Supplementary appendix

How many of persistent coughers have pulmonary tuberculosis? A population-based cohort study in Ethiopia

Abiot Bezabeh Banti[1,2], Daniel Gemechu Datiko[3], Sven Gudmund Hinderaker[2], Einar Heldal[4], Mesay Hailu Dangiso[6], Gebeyehu Assefa Mitiku[7], Richard Aubrey White[4] and Brita Askeland Winje[4,5*]

File S1 Strobe checklist

File S2 Adult symptom screening questionnaire

File S3 Statistical method

Table S1 Risk factors for all-type pulmonary tuberculosis in Dale, Ethiopia, 2016–2017

Table S2 Comparison of current versus previous study in six kebeles included in the current study 5-years earlier.

File S1 Strobe checklist

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation	
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	4
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	6
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed	na
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	7
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	7
Bias	9	Describe any efforts to address potential sources of bias	8
Study size	10	Explain how the study size was arrived at	6
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	7
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	7
		(*b*) Describe any methods used to examine subgroups and interactions	7
		(*c*) Explain how missing data were addressed	7
		(*d*) If applicable, explain how loss to follow-up was addressed	
		(*e*) Describe any sensitivity analyses	
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	8
		(b) Give reasons for non-participation at each stage	8
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	8
		(b) Indicate number of participants with missing data for each variable of interest	10/13
		(c) Summarise follow-up time (eg, average and total amount)	8
Outcome data	15*	Report numbers of outcome events or summary measures over time	8
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	10/13

		(*b*) Report category boundaries when continuous variables were categorized	10/13
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	8/9
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	8/Suppl
Discussion			
Key results	18	Summarise key results with reference to study objectives	14
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	3+14
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	14-17
Generalisability	21	Discuss the generalisability (external validity) of the study results	17
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	18

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

File S2. Adult symptom screening questionnaire

APPENDIX 2: Adult – tuberculosis symptom based questionnaire – Sidama cluster-study

Questionnaire n°. ☐☐☐☐ **Name of interviewer** _____

Date ☐☐.☐☐.☐☐ (dd.mm.yy)

_____

1. Socio-demographic variables

1.1 Name _____ 1.2. Age: ☐☐.☐☐.☐☐

1.3. Gender (male=1, female=2) ☐ 1.4.Cellphone. ☐☐☐☐☐☐☐☐☐

1.5. Address: Woreda _____Kebele _____Village _____

1.6. Marital status: (single=1, married=2, divorced=3, widowed=4, other=9) ☐

If other, specify: _____

1.7. Total number of people in the household: ☐☐

Among them, number of children < 5 years of age: ☐☐

Their age(s): ☐☐, ☐☐ , ☐☐ , ☐☐ , ☐☐

1.8. Highest grade completed education: ☐☐,

No schooling (yes=1, no=0. other=9) ☐

If other, specify _____

1.9. Occupation:

(farmer=1, housewife= 2, merchant=3, student=4, government employee=5, daily labourer=6, other=9) ☐

If other,(specify) _____

2. Socio-economic variables

2.1. Number of rooms in the house: ☐☐

2.2. Wall type: (wood with mud/cement or brick=1, wood only=2, other=9) ☐

If other, specify_____

2.3. Roof type: (corrugated iron sheet=1, thatched/leaf=2, other=9) ☐

If other, specify:_____

2.4. Type of fuel for cooking

(electricity=1, kerosone=2, charcoal=3, wood=4, cow doug=5, agriculture by-product=6, no cooking in the household=7, other=9): ☐

If other, specify:_____

Appendix 2: ENGLISH Last updated: 250315

APPENDIX 2: Adult – tuberculosis symptom based questionnaire – Sidama cluster-study

2.5. Light source: Electricity (yes=1, no=0, other=9): ☐

If Other, specify: _____

2.6. Variables for household wealth index (present=1, absent=0, unknown=9)

Household		Present	Absent	Unknown
2.6.1	Separate kitchen in the household			
2.6.2	Cooking room ventilation			
2.6.3.	Heating in the house			
2.6.4.	Radio			
2.6.5.	Television			
2.6.6.	Mobile phone			
2.6.7.	Refrigerator			
2.6.8.	Land for agriculture			
2.6.9.	Bank account			

3. Symptoms of tuberculosis (yes=1, no=0)

Symptoms		Yes	No	Comment (i.e. duration in weeks, dates. enter NK if not known)
3.1.	How long have you been coughing?			
3.2	Is this cough productive of sputum?			
3.3.	Does the sputum contain blood?			
3.4.	Do you have fever?			
3.5.	Do you have night sweats?			
3.6.	Have you lost your appetite?			
3.7.	Have you lost weight?			
3.8.	Do you have chest pain or difficulty of breathing?			
3.9.	Did you visit a health facility for your current illness?			
3.10.	If no in 3.9, reasons for not seeking health care (enter 1 in any applicable): Not knowing that it could be TB ☐, not knowing where to go for care ☐, distance to health facility ☐, having to take transport ☐, getting permission to go for care ☐ costs ☐, other ☐ If other specify:_____			

APPENDIX 2: Adult – tuberculosis symptom based questionnaire – Sidama cluster-study

4. Contact-history and risk-factors (yes=1, no=0)

Contact-history and risk-factors		Yes	No	Comment (i.e. duration in weeks, dates. enter NK if not known)
4.1.	Were you treated for tuberculosis before?			
4.2.	Any TB case in the household in the past 5 years			
4.3.	Did you live with a person who has a chronic cough?			
4.4.	Were you tested for HIV in the past year			
4.5.	Ever alcohol-drinker			
4.6.	Ever chewed Khat			
4.7.	Ever smoker			
4.8.	Smoker currently in the household			
4.9.	Smoker previously in the household			

5. Clinical and diagnostic information

5.1. Height (cm) ☐☐☐ 2.2. Weight (kg) ☐☐.☐ 2.3. MUAC (cm): ☐☐

5.2. BCG-scar (yes=1, no=0, unknown=9): ☐

5.3. Sputum sample collection/s

Date of test I: ☐☐.☐☐.☐☐, II: ☐☐.☐☐.☐☐, III: ☐☐.☐☐.☐☐

5.4. Date of smear result: ☐☐.☐☐.☐☐,

Result (positive=1, negative=0): ☐, if positive, grade: ☐☐☐

5.5. Date of GeneXpert result: ☐☐.☐☐.☐☐,

5.6. GeneXpert detection of *M. tb complex* (yes=1, no=0): ☐

5.7. GeneXpert detection of rifampicin-resistance (yes=1, no=0): ☐

5.8. Date of culture result: ☐☐.☐☐.☐☐,

5.9. Detection of *M. tb complex* (yes=1, no=0): ☐

6. Treatment for tuberculosis disease and follow-up

6.1. Date of registration: ☐☐.☐☐.☐☐ (dd.mm.yy)

6.2. Date of treatment initiation: ☐☐.☐☐.☐☐ (dd.mm.yy)

6.3.Treatment outcome: (cured=1, completed=2, failure=3, death=4, default=5, transfer out=6) ☐, Other, specify:_____

7. Contact screening

In household with symptomatic cases, is a contact form with names and age of household contacts filled: Yes ☐ No ☐

Appendix 2: ENGLISH Last updated: 250315

File S3. Statistical method

All analyses were complete case analyses. For each exposure, a univariable Cox proportional hazards regression model was run. For continuous exposures, the HR and corresponding p-value represent an effect size of 1-unit increase in the exposure. For binary exposures, the HR and corresponding p value represent a comparison between TRUE (exposure) and FALSE (reference). For categorical exposures with 3 or more levels, the HR and corresponding p value represent a comparison between the categorical level and the baseline. There is a second p-value that tests the significance of the overall category. If this second p value is not significant, then the within-category-levels analysis should not be considered.

For each exposure group (risk factors, demographic, socio economic, clinical information, and risk behavior) we ran a multivariable Cox proportional hazards regression model containing all of the exposures in the exposure group. For continuous exposures, the HR and corresponding p value represent an effect size of 1-unit increase in the exposure. For binary exposures, the HR and corresponding p value represent a comparison between TRUE (exposure) and FALSE (reference). For categorical exposures with 3 or more levels, the HR and corresponding p value represent a comparison between the categorical level and the baseline. There is a second p value that tests the significance of the overall category. If this second p value is not significant, then the within-category-levels analysis should not be considered.

For each exposure group (risk factors, demographic, socio economic, clinical information, and risk behaviour) we ran a multivariable Cox proportional hazards Lasso (penalized) regression model containing all of the exposures in the exposure group. A lasso regression performs automated variable selection, and only reports penalized hazard ratios. The hazard ratios are interpreted in a similar manner to the other analyses. There are no confidence intervals or p-values in this analysis. If the hazard ratio is not 1, then it is considered significant.

Interpretation of Lasso versus fully adjusted HR

Lasso may have bias, but reduces variability. Multivariate regression on the other hand may have less bias, but high variability. This means that if you repeat the analyses 100 times, lasso will continue to produce similar results, whereas repeated regression analyses will produce a range of different results. The choice of method is really a trade-off between bias and stability of the results. Lasso is the preferred method when in analyses where you want to adjust for many exposure variables, while the number of cases is limited (as in the current paper).

Lasso regression

If fully penalized, there will be no effects at all; effect size is fixed at 1. In the current model, the analysis is repeated several times and the model decides what the best prediction validity is. Compared to other adjusted regression models, lasso may have bias, but reduces variability. It is therefore the most trustworthy estimate. Lasso estimates should be treated as significant predictors although they come without p values or CI intervals and should report the overall p-value for the fully adjusted HR. Values should be interpreted similar to traditional regression models. Lasso results will be the main results in the paper and the ones that should be reported in the abstract. It also should be presented alongside crude and adjusted traditional regression results to provide the full picture. Variance may explain discordance between fully adjusted HR and Lasso.

Table S1, Risk factors for all-type pulmonary tuberculosis in Dale, Ethiopia, 2016–2017

Covariates		N	TB	Crude HR	p	Adjusted HR	p	Lasso HR
		3484	180	(95% CI)		(95% CI)		
Socio-demographic								
Age-groups (years)		3484	180		0.001 q		0.001 q	
	15 to 34 years	1238	108	1		1		
	35 to 64 years	1945	65	0.37 (0.27, 0.50)	0.001	0.34 (0.24, 0.48)	0.001	0.46
	65+ years	301	7	0.24 (0.11, 0.52)	0.001	0.19 (0.09, 0.42)	0.001	0.39
Sex								
	Female	2039	82	1		1		
	Male	1445	98	1.72 (1.28, 2.30)	0.001	2.09 (1.50, 2.92)	0.001	1.63
Catchment area					0.001 q		0.001 q	
	Semen Mesenkala	416	29	1		1		
	Magara	220	12	0.83 (0.43, 1.63)	0.597	1.06 (0.53, 2.13)	0.866	
	Hida Kaliti	77	9	1.63 (0.77, 3.44)	0.203	2.17 (0.96, 4.91)	0.062	1.37
	Bera Tadicho	446	31	1.04 (0.63, 1.72)	0.884	2.62 (1.51, 4.55)	0.001	1.55
	Goida	267	16	0.92 (0.50, 1.70)	0.791	1.77 (0.93, 3.37)	0.083	1.1
	Boa Badagalo	675	25	0.56 (0.33, 0.95)	0.032	0.87 (0.50, 1.54)	0.643	0.88
	Dagyia	380	5	0.21 (0.08, 0.54)	0.001	0.52 (0.20, 1.39)	0.193	0.7
	Gidamo	201	11	0.74 (0.37, 1.49)	0.406	2.05 (0.98, 4.30)	0.057	-
	Moto	466	21	0.66 (0.38, 1.15)	0.144	1.02 (0.57, 1.83)	0.948	-
	Semen Kege	336	21	0.98 (0.56, 1.72)	0.939	1.77 (0.96, 3.26)	0.069	1.02
Occupation					0.039 q		0.363 q	
	Farmer	2328	123	1		1		
	Housewife	883	34	0.76 (0.52, 1.11)	0.16	0.88 (0.57, 1.35)	0.557	-
	Merchant	70	3	0.79 (0.25, 2.48)	0.686	0.56 (0.17, 1.90)	0.354	-
	Student	169	18	2.04 (1.25, 3.35)	0.005	1.90 (1.08, 3.35)	0.027	1.42
	GO	16	2	2.28 (0.56, 9.21)	0.248	2.21 (0.53, 9.26)	0.278	1.01
	Daily labourer	10	0	-	0.993	-	0.995	0.92
	Other	8	0	-	0.993	-	0.996	-
Marital status								
	Married	2801	132	1		1		
	Not-married	683	63	1.49 (1.07, 2.07)	0.019	0.87 (0.58, 1.30)	0.492	-
n of household members		3484		1.11 (1.00, 1.24)	0.06	1.10 (0.98, 1.23)	0.114	1.04
Years completed		3481	180	0.82 (0.78, 0.86)	0.001	0.77 (0.72, 0.82)	0.001	0.82
Clinical information								
BMI >= 18.5 kg/m^2		1396	57	1		1		
BMI < 18.5 kg/m^2		2077	123	1.51 (1.10, 2.07)	0.01	1.29 (0.92, 1.79)	0.136	1.07

MUAC in cm		3453	180	0.81 (0.76, 0.86)	0.001	0.79 (0.74, 0.85)	0.001	0.83
Risk factors								
Previous history of TB								
	no	2945	150					
	yes	537	30	1.00 (0.67, 1.47)	0.982	1.29 (0.81, 2.05)	0.279	-
History of TB case in household								
	no	2977	153					
	yes	505	27	0.98 (0.65, 1.47)	0.914	1.44 (0.86, 2.40)	0.165	0.96
Living with chronic cougher								
	no	2539	150	1				
	yes	945	30	0.51 (0.34, 0.75)	0.001	0.40 (0.25, 0.64)	0.001	0.62
HIV test		3483						
	No	2754	131	1				
	yes	729	49	1.42 (1.02, 1.97)	0.037	1.85 (1.31, 2.60)	0.001	1.39
Ever drink alcohol								
	no	3295	173					
	yes	189	7	0.66 (0.31, 1.41)	0.288	0.62 (0.24, 1.58)	0.315	0.86
Ever chew chat								
	No	3268	172	1		1		
	yes	216	8	0.67 (0.33, 1.36)	0.263	0.73 (0.27, 1.99)	0.544	96
Ever smoke								
	No	3355	175	1		1		
	yes	129	5	0.72 (0.29, 1.74)	0.461	0.80 (0.23, 2.85)	0.736	0.99
Currently smoking								
	no	3235	173					
	yes	247	7	0.52 (0.24, 1.11)	0.089	0.70 (0.26, 1.91)	0.484	0.83
Previously smoke smoker								
	No	3183	167					
	Yes	299	12	0.75 (0.42, 1.35)	0.344	0.91 (0.42, 1.96)	0.808	-
Economic indicators								
Number of rooms		3484		0.85 (0.71, 1.02)	0.074	0.98 (0.79, 1.20)	0.814	
Wall type								
	Wood only /other	1839	121	1		1		
	Wood with/mud/brick /cement	1645	59	0.52(0.38, 0.71)	0	0.68 (0.47, 0.97)	0.035	0.73
Roof type								
	Leaf/tached/other	1222	37	1				

	Corrugated iron	2262	143	0.45 (0.32, 0.65)	0.001	0.70 (0.45, 1.11)	0.128	0.86
Electricity access								
	No	2864	151	1				
	Yes	620	29	0.89 (0.60, 1.33)	0.57	1.34 (0.86, 2.10)	0.192	-
Fuel for cooking								
	Other	3457	180					
	Electricity	27	0	0		0.992		0.993
Separate kitchen								
	No	2746	166	1			1	
	yes	738	14	0.30 (0.17, 0.51)	0.001	0.70 (0.45, 1.11)	0.128	0.63
Ventilation								
	No	3098	165					
	yes	386	15	0.30 (0.17, 0.51)	0.001	0.70 (0.45, 1.11)	0.128	0.63
Heating								
	No	3394	179					
	Yes	90	1	0.22 (0.03, 1.56)	0.129	0.21 (0.03, 1.64)	0.137	0.88
Bank account								
	No	3350	176					
	Yes	133	3	0.40 (0.13, 1.25)	0.116	0.53 (0.12, 2.22)	0.382	-
Land agriculture								
	No	444	36					
	Yes	3040	144	0.54 (0.37, 0.78)	0.001	0.80 (0.53, 1.19)	0.27	0.87
Mobile								
	No	2862	150					
	Yes	622	30	0.85 (0.58, 1.26)	0.429	0.93 (0.60, 1.43)	0.738	-
TV								
	No	3415	178					
	Yes	69	2	0.55 (0.14, 2.22)	0.403	1.03 (0.17, 6.11)	0.972	-
Radio								
	No	2995	160					
	Yes	489	20	0.71 (0.45, 1.13)	0.152	1.08 (0.64, 1.84)	0.769	-
Refrigerator								
	No	3450	178					
	Yes	34	2	1.15(0.28,4.63	0.848	2.52 (0.42,15.03)	0.309	-

TB, tuberculosis; HR, hazard ratio; CI, confidence interval; GO, government employee; BMI, body mass index; MUAC, middle-upper arm circumference; TV, television; *n*=number

[q]*p* value for the variable as a whole for variables with more than one value.

Table S2, Comparison of current versus previous study in six kebeles included in the current study 5-years earlier.

	Six kebeles		
Categories	Current	Previous	Reduced
Number of kebeles	6	6	
Population >=14 years previous />=15 years current	24117	21774	
Persistent coughers	273	724	2.7
Smear positive TB cases diagnosed	5	23	4.6
Observation years	333	588	1.8
TB per 100 000 observation years	1502	3912	2.6