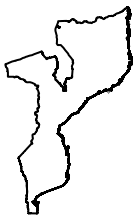
Estimating the catastrophic costs of TB in Mozambique - methods

Alberto García-Basteiro and Joe Brew

Contents



Study objective

Estimate the proportion of households facing catastrophic costs due to TB among those affected by TB.

Approach

Three-stage sampling

- 1st stage units: provinces
- 2nd stage units: districts
- 3rd stage units: health facilities (NTP network)

Stratification

Provinces will be selected taking into account their location (north, centre and south). Districts will be selected according to their type (rural or urban).

The objective of stratification is to improve the representativeness of the sample. We will force the proportion of people diagnosed from TB in each location to be equal to the real proportion. This technique is called proportionate stratification.

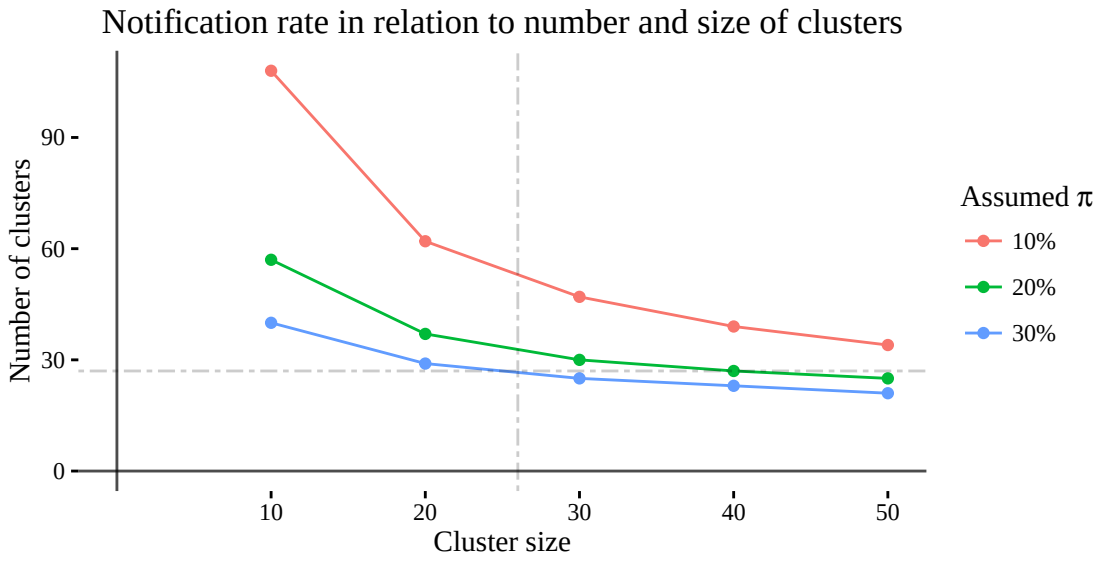
Stratification leads to a smaller sample size, as a result of reducing the variability. But we are not able to estimate this reduction.

Assumptions

Using the formula in the protocol (v.15, section 3.5), we assume:

- A regular cluster sample survey.
- Three different potential notification rates (gamma/prevalences) of households facing catastrophic costs: 10%, 20% and 30%.
- A relative precision (d) of 0.2 (20%), as recommended in the protocol.

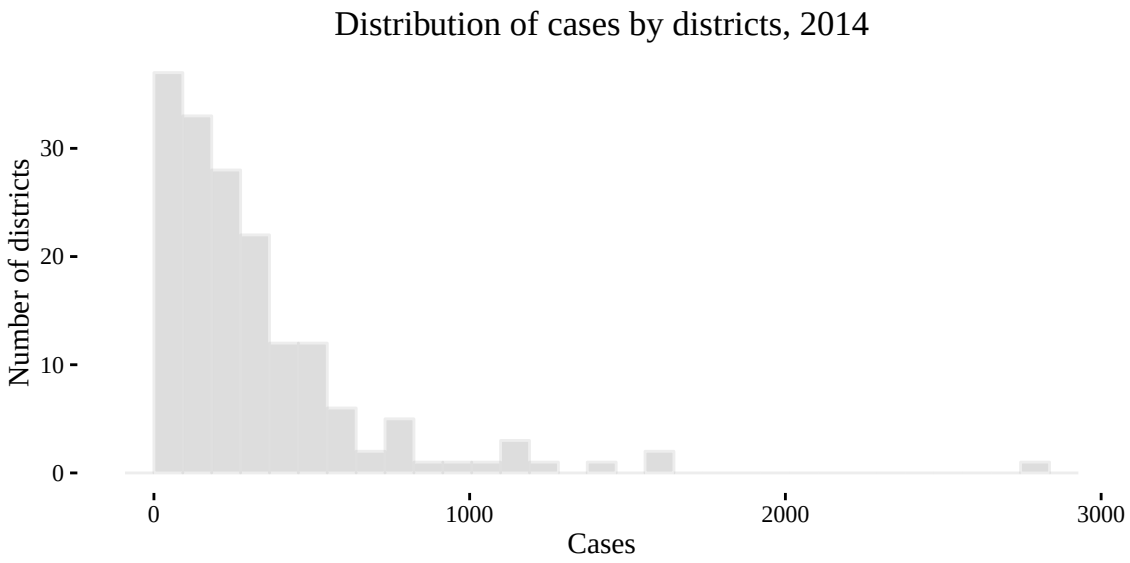
Under these assumptions, our cluster-specific sample size varies as a function to the notification rate and number of clusters:



Given that our target precision is 0.2, and that we wish to limit the number of health facilities to 27 (slightly above the recommended 25), we arrive at a minimum facility-specific sample size of 24, which we buffer to 26 to account for potential data quality issues.

Facility-specific sample size	Number of facilities
20	29
21	28
22	28
23	27
24	27
25	26
26	26
27	26
28	25
29	25
30	25
31	24
32	24
33	24
34	24
35	24
36	23
37	23
38	23
39	23
40	23

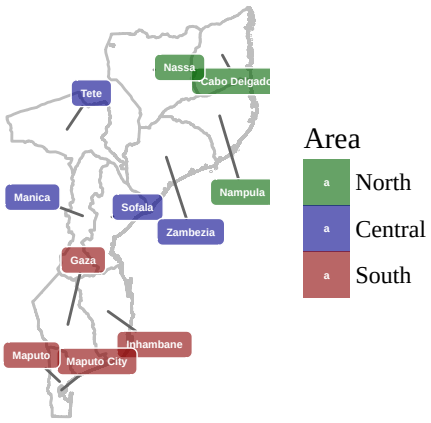
Given that previous studies have suggest a gamma of 30%-50%, we assume a gamma of 30% (blue line in the above chart), suggesting a necessary minimum cluster size of 26 per participating health center. This is feasible, as 100% of provinces saw an average of more than 26 cases per district in 2014 (the last year for which data are available), and many see an annual number of cases far in excess of our minimum:



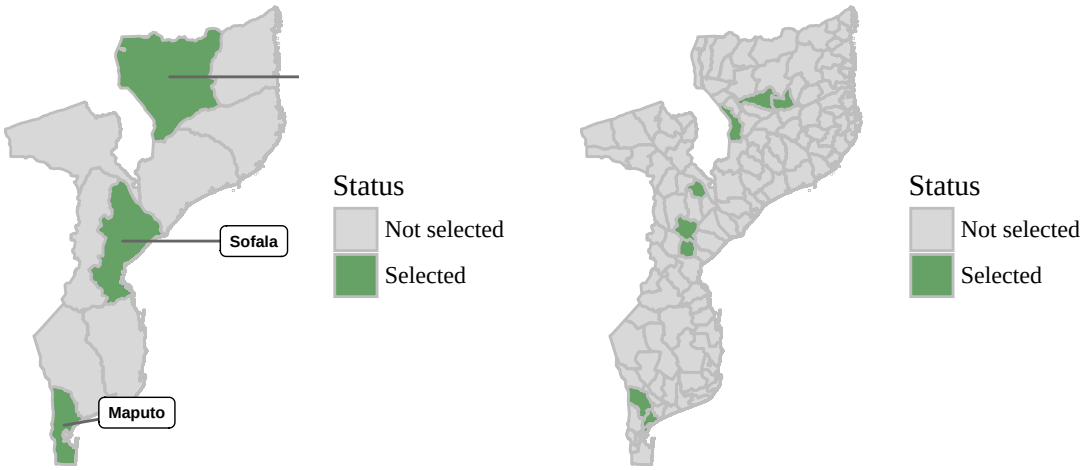
Selection overview

We will employ a multi-staged clustering method which balances the need for geographic representativeness with the economic necessity of limiting the total number of participating health facilities to 27. In order to ensure that our results reflect the true distribution throughout the country, we split our clustering at 3 strata - the "area", province, and district levels.

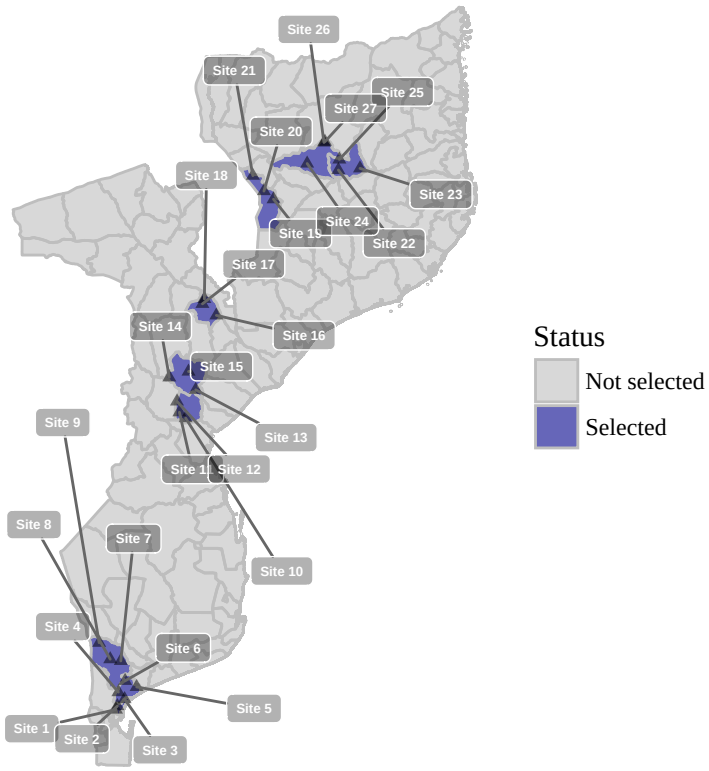
First, we divide the country into three "areas" - north, central, and south:



Each area contains between 3 and 4 provinces. From each area, 1 province will be randomly selected, yielding a total of 3 provinces (left). Provinces contain approximately 8-10 districts. 3 districts will be randomly selected for each province, yielding a total of 9 districts (right):

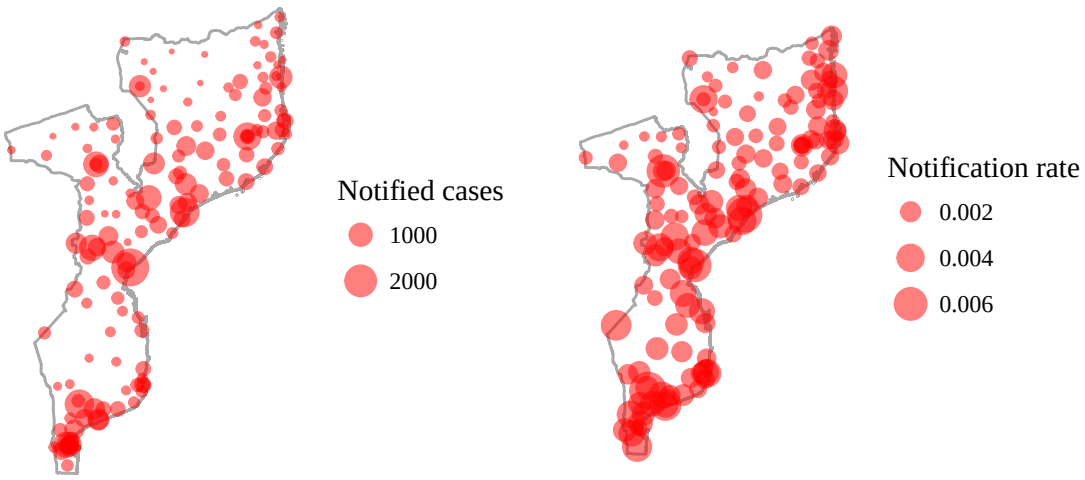


Each district contains 1-2 national TB control offices, and 8-10 peripheral branches. In the final step, 3 offices/branches will be selected from each of the 9 districts, yielding a total of 27 participating sites:



Selection details

The crude number of TB notifications (left) as well as the population-adjusted notification rate for TB (right) vary significantly by geography, as seen below (2014 data):

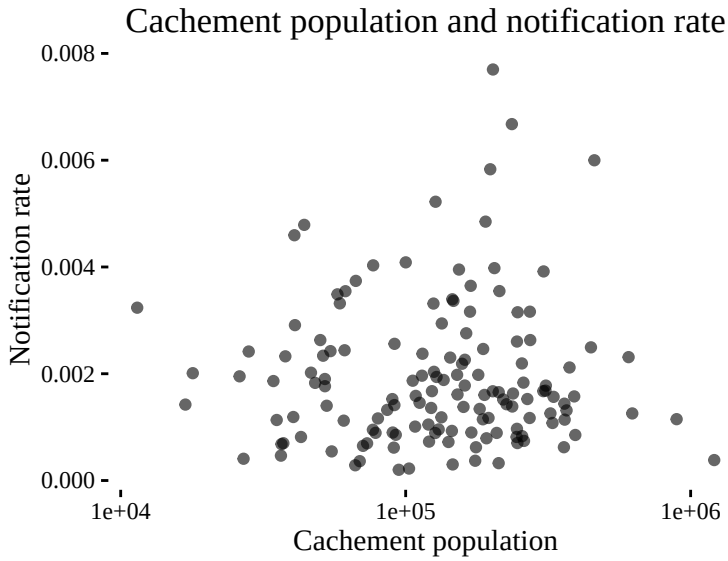


In order to account for heterogeneous notification rates by geography, we follow the methodology of WHO in similar studies. An "initial" randomization assigns the 3 participating provinces and 9 participating districts. From these 9 participating districts, a "secondary" randomization assigns the 27 particpating facilities. The secondary randomization differs from the initial randomization in two important ways:

1. The secondary randomization employs a propability proportional to size (PPS) approach. In other words, assignment is weighted by the number of cases previously detected in the district. The weighting, modeled off previous WHO studies (Myanmar) will be linear, ie a facility's likelihood of being recruited is directly proportional to its size (in terms of likelihood of notififcation).

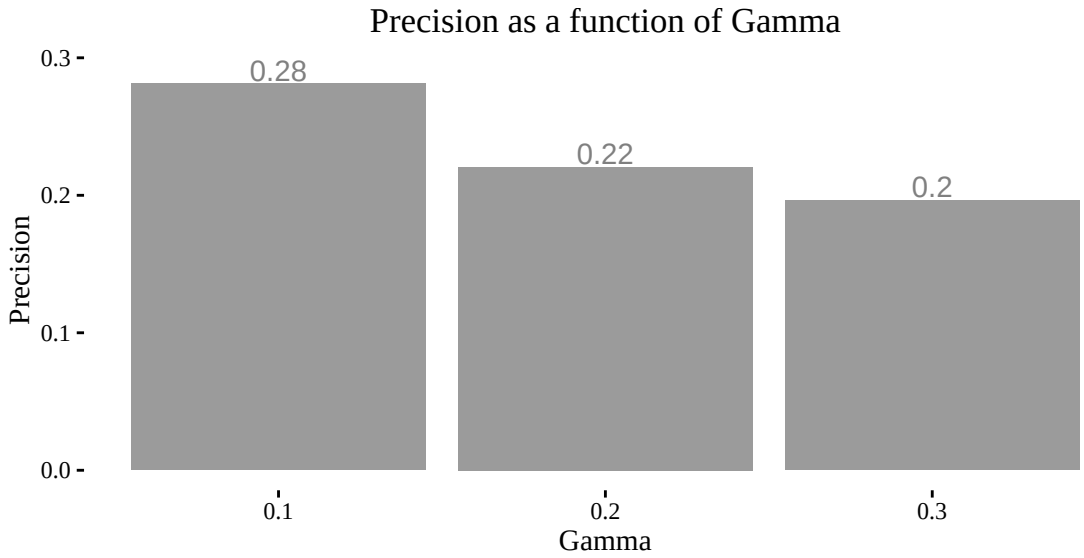
2. The secondary randomization has a simple exclusion criteria - for a facility to be eligible for assignment, it must have a previous annual notification rate of at least 100 cases per year. This virtually ensures that we will arrive to our minimal cluster-specific sample size of 26 in the 4 months during which data collection will take place.

This step is necessary in order to avoid selection of facilities which will not produce enough cases so as to reach the level of statistical significance. It is justifiable given the non-correlation between district cachement area and notification rate (below):



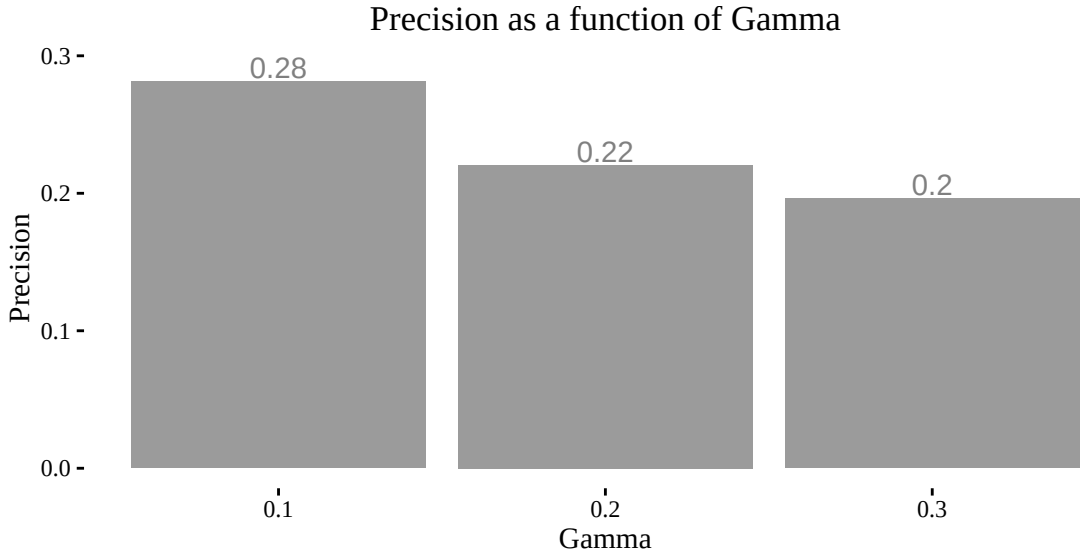
Number of patients

Having established the number of clusters at 27 (previous section) and the minimum cluster size at 26 ("Assumptions" section), we can estimate relative precision and quantify uncertainty at various gamma levels (assuming a 10% non-response rate).



Given that the goal is to reach a precision of approximately 0.2 (20%), and that even our high-end gamma estimate (0.3) is low relative to other similar studies (0.3-0.5), our proposal to recruit 26 patients at each of the 27 sites should be wholly sufficient.

If, on the other hand, we wanted to afix precision at 0.2, we could examine how cluster-specific sample size changes as a function of gamma (the percentage of households facing catastrophic costs):



Limitations

Our approach is not without limitations.

Representativeness: The cost of training workers is a function of the number of sites (rather than the number of workers), thereby requiring a multi-tiered clustering approach. In order to arrive at minimal cluster-specific sample size (26), we have to only include facilities with a higher case load (100 annual notifications). This will bias selection in favor of facilities serving larger cachement areas, or more endemic regions. Though we believe that the clustering methodology will yield results which are generalizable at the national scale, significant inter-district variability - or simply a "bad draw" - could bias results.

Timeliness: Our population and previous notification rates information are based on previous prevalence data. Though the correlation between preveious trends and the present is likely tight, migratory activity and epidemiological change may mean that our sampling strategy is no longer the right fit.

Recruitment assumptions: We assume a 90% response rate. A significantly lower response rate could mean difficulty in meeting the cluster-specific minimum recruitment numbers in the study time period allotted (this likely only applies at the least populous facilities).

Facility-based recruitment: In order to recruit the number of TB patients necessary for meaningful analysis, we rely on the health system infrastructure. Patients who are attended in health facilities are qualitatively different from those that are not treated, or treated by alternative means. This could have important negative implications for the generalizability of our results.

Details

The data and code for the calculations mentioned herein can all be found at http://github.com/joebrew/tb_catastrophic_costs.

Aina Castellas performed the original calculations and wrote the code on the multitude of potential scenarios and strategies that allowed us to hone in on the approach laid out in this document.