scientific reports

Check for updates

OPEN

Data-driven identification of communities with high levels of tuberculosis infection in the Democratic Republic of Congo

Mauro Faccin[1]✉, Olivier Rusumba[2], Alfred Ushindi[2], Mireille Riziki[2], Tresor Habiragi[2], Fairouz Boutachkourt[3] & Emmanuel André[4,5]

When access to diagnosis and treatment of tuberculosis is disrupted by poverty or unequal access to health services, marginalized communities not only endorse the burden of preventable deaths, but also suffer from the dramatic consequences of a disease which impacts one's ability to access education and minimal financial incomes. Unfortunately, these pockets are often left unrecognized in the flow of data collected in national tuberculosis reports, as localized hotspots are diluted in aggregated reports focusing on notified cases. Such system is therefore profoundly inadequate to identify these marginalized groups, which urgently require adapted interventions. We computed an estimated incidence-rate map for the South-Kivu province of the Democratic Republic of Congo, a province of 5.8 million inhabitants, leveraging available data including notified incidence, level of access to health care and exposition to identifiable risk factors. These estimations were validated in a prospective multi-centric study. We could demonstrate that combining different sources of openly-available data allows to precisely identify pockets of the population which endorses the biggest part of the burden of disease. We could precisely identify areas with a predicted annual incidence higher than 1%, a value three times higher than the national estimates. While hosting only 2.5% of the total population, we estimated that these areas were responsible for 23.5% of the actual tuberculosis cases of the province. The bacteriological results obtained from systematic screenings strongly correlated with the estimated incidence (r = 0.86), and much less with the incidence reported by epidemiological reports (r = 0.77), highlighting the inadequacy of these reports when used alone to guide disease control programs.

Worldwide, it is estimated that up to 4 million patients affected by active tuberculosis disease (TB) are left untreated every year. In Africa, up to 50% of patients requiring care are left undiagnosed today, and while drug-resistant TB remains a major concern, there are probably more transmission and deaths related to TB under-detection than due to drug resistance[1].

The direct and indirect costs supported by these millions of people in need of care are both the cause and the consequence of the socio-economic impact of TB on the poorest populations[2–4]. In this sense, inadequate access to care is a major contributor to the dramatic spiral of poverty and uncontrolled disease transmission.

Recognizing the importance of this problem, the World Health Organization (WHO) recommends performing systematic TB screening in high-risk communities, and since a decade, numerous "active case finding" (ACF) pilot projects have been supported and implemented in high burden countries. The very irregular level of efficacy of these programs is intrinsically related to the difficulty to identify pockets of the population with the highest burden of disease. Performing systematic screening in a population with low incidence or having a facilitated

[1]ICTEAM, Université Catholique de Louvain, Louvain-la-Neuve, Belgium. [2]Ambassadeurs de la lutte contre la Tuberculose, Bukavu, Democratic Republic of Congo. [3]Institut de Recherche Expérimentale et Clinique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium. [4]Department of Microbiology, Immunology and Transplantation, Katholieke Universiteit Leuven, Leuven, Belgium. [5]Laboratoriumgeneeskunde, Universitair Ziekenhuis Leuven, Leuven, Belgium. ✉email: mauro.fccn@gmail.com

access to health services outside screening campaigns will lead to very limited additional cases found. Additionally, it will typically impact staff motivation and require difficultly scalable human and financial resources[1–6].

Pushed by the necessity to achieve a significant yield, a typical recommendation is to perform systematic screening in households of TB patients and among people living with HIV[7–9]. Although these indications are well proven to be useful, restricting systematic screening to these very limited groups will structurally miss the opportunity to detect the majority of TB cases, in particular in the context of the structural under-diagnosis of TB and HIV such as experienced in the DRC[10–12]. It has further been described that within households of active TB cases, the source of new infections is very often different from the presumed index case[13]. These observations illustrate the need to extend systematic screening beyond the current recommended perimeter, in particular to areas with very high incidence of the disease[14].

These recommendations are in practice inapplicable, as countries lack tools to identify pockets of the population where high level of under-notification hides dramatic incidence rates of TB. We develop a tool which would precisely identify the pockets of the population where the majority of TB cases are undetected. In this context, such tool would allow guiding highly efficient ACF interventions, and would allow avoiding over- or under-utilization of community workers and medical resources[15].

Methods

We introduce a new approach to ACF planning which is in two-folds: it combines a data-driven detection of high-incidence pockets of the disease and a digital assessment of the individual risk as triage.

Firstly, we collect openly-available datasets that describe environment characteristics such as the population density, the presence of the local health care system or the closeness to mining sites. A data-driven prediction algorithm combines these datasets with the information from the local TB reports to precisely identify localized pockets of very high circulation of TB which can be defined as an incidence rate above 1000/100,000 (1%). A representation of these calculations on a map of the South-Kivu province of the Democratic Republic of Congo (DRC) is illustrated in Fig. 1. In this map, color codes represent the estimated incidence rate for active pulmonary tuberculosis. This map illustrates the significant variations in the predicted incidence within the province: while the vast majority of the surface of the province shows predicted incidence rates below 0.1%, only very limited areas actually show a risk of above 1%. In South-Kivu, the share of population living in high-risk zones, with a predicted incidence rate above 1%, is only about 2.5%. This same population is expected to host more than 20% of active TB cases.

Secondly, in the communities highlighted by the estimated incidence rate an ACF intervention is performed with the aid of a digitally supported questionnaire and the MediScout application stack as triage.

Study setting and study design. We performed a multi-centric prospective study in the South-Kivu province of the DRC, a region at the border with Rwanda and Burundi and facing a high burden of TB, partly due to a situation of over 20 years of conflicts and population displacements. As most eastern provinces of DRC, South-Kivu has important artisanal and industrial mining sectors. The South-Kivu province hosts a population of 5.8 million inhabitants divided in 34 health zones. In total, 113 health facilities providing basic TB diagnostic and treatment services.

We evaluated two main elements in this study.

The first element under evaluation is the map of predicted incidence rate that represents a risk measure for the local communities. The major outcome is the ability to segregate communities with a high rate of TB ($> 1\%$) from other communities not systematically eligible for ACF interventions as per the WHO criteria. To do so, we included 11 urban, semi-urban and rural communities distributed in 11 health zones of the province. The choice of these communities was made in order to cover a wide range of TB predicted incidence, ranging from 0.1 to 2.3% (see Table 1). Some zones at high predicted risk such as Itombwe and Minembwe had to be excluded due to ongoing insecurity issues. We included in this study remote communities such as Matili and Lugushwa which were only accessible using local planes combined by 2 days of travel on a motorcycle. Table 2 provides detailed information on the notified incidence rate, the number of cases reported by the local health facilities divided by the population covered by such facilities, and the predicted incidence rate for each of these communities. In this table, we observe sensible differences between the notified incidence rate and the predicted incidence rate, the predicted incidence rate being up to ten times higher than the notified incidence rate in the same area. An example of this situation is the city of Shabunda, a rural and mining area with a very low coverage of health structures.

The second element under evaluation is the individual risk assessment tool represented by the MediScout mobile app, including the built-in questionnaire (see Fig. 2). The questionnaire, based on the weighted evaluation of TB-related individual risks, such symptoms, exposure and personal history, acts as a digital-supported triage system.

We partnered with a local organization experienced in TB ACF and trained a project coordinator and 20 research assistants, involved in community-outreach activities, for the utilization of the MediScout applications. This training included the initiation of ACF missions, individual screenings and referral of patients to local health clinics and was held in Bukavu on the 20th and on the 21st of March 2019. These screeners, originating from the provincial capital city Bukavu, visited all study areas, where they were accompanied by local community health workers to facilitate the collaboration of the community. The data collection was performed during the period from Mar 2019 to Jan 2020.

Relative incidence rate prediction. We gathered openly available data linked to the risk of TB to compute a relative risk level for each precise location. The different data sources and their utilization in our model are described hereunder. All datasets were retrieved in January 2019.

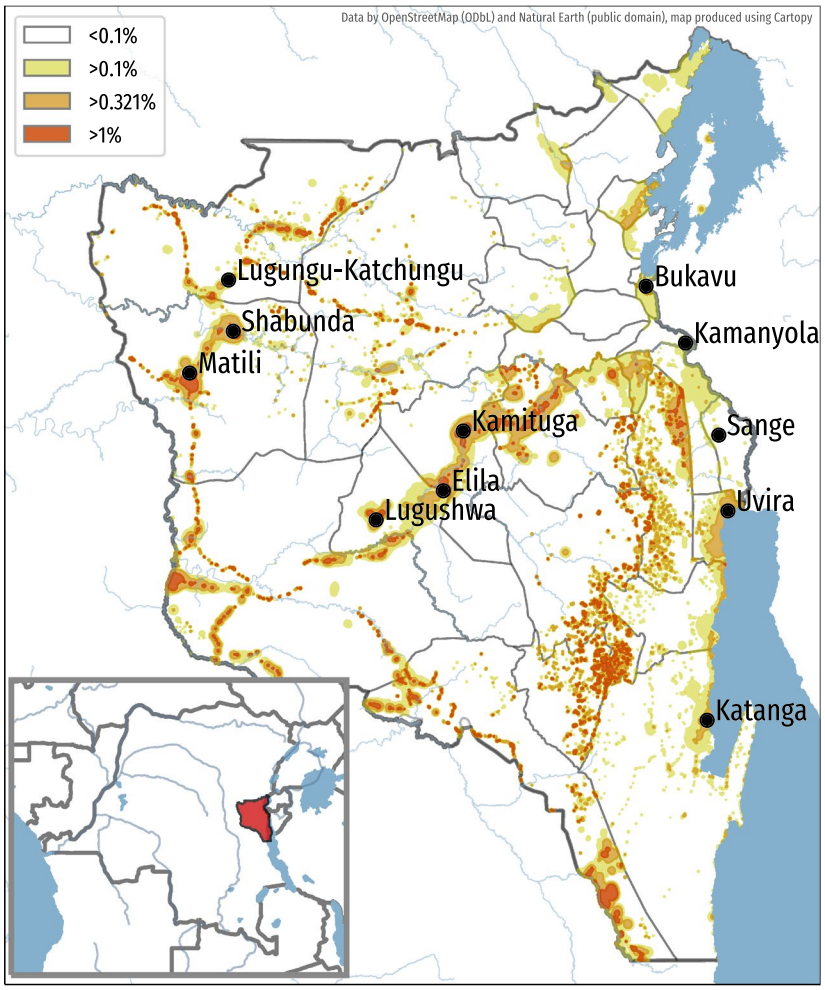

Figure 1. Mapping of TB predicted incidence rate for the South-Kivu province of the Democratic Republic of Congo (inlay). No color: below 0.1%; Yellow : between 0.1% and national average (0.322%); Orange : between 0.322% and 1%; Red : above 1%. Geospatial data provided by OpenStreetMap (under ODbL license) and Natural Earth (public domain), map produced using Cartopy `v0.20` from https://scitools.org.uk/cartopy.

	Predicted incidence (%)	Number of screenings	Average individual score	Laboratory confirmed	Laboratory tested
Bukavu	0.162	2156	3.365	4	56
Sange	0.131	225	4.311	0	39
Uvira	0.664	475	5.318	1	79
Katanga	0.427	1174	3.590	6	121
Kamanyola	0.107	226	4.049	2	30
Kamituga	1.129	1965	6.721	15	135
Lugushwa	0.936	1383	7.120	19	129
Elila	1.149	498	8.040	8	58
Shabunda	2.341	1254	4.910	29	NA
Lugungu-Katchungu	2.192	662	5.735	19	NA
Matili	2.272	255	3.902	9	NA

Table 1. Project statistics disaggregated by location.

Communities (Health zone)	Reported incidence (2017)	Estimated incidence	Current detection rate (estimated)
Kamanyola (Nyangezi)	100	110	91%
Sange (Ruzizi)	40	130	31%
Bukavu (Ibanda, Kadutu, Bagira)	100	160	63%
Katanga (Fizi)	150	430	35%
Uvira (Uvira)	170	660	26%
Lugushwa (Kitutu)	170	940	18%
Kamituga (Kamituga)	240	1129	21%
Elila (Kitutu)	260	1149	23%
Lugungu-Katchungu (Lulingu)	250	192	11%
Matili (Shabunda)	260	272	11%
Shabunda (Shabunda)	170	341	7%

Table 2. Distribution of reported and estimated incidence on the communities interested in this work.

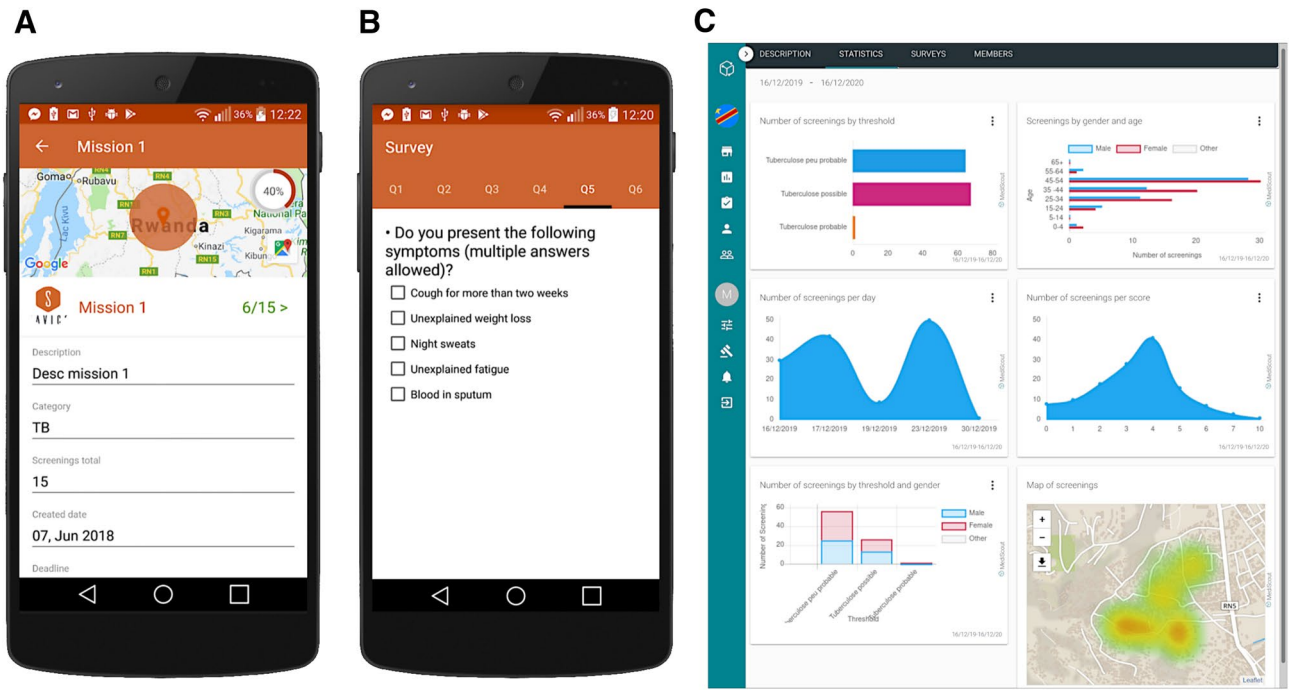

Figure 2. The Mediscout mobile and web applications. (**A**) Mobile application with interface for mission acceptance by the community health-worker. (**B**) Questionnaire in the mobile application. (**C**) Statistics of the missions in the web application (note the mapping of the screenings at the bottom).

Population distribution and density information was extracted from the Worldpop project[16,17], an initiative which provides an estimation of the population density with a resolution of approximately 100 m. These estimations result from combined geospatial datasets with available aggregated count data[18]. Further, additional information on the location of urban and residential environments was gathered from the OpenStreetMap project[19].

We then combined the demographic data with the most granular level of epidemiological surveillance, being, in the context of DRC where this study took place, the Health Zone quarterly notification reports. These were used to make a baseline distribution of the TB cases. For this baseline distribution, we assumed areas with low population density have a lower TB incidence-rate compared to higher density areas. Locations with a population density lower than 10 people per square kilometer were ignored in this distribution model.

We further used the location and type of health facilities extracted from the Global Healthsites Mapping Project[20] in collaboration with OpenStreetMap[19]. We used the distance between each community and the nearest health facility to estimate the phenomenon of under-detection of TB, which is correlated to the difficulties for

Item evaluated during questionnaire	Score	
Unexplained fatigue	1	Symptoms (maxscore = 10)
Unexplained weight loss	1	
Night sweats	1	
Temperature above 38C	1	
Cough (< 15 days)	1	
Cough (> 15 days)	2	
Blood in sputum	2	
Chest pain	1	
Close and persistent contact	2	Individual TB exposure (max score = 8)
Occasional close contacts	1	
Contacts in poorly ventilated area	1	
Contacts occurred less than 2 years ago	1	
Living or lived in military camp	1	
Living or lived in mining camp	1	
Living or lived in prison	1	
History of abandoned treatment	1	Individual TB history (max score = 2)
History of failed treatment	1	

Table 3. Elements and corresponding weights used to evaluate individual risk for TB in the questionnaire.

each community to access health services, assuming that "far" from health facilities, 50% of the real TB cases are missed. Furthermore, the type of health facility, local clinical versus hospitals, was also used in the risk estimation.

Finally, silica exposure, in particular when linked to mining activities, is a risk factor for TB[21–23]. Since the South-Kivu province, where this study took place, has an important mining activity, we integrated in the risk computation the location and size of mines, accessible through the IPIS Research project[24] and OpenStreetMap[19]. The predicted risk of TB was correlated with the presence of mines in the same geographical areas. We assumed that in mining environments, the incidence rate of TB cannot be lower than of 0.5% or 500/100,000.

Individual risk prediction. The main output that was to be achieved by the individual assessment questionnaire was to be able to yield a positivity rate greater than 10% among people identified as being at high risk for active TB.

To build this individual assessment tool, we took into consideration previous observations highlighting that individual risk for TB infection can be extrapolated from a combination of elements including symptoms and the type of exposition to TB[15].

Our questionnaire for evaluating the individual risk for TB was therefore based on the presence or absence of several TB-related symptoms and the precise context of an eventual exposition to TB. The elements captured and the weight given to each element is described in Table 3. The survey generates a total score in the range 0–20. In this study, all people with a score higher than or equal to 4 and those with a cough (regardless of the total score) were referred for a laboratory test.

MediScout solution. The MediScout web application (developed by Savics, Belgium) was used as an integrated end-user interface allowing to visualize the incidence prediction map and plan geographically localized ACF campaigns. The MediScout mobile application (Savics, Belgium) was used to guide community-based health workers in the restricted area of interest, while receiving from the remote web application ACF missions to perform. Further, the app was used to go through the individual questionnaires when in direct contact with the study participants, to automatically compute the individual risk and send the results of these questionnaires back to the remote server for later analysis, see Fig. 2 for a visualization (Fig. 2A: mobile application mission interface, Fig. 2B: mobile application questionnaire interface). The MediScout mobile app allows uploading information to the MediScout web interface when internet or 3G connectivity were available. This system allows full traceability of all community-based screening events, including GPS location, demographics of the population covered and individual scores.

Based on these reports, aggregated statistics and geographic locations are reported automatically and in real-time (Fig. 2C: web app interface with example of data visualization tools and intervention mapping at the bottom).

Cases distribution from population density. The cases reported by the local health system are disaggregated according to the local population density inspired by the equilibrium solution of a simple compartmental model (SIS) of an endemic infected population[25]:

$$r(x) \propto p(x)\frac{p(x) - p_0}{p(x) + p_1 - 2p_0}$$

Estimated incidence rate	Participants with score < 4	Participants with score ≥ 4	Total number of participants
< 1%	55.5% (4614)	44.5% (3708)	8322
≥ 1%	32.5% (1792)	67.5% (3727)	5519

Table 4. Participant distribution over high and low predicted incidence areas.

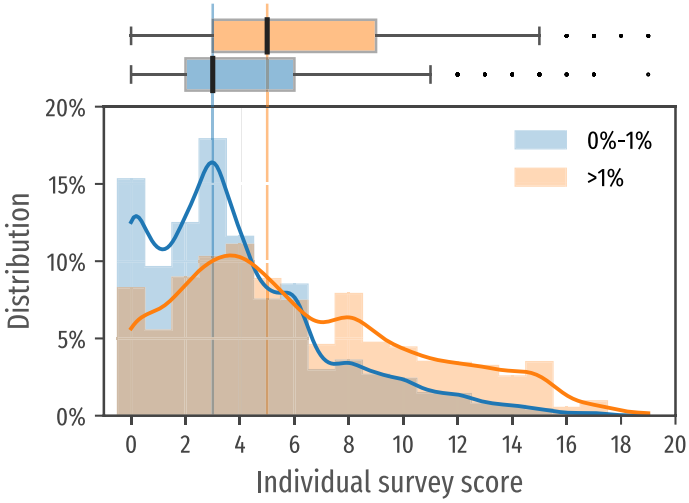

Figure 3. Distribution of individual risk for the population located in low risk zones (blue) where the predicted incidence rate is lower than 1% and for the population in high risk zones (orange). Note that the proportion of lower scores (< 4) is higher in the lower risk communities (bottom). Score medians in both subpopulations (3 and 5 respectively) with quartiles are reported in the upper box plot.

where p_0 represents the minimal value of the population density for which the endemic infected population survive. Furthermore, p_1 accounts for the relation between the population density and the transmission parameter of the disease. A normalization is applied in order to recover the same aggregated figures. We choose $p_0 = 10$ inhabitants per square km and $p_1 = 1000$ inhabitants per square km.

Ethical approval. The study and its protocols have been approved by the *Comité Institutionnel d'Ethique de la Santé* of the *Université Catholique de Bukavu* with reference number UCB/CIES/NC/07/2019.

Informed consent. All subjects participating in this study (or their legal guardians) gave its informed consent. All methods were carried out in accordance with the relevant guidelines and regulations.

Results

Performance of the relative incidence rate prediction tool. We used the individual risk-assessment questionnaire to evaluate the performance of the incidence rate prediction algorithm. This questionnaire, based on a combination of TB-related symptoms, personal exposition to TB and personal history of TB, was considered as a good proxy for estimating the level of circulation of TB in a community (see Table 3). Overall, in the 11 study sites, 13841 people agreed to respond to the individual risk-assessment questionnaire. 8322 (60.1% of total) questionnaires originated from areas at low predicted incidence rate (< 1%) and 5519 (39.9%) questionnaires originated from areas at high predicted incidence rate (> 1%) (see Table 4). In areas with low predicted incidence, 55.5% of the responders had a low risk (score < 4) and 44.5% had a high risk for TB (score ≥ 4). In comparison, in areas with higher predicted incidence, 32.5% of the responders had a low risk (score < 4) and 67.5% had a high risk for TB (score ≥ 4).

The individual risk score is comprised between 0 and 20, as is computed based on a diversity of questions including related to the presence of several TB-related symptoms and different elements reflecting the intensity of exposition to TB. We looked at the distribution of these scores in the high and low predicted incidence areas.

The median and the interquartile ranges of the risk score in low or high incidence predicted areas was of 3 (2–6) and 5 (3–9) respectively, highlighting the different, although partly overlapping, distribution of scores in the two populations (p-value = 0.000). Figure 3 shows the distribution of the risk score in the two types of settings.

Figure 3 (top) illustrates how the individual risk score in the two populations at high and low risk follows different distributions (with p-value = 0.000). It also illustrates the high dispersion of risk scores within these areas, reflecting the heterogeneity of individual risk in each community: the individual score median as well as the confidence intervals increase in areas at higher risk.

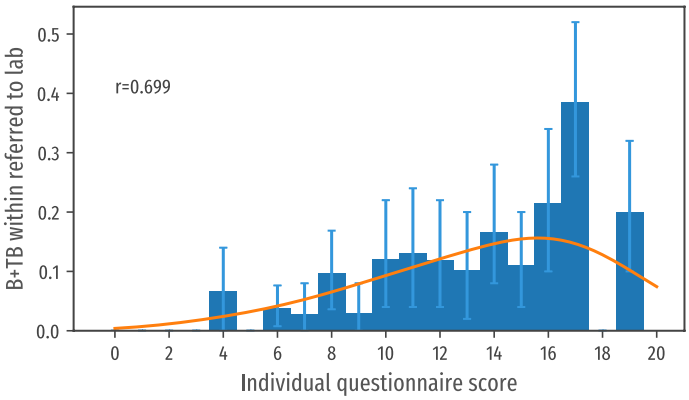

Figure 4. Proportion on laboratory confirmed TB cases per score class.

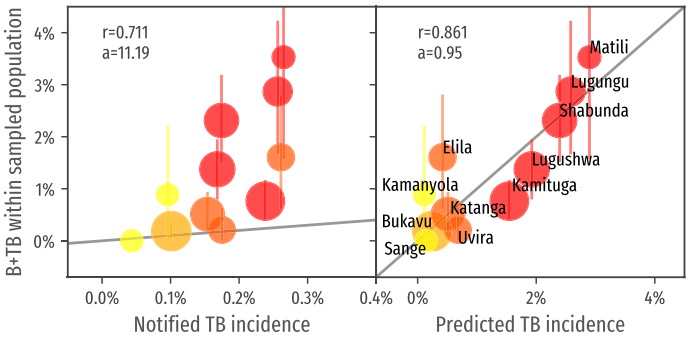

Figure 5. Correlation of the predicted incidence with the measured incidence within sampled population (right). Note the high value of the correlation coefficient *r*. The parameter *a* represent the slope of the fitting line. As a term of comparison, the correlation of the incidence extracted from the health system reports with the measured incidence within the sampled population (left). Note in this case the lower correlation coefficient and the slope of the fitting line (the measured incidence surpass 10x the incidence of reported cases).

Performance of the individual-risk assessment questionnaire. We used Ziehl–Neelsen microscopy as a reference method to evaluate the performance of the individual-risk assessment questionnaire. A method with higher sensibility such as GeneXpert would have been preferred, but, unfortunately, it wasn't available to the local health care system.

In total 1153 laboratory tests were performed. Screeners were trained to suggest a laboratory test only among people presenting a cough (regardless of the total score) or having an individual risk score ≥ 4. Within the performed tests, 112 individuals were diagnosed with laboratory-confirmed pulmonary TB (9.7% positivity). Unfortunately, some laboratory outcomes in few remote facilities resulted unlinked from the corresponding questionnaire and were discarded from this analysis (we know only the aggregated picture of 449 negative tests within Matili, Lugungu–Katchungu and Shabunda). The positivity rate was 8.8% (n = 55/626) for those with a score ≥ 4 and 12.3% for those with a score ≥ 8 (n = 48/389).

No positive laboratory results were reported among people with a score lower than 4, regardless of the presence of cough (21 lab tests). The proportion of positive smears increased together with the risk score as illustrated in Fig. 4, supporting the effectiveness of the mobile questionnaire as triage.

Performance of the integrated prediction as a policy-decision tool for ACF disease control programs. To assess the performance of MediScout as a technical tool that can be used by TB-program managers and community-outreach organizations to optimize the efficiency of ACF interventions, we compared the predicted incidence rate with the reported incidence of bacteriologically-confirmed pulmonary cases.

Although the populations originating from areas with high risk (> 1% predicted incidence) and areas with low risk were of similar size, of the 112 individuals diagnosed with bacteriologically-confirmed pulmonary TB, 91 (81.25%) originated from the predicted zones at very high risk and 21 (18.75%) originated from zones predicted to have an incidence rate <1%.

Figure 5 (right) shows a strong association (r = 0.861) between the predicted incidence rate and the proportion of bacteriologically-confirmed pulmonary TB (B + TB) cases identified, with a nearly linear correlation between the predicted incidence and the observed yield of the ACF interventions (with a fitting line with a slope of 0.95).

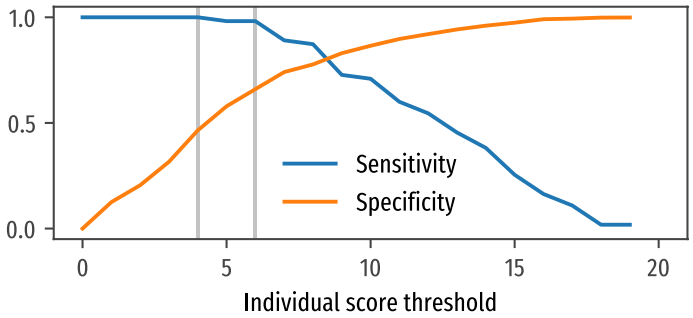

Figure 6. Expected sensitivity and specificity if the threshold to refer subjects to lab is changed to a different value. The vertical lines correspond to a threshold of 4 and 6. A threshold of 4 is maybe too conservative; in other contests, one can safely use 6.

In comparison, Fig. 5 (left) illustrates a much less performant correlation between the historically notified cases and the actual incidence observed in this study associated with the heterogeneous level of under detection of TB.

Choice of individual risk threshold. In the present prospective study, the community health workers suggested single individuals to refer to a laboratory for a microscopy test if their individual-risk score reached at least the conservative threshold value of 4. The choice of such conservative value is dictated by the specific testing aims of the current study. We computed the sensitivity and specificity of the survey based on different thresholds (see Fig. 6). The chosen threshold presents a sensitivity = 1.0 (no screening with individual score below 4 where confirmed positive to TB in laboratory) and a specificity = 0.46. The choice of a threshold value equal to 6 would still have high sensitivity = 0.98 while increasing the specificity = 0.67 (and decreasing the number needed to test in laboratory to find a positive case). Applying a threshold at 6 would have saved 2692 (36%) tests, but only 1 case ($\sim$ 0%) would have been missed.

Discussion

One of the major limitations of this study is that the diagnosis of TB relied uniquely on series of three Ziehl-Neelsen microscopy tests per patient, the only method available to the local health system. This technique probably underestimated the level of active TB disease as it is known to have a limited sensitivity and would therefore miss a proportion of paucibacillary infections. The real incidence may thus have been higher than reported in this study. Despite this limitation, we believe that performing this evaluation in a real-life setting, which includes the technical limitations experienced by health workers, is probably more informative than a study which would create an artificial bias by including technologies which are currently inaccessible for the vast majority of the population.

Very interestingly, we could show that using data-driven predictions and setting a threshold at a predicted incidence of 1% allows prioritizing ACF interventions in well-circumcised pockets of the population where over 80% of the cases, found through ACF, reside. This illustrates that in setting with uncontrolled transmission, the majority of the people experiencing active TB disease will not access the health system if they are not actively identified and supported by outreach interventions.

Determining and disentangling the real factors that trigger the spread of the disease in each case may be an unachievable task. Indeed, our prediction grasps the most important factors that, in this framework, include poverty or difficulty to access to health care[12] (two highly entangled factors) and closeness to mining activities[21–23]. The efficiency of this data-driven approach emerges from the comparison to similar ACF activities (in mining communities). In particular, in location predicted to be at high risk, the number needed to be screened reached 61 (5519 screenings/91 positive cases) while the number needed to test was as low as 9 (770 tested/91 positive cases). The same statistics in similar settings on other works are considerably higher (110 and 25[23], 151 and 12[10]).

Another major consequence of the described findings is that historical notification reports may fail in uniformly sample the disease incidence. Planning and prioritization of TB control interventions should, therefore, not rely only on those reports. In such cases, the decision process would contribute to under-estimate the actual level of disease transmission in the most vulnerable pockets of the population without or with limited access to the health system. This study highlights that these historical notifications reports should be integrated with demographic, geographical and social data in order to optimally inform public health authorities and funders about where to prioritize the implementation of complementary interventions.

Current notification-based epidemiological surveillance approaches have structurally failed to quantify the level of tuberculosis under-detection. The roots of this problem are multiple and include the fact that populations affected by a high burden of tuberculosis are concomitantly affected by a multitude of other poverty-related problematics in particular facilitated access to prevention and curative health services. In this study, we demonstrate that this major problem, that prevents any TB-eradication program to achieve its objective, can be overcome by analyzing TB notification data together with other publicly-available data such as population density, environmental risk factors for tuberculosis and socio-economic indicators. These results support a One Health approach to tuberculosis control, which aims to integrate the patient and his disease in a broader context.

Received: 10 September 2021; Accepted: 22 February 2022
Published online: 10 March 2022

References

1. WHO. *Global Tuberculosis Report 2019*. (World Health Organization, 2019).
2. Ghazy, R. M. *et al.* A systematic review and meta-analysis of the catastrophic costs incurred by tuberculosis patients. *Sci. Rep.* https://doi.org/10.1038/s41598-021-04345-x (2022).
3. Lu, L. *et al.* Catastrophic costs of tuberculosis care in a population with internal migrants in china. *BMC Health Services Res.* https://doi.org/10.1186/s12913-020-05686-5 (2020).
4. Timire, C. *et al.* Catastrophic costs among tuberculosis-affected households in Zimbabwe: A national health facility-based survey. *Trop. Med. Int. Health.* **26**, 1248–1255. https://doi.org/10.1111/tmi.13647 (2021).
5. Chen, J.-O. *et al.* Role of community-based active case finding in screening tuberculosis in Yunnan Province of China. *Infect. Dis. Poverty*. https://doi.org/10.1186/s40249-019-0602-0 (2019).
6. Su, Y. *et al.* Tracking total spending on tuberculosis by source and function in 135 low-income and middle-income countries, 2000–2017: A financial modelling study. *Lancet Infect. Diseases.* **20**, 929–942. https://doi.org/10.1016/s1473-3099(20)30124-9 (2020).
7. Pande, T. *et al.* Finding the missing millions: Lessons from 10 active case finding interventions in high tuberculosis burden countries. *BMJ Glob. Health.* **5**, e00383. https://doi.org/10.1136/bmjgh-2020-003835 (2020).
8. Hanson, C., Osberg, M., Brown, J., Durham, G. & Chin, D. P. Finding the missing patients with tuberculosis: Lessons learned from patient-pathway analyses in 5 countries. *J. Infect. Diseases.* **216**, S686–S695. https://doi.org/10.1093/infdis/jix388 (2017).
9. Burke, R. M. *et al.* Community-based active case-finding interventions for tuberculosis: A systematic review. *Lancet Public Health.* **6**, e283–e299. https://doi.org/10.1016/s2468-2667(21)00033-5 (2021).
10. André, E. *et al.* Patient-led active tuberculosis case-finding in the democratic republic of the Congo. *Bull. World Health Organ.* **96**, 522–530. https://doi.org/10.2471/blt.17.203968 (2018).
11. Sinha, P., Shenoi, S. V. & Friedland, G. H. Opportunities for community health workers to contribute to global efforts to end tuberculosis. *Glob. Public Health.* **15**, 474–484. https://doi.org/10.1080/17441692.2019.1663361 (2019).
12. André, E. *et al.* Prediction of under-detection of pediatric tuberculosis in the Democratic Republic of Congo: Experience of six years in the South-Kivu Province. *PLoS One*. https://doi.org/10.1371/journal.pone.0169014 (2017).
13. Buu, T. N. *et al.* Tuberculosis acquired outside of households, rural Vietnam. *Emerg. Infect. Dis.* **16**, 1466–1468. https://doi.org/10.3201/eid1609.100281 (2010).
14. Koura, K. G., Trébucq, A. & Schwoebel, V. Do active case-finding projects increase the number of tuberculosis cases notified at national level?. *Int. J. Tuberc. Lung Dis.* **21**, 73–78. https://doi.org/10.5588/ijtld.16.0653 (2017).
15. Saunders, M. J. *et al.* A household-level score to predict the risk of tuberculosis among contacts of patients with tuberculosis: A derivation and external validation prospective cohort study. *Lancet Infect. Diseases.* **20**, 110–122. https://doi.org/10.1016/s1473-3099(19)30423-2 (2020).
16. Tatem, A. Worldpop. https://www.worldpop.org (2021).
17. Tatem, A. J. WorldPop, open data for spatial demography. *Sci. Data*. https://doi.org/10.1038/sdata.2017.4 (2017).
18. Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0107042 (2015).
19. OpenStreetMap Foundation. Openstreetmap. https://www.openstreetmap.org (2021).
20. Global Healthsites Mapping Project. Healthsites.io. https://healthsites.io (2021).
21. Stuckler, D., Basu, S., McKee, M. & Lurie, M. Mining and risk of tuberculosis in Sub-Saharan Africa. *Am. J. Public Health.* **101**, 524–530. https://doi.org/10.2105/ajph.2009.175646 (2011).
22. Gottesfeld, P., Andrew, D. & Dalhoff, J. Silica exposures in artisanal small-scale gold mining in Tanzania and implications for tuberculosis prevention. *J. Occupat. Environ. Hygiene.* **12**, 647–653. https://doi.org/10.1080/15459624.2015.1029617 (2015).
23. Ohene, S.-A., Bonsu, F., Adusi-Poku, Y., Dzata, F. & Bakker, M. Case finding of tuberculosis among mining communities in Ghana. *PLoS One.* **16**, e0248718. https://doi.org/10.1371/journal.pone.0248718 (2021).
24. International Peace Information Service. IPIS. https://www.ipisresearch.be (2021).
25. Barrat, A., Barthelemy, M. & Vespignani, A. *Dynamical Processes on Complex Networks* Vol. 1 (Cambridge University Press, 2008).

Acknowledgements

M.F. was partially funded by Innoviris Grant number D1.31402.007. M.F.'s current affiliation is Institut de Recherche pour le Développement (IRD) and Centre Population et Développement (CEPED), Université de Paris, France.

Author contributions

M.F.: Study design, data analysis, writing; O.R.: Study design, data collection; A.U., M.R., T.H.: Data collection; F.B.: Data collection, data analysis; E.A.: Study design; conceptualization; formal analysis; writing.

Competing interests

E.A. provided strategic advice to Savics from October 2019 to March 2020. The rest of the authors have no conflict of interest.

Additional information

Correspondence and requests for materials should be addressed to M.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.