

All code and derived data are publicly available so that these findings can be verified and extended.

Declarations

Data and code availability. All analysis code and derived datasets are available at <https://github.com/ArsalanKaleem/Rural-Healthcare-Accessibility-Index-RHAL>. Raw third-party inputs are not redistributed; sources and vintages are listed in Table 1.

Funding. This research received no external funding.

Conflicts of interest. The author declares no conflict of interest.

Ethical considerations. This study used only publicly available and aggregate data. No

individual-level data were used and no human subjects were involved. No individual facility is identified as absent or non-functional; all such findings are reported as district-level rates.

Acknowledgements. This work builds on the OpenStreetMap, GeoNames and healthsites.io communities, whose freely licensed data made the analysis possible, and on the People's Primary Healthcare Initiative Sindh, whose published facility lists provided the denominator against which open data was assessed.

REFERENCES

- Kumar, R., Ahmed, J., Anwar, F., & Somrongthong, R. (2019). Availability of emergency obstetric and newborn care services at public health facilities of Sindh province in Pakistan. *BMC Health Services Research*, 19, 968. <https://doi.org/10.1186/s12913-019-4830-6>
- Ansari, M. S., Manzoor, R., Siddiqui, N., & Ahmed, A. M. (2015). Access to comprehensive emergency obstetric and newborn care facilities in three rural districts of Sindh province, Pakistan. *Health Research Policy and Systems*, 13(Suppl 1), 55. <https://doi.org/10.1186/s12961-015-0042-7>
- Baig, N., Ali, Z., Ahmed, M., Rahoojo, R. H., Jamal, W. Z., Usmani, B. A., Rehman, S. U., Jagirani, J. A., & Fatmi, Z. (2025). Subnational malaria burden in Sindh, Pakistan: Over a decade of evidence for tailored strategies. *Malaria Journal*, 24, 265. <https://doi.org/10.1186/s12936-025-05505-4>
- Barbieri, S., & Jorm, L. (2019). Travel times to hospitals in Australia. *Scientific Data*, 6, 248. <https://doi.org/10.1038/s41597-019-0266-4>

- Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65, 126–139. <https://doi.org/10.1016/j.compenvurbsys.2017.05.004>
- Boeing, G. (2025). Modeling and analyzing urban networks and amenities with OSMnx. *Geographical Analysis*, 57(4), 567–577. <https://doi.org/10.1111/gean.70009>
- Falchetta, G., Hammad, A. T., & Shayegh, S. (2020). Planning universal accessibility to public health care in sub-Saharan Africa. *Proceedings of the National Academy of Sciences*, 117(50), 31760–31769. <https://doi.org/10.1073/pnas.2009172117>
- Friesen, J., Georganos, S., & Haas, J. (2025). Differences in walking access to healthcare facilities between formal and informal areas in 19 sub-Saharan African cities. *Communications Medicine*, 5, 41. <https://doi.org/10.1038/s43856-025-00746-5>
- Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., & Zipf, A. (2023). A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. *Nature Communications*, 14, 3985. <https://doi.org/10.1038/s41467-023-39698-6>
- Hulland, E. N., Wiens, K. E., Shirude, S., Morgan, J. D., Bertozzi-Villa, A., Farag, T. H., Fullman, N., Kraemer, M. U. G., Miller-Petrie, M. K., Gupta, V., Reiner, R. C., Rabinowitz, P., Wasserheit, J. N., Bell, B. P., Hay, S. I., Pigott, D. M., & Graetz, N. (2019). Travel time to health facilities in areas of outbreak potential: Maps for guiding local preparedness and response. *BMC Medicine*, 17, 232. <https://doi.org/10.1186/s12916-019-1459-6>
- Kamal, A., Shah, G. H., Hafeez, A., Siddiqua, M., & Owens, C. (2025). Mapping geographic disparities in healthcare access barriers among married women in Pakistan: Evidence from a nationally representative survey. *Healthcare*, 13(19), 2448. <https://doi.org/10.3390/healthcare13192448>
- Macharia, P. M., Beňová, L., Ray, N., Semaan, A., Musau, M. M., Kipterer, J. K., Herringer, M., Snow, R. W., & Okiro, E. A. (2025). Putting health facilities on the map: A renewed call to create geolocated, comprehensive, updated, openly licensed dataset of health facilities in sub-Saharan African countries. *BMC Medicine*, 23, 211. <https://doi.org/10.1186/s12916-025-04023-z>
- Maina, J., Ouma, P. O., Macharia, P. M., Alegana, V. A., Mitto, B., Fall, I. S., Noor, A. M., Snow, R. W., & Okiro, E. A. (2019). A spatial database of health facilities managed by the public health sector in sub-Saharan Africa. *Scientific Data*, 6, 134. <https://doi.org/10.1038/s41597-019-0142-2>
- Malik, N. A., & Shaikh, M. A. (2019). Spatial distribution and accessibility to public sector tertiary care teaching hospitals: Case study from Pakistan. *Eastern Mediterranean Health Journal*, 25(6), 431–434. <https://doi.org/10.26719/emhj.18.049>
- Saameli, R., Kalubi, D., Herringer, M., Sutton, T., & de Roodenbeke, E. (2018). Healthsites.io: The global healthsites mapping project. In S. Hostettler, S. Najih Besson, & J.-C. Bolay (Eds.), *Technologies for development: From innovation to social impact* (pp. 53–59). Springer. https://doi.org/10.1007/978-3-319-91068-0_5

Singh, A., & Guite, L. T. S. (2023). Spatial accessibility analysis of public healthcare resources in Muktsar district of Punjab, India using geospatial technology. *GeoJournal*, 88(5), 5031–5042. <https://doi.org/10.1007/s10708-023-10907-5>

Weiss, D. J., Nelson, A., Vargas-Ruiz, C. A., Gligorić, K., Bavadekar, S., Gabrilovich, E., Bertozzi-Villa, A., Rozier, J., Gibson, H. S., Shekel, T., Kamath, C., Lieber, A., Schulman, K., Shao, Y., Qarkaxhija, V., Nandi, A. K., Keddie, S. H., Rumisha, S., Amratia, P., ... Gething, P. W. (2020). Global maps of travel time to healthcare facilities. *Nature Medicine*, 26(12), 1835–1838. <https://doi.org/10.1038/s41591-020-1059-1>

Ye, P., Ye, Z., Xia, J., Zhong, L., Zhang, M., Lv, L., Tu, W., Yue, Y., & Li, Q. (2024). National-scale 1-km maps of hospital travel time and hospital accessibility in China. *Scientific Data*, 11(1), 1130. <https://doi.org/10.1038/s41597-024-03981-y>

Yeboah, G., Porto de Albuquerque, J., Pitidis, V., & Ahmed, S. A. K. S. (2021). Analysis of OpenStreetMap data quality at different stages of a participatory mapping process: Evidence from slums in Africa and Asia. *ISPRS International Journal of Geo-Information*, 10(4), 265. <https://doi.org/10.3390/ijgi10040265>