Statistical modelling to predict silicosis risk in deceased Southern African gold miners without medical evaluation

AUTHORS:
Jonathan E. Myers[1]
Mary Lou Thompson[2]

AFFILIATIONS:
[1]School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa
[2]Department of Biostatistics, University of Washington, Seattle, Washington, USA

CORRESPONDENCE TO:
Jonathan Myers

EMAIL:
myers.jonny@gmail.com

DATES:
Received: 06 Oct. 2021
Revised: 28 Feb. 2022
Accepted: 05 Apr. 2022
Published: 28 July 2022

HOW TO CITE:
Myers JE, Thompson ML. Statistical modelling to predict silicosis risk in deceased Southern African gold miners without medical evaluation. S Afr J Sci. 2022;118(7/8), Art. #12502. https://doi.org/10.17159/sajs.2022/12502

ARTICLE INCLUDES:
☒ Peer review
☒ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☐ On request from author(s)
☒ Not available
☐ Not applicable

EDITOR:
Pascal Bessong

KEYWORDS:
gold miner, silicosis, risk prediction, statistical model, classification accuracy

FUNDING:
Qhubeka Trust

The Qhubeka Trust was established in 2016 in a legal settlement on behalf of former gold miners seeking compensation for silicosis contracted on the South African mines. Settlements resulting from lawsuits on behalf of gold miners aim to provide fair compensation. However, occupational exposure and medical records kept by South African mining companies for their employees have been very limited. Some claimants to the Qhubeka Trust died before medical evaluation was possible, thus potentially disadvantaging their dependants from receiving any compensation. With medical evaluation no longer possible, a statistical approach to this problem was developed. The records for claimants with medical evaluation were used to develop a logistic regression prediction model for the likelihood of silicosis, based on the potential predictors: cumulative exposure to respirable dust, age, years since first exposure, years of life lost prematurely, vital status at 31 December 2019, and a history of tuberculosis diagnosis. The prediction model allowed estimation of the likelihood of silicosis for each miner who had died without medical evaluation and is a novel approach in this setting. In addition, we were able to quantitatively evaluate the trade-offs of different silicosis risk classification thresholds in terms of true and false positives and negatives.

Significance:

- A statistical approach can be used for risk estimation in settings where the outcome of interest is unknown for some members of a class.
- The likelihood of silicosis in deceased miners without medical evaluation in the Qhubeka Trust can be accurately estimated, using information from finalised claims.
- Strategies for classifying the silicosis status of deceased miners without medical evaluation in the Qhubeka Trust can be assessed in a rigorous, quantitative framework.

Introduction

A settlement arising out of a lawsuit against Anglo American South Africa Limited and AngloGold Ashanti led to the establishment of the Qhubeka Trust to process the claims for silicosis of a closed list of 4365 named gold miner claimants. The claims process required each claimant, or, where the claimant was deceased, their dependant, to formally lodge a claim, and for the claimant to be medically examined to determine whether they had a compensable silica-related disease as set out in the Trust Deed. In the case of deceased claimants, the dependants were required to provide medical information on which to base a diagnosis.

Provision16.2 (ii) (1) of the Trust Deed specifies that dependants may qualify for compensation (but only at the lowest level) based on medical records, employment records, or other evidence that the Trustees deem 'credible and reliable'. There are 466 deceased Qhubeka Trust claimants for whom a claim was lodged by their dependants, but for whom there is no, or insufficient, medical information on which to base a diagnosis of silicosis. In addition to requesting medical records from the deceased's dependants, various avenues of enquiry undertaken by the Qhubeka Trust to obtain medical information for these claimants have had limited or no success. It is notable that the silica exposure and occupational medical information kept by mining companies for their employees has historically been incomplete, and, in many cases, non-existent. Without such medical documentation, or any other means to determine the possible presence of silicosis, family members of deceased claimants, or, as described in the Trust Deed, 'dependent claimants', may be disadvantaged from receiving compensation to which they may be entitled.

The benefit award to Qhubeka Trust qualifying claimants is paid in two tranches – the first when the claim has been processed and the claimant is determined to have met all qualifying criteria, and the second and larger when the totality of claims for all claimants have been medically assessed and finalised. Consequently, by the end of 2019, by which time all the living claimants had been medically evaluated, finalisation of payment of equitable benefits to all claimants was being held up by the indeterminate silicosis status of the 466. There was therefore some urgency to identify an approach that would enable equitable evaluation of claims from this group of deceased ex-miners, as the second and final payment to qualifying claimants could not be made until all outstanding claims (the 466 deceased claims) were finalised. At that point, all available benefit funds could be distributed equitably.

In light of challenges with obtaining medical information for deceased claimants, the Trustees approached us to determine whether it was possible, by means of statistical modelling, to estimate the probability of deceased claimants with insufficient medical information having had silicosis at the time of their deaths.

A statistical prediction model, novel in this context, but using well-established statistical methodology, was developed to predict likelihood of silicosis for each deceased miner without medical evaluation.

Materials and methods

A Microsoft Excel database with anonymised available information on all claimants was supplied by the Qhubeka Trust for the purpose of this study only. The data otherwise remain under the confidential control of the Trust. Clause 12.4 of the Qhubeka Trust Deed provides that:

> *The trustees have great power to use, publish or disseminate in anonymised form and without profit or payment, the information and data gathered during the course of the operation of the Trust solely for the purpose of advancing public health.*

A prediction model for the risk of silicosis was developed, based on the silicosis status of the 'Modelling group', consisting of those claimants who had been medically assessed for the presence of silicosis, and for whom there was complete information on the potential predictors. The 'Prediction group' consisted of the deceased miners without medical evaluation.

The statistical methodology used is well established, but the application in this setting is novel.

Based on associations reported in the literature[1-15] and data availability, the potential predictors of silicosis risk that we considered are listed in Table 1. All variables were examined for potential errors and missing values which might limit numbers of individuals who could be included in the analyses.

Table 1: Potential predictors of silicosis diagnosis

Predictor variable	Description
Cumulative exposure	This was calculated [by JEM] from each claimant's total years of recorded service multiplied by the average respirable dust concentration in mg-years/m^3 (as reported in the literature[2-7,9,11,12]) for their main job
Age	Age in years at 31 December 2019 (for deceased claimants this was the age they would have attained at this date)
Latency	Years from start date working on mines to claim date at the Qhubeka Trust
Years of life lost	Years of life lost prematurely by those who died before 31 December 2019. For claimants alive at that date this is zero.
Vital status	Vital status at 31 December 2019 (0 = alive, 1 = deceased)
History of tuberculosis	Tuberculosis diagnosis (0 = no, 1 = yes)

Initial exploratory data analysis was conducted in the Modelling group, including cross-tabulations of silicosis status with categorical predictor variables (vital status at 31 December 2019 and historical diagnosis of tuberculosis), and non-parametric smoothed plots versus risk of silicosis for continuous predictor variables (cumulative exposure to respirable dust, age at 31 December 2019, latency, and years of life lost).

Because all claimants in the Prediction group are deceased, one option considered was to base prediction modelling just on the deceased in the Modelling group, this being the subgroup that best parallels the Prediction group. We elected rather to develop prediction models based on all claimants in the Modelling group, while including vital status as a predictor, because information from alive claimants would inform other aspects of silicosis risk such as cumulative exposure.

We considered the following predictors in three logistic regression models with silicosis (yes/no) as the outcome:

M1: cumulative dust exposure, age, years of life lost, latency, vital status, tuberculosis

M2 (allowing for non-linearity of the continuous predictors, cumulative exposure and age): linear + quadratic cumulative dust exposure, linear + quadratic age, years of life lost, latency, vital status, tuberculosis

M3 (without latency): linear + quadratic cumulative dust exposure, linear + quadratic age, years of life lost, vital status, tuberculosis

The fitted models yielded predicted risks of silicosis for all claimants in both the Modelling and Prediction groups. Towards exploring risk thresholds to classify claimants in the Prediction group as silicosis positive/negative, Tables 2 and 3 introduce and define the terminology associated with classification accuracy. True positives (TPs) and negatives (TNs), false positives (FPs) and negatives (FNs) are the building blocks for calculating sensitivity, specificity, positive and negative predictive values (PPV and NPV) and the proportion correctly classified. PPV, NPV and proportion correctly classified depend on the prevalence (n_1/N) of the disease under consideration, while sensitivity and specificity do not.

Table 2: Terminology used in assessing classification accuracy

		True disease status		
		Positive	**Negative**	
Model disease classification	**Positive**	True positive (TP)	False positive (FP)	A
	Negative	False negative (FN)	True negative (TN)	B
		n_1	n_2	N

Table 3: Definitions of classification accuracy characteristics

Term	Definition
Risk classification threshold	A classification rule such that all individuals with a predicted risk at or above the threshold would be classified as positive for disease
Sensitivity	The proportion of individuals who are classified positive among all those who actually have the disease (TP/n_1)
Specificity	The proportion of individuals who are classified negative among all those who actually do not have the disease (TN/n_2)
Positive predictive value (PPV)	The proportion of individuals who actually have the disease among all those who are classified positive (TP/A)
Negative predictive value (NPV)	The proportion of individuals who actually do not have the disease among all those who are classified negative (TN/B)
Proportion correctly classified	The proportion of individuals who are correctly classified (TP+TN)/N

The receiver operating characteristic (ROC) curve plots sensitivity versus 1-specificity at each possible risk classification threshold. For each of the models we considered, we estimated the associated area under the ROC curve (AUC). AUCs can take on values from 0.5 to 1.0 and provide a measure of the model's ability to discriminate between those subjects who experience the outcome of interest (here, silicosis) versus those who do not. A model with AUC of 0.5 is no better than a coin toss; AUCs in the range 0.7–0.8 are regarded as reflecting a prediction model with acceptable discrimination, 0.8–0.9 as excellent discrimination and above 0.9 as outstanding discrimination.[16] We used likelihood ratio tests to compare models and the final model was chosen as the model with the minimum Akaike Information Criterion (AIC)[17] among competing models.

For the final model, we assessed the goodness of fit of the silicosis risk predictions applied to the Modelling group graphically and quantitatively by the Hosmer–Lemeshow and Stukel goodness-of-fit tests.[16,18,19] We estimated the accuracy characteristics of the final model (sensitivity, specificity, PPV, NPV and per cent correctly classified) for silicosis classifications, according to different thresholds of predicted risk, focusing on the subgroup of deceased claimants in the Modelling group, which was considered to be most comparable to the deceased claimants in the Prediction group, including with regard to likely prevalence of silicosis.

Accuracy characteristics estimated from the same data on which a model was developed may be 'optimistic', i.e. may be more favourable than if the model was applied to new data. We assessed optimism in measures of AUC and accuracy by bootstrapping.[20]

STATA version 13.1 was used for analysis.

Results

Figure 1 shows the finalisation status of Qhubeka Trust claims at the end of December 2019. At that date, 3369 claimants (2993 living and 376 deceased) had been medically assessed and their claims finalised. Of the claimants with finalised claims, 58% were diagnosed with silicosis and 63% had a history of tuberculosis diagnosis.

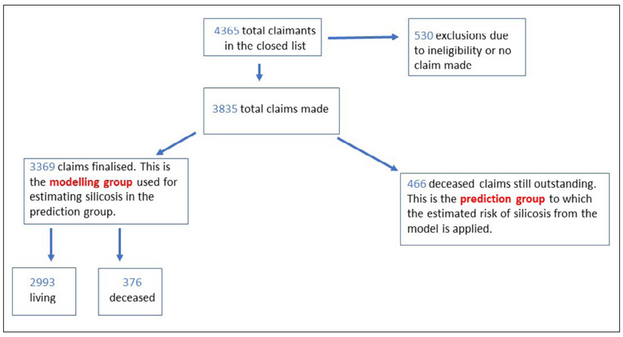
4365 total claimants in the closed list
530 exclusions due to ineligibility or no claim made
3835 total claims made
3369 claims finalised. This is the modelling group used for estimating silicosis in the prediction group.
466 deceased claims still outstanding. This is the prediction group to which the estimated risk of silicosis from the model is applied.
2993 living
376 deceased

Figure 1: Finalisation status of Qhubeka Trust claimants as of 31 December 2019.

Of the 3369 claimants with diagnosis in the Modelling data, there were 220 for whom cumulative exposure could not be determined (because length of service was not known) and 16 for whom age (at 31 December 2019) could not be determined (because date of birth was not known). Additionally, latency could not be determined for a further 22 of these claimants, because either start date of employment at the mines or claim date was not known. Analyses were restricted to claimants with complete records, i.e. $n=3111$ for analyses including latency, and $n=3133$ for analyses without this variable.

Table 4 shows summaries of silicosis predictor variables in the Modelling group by silicosis diagnosis status, for the 3133 claimants with complete data on all predictor variables except latency (22 missing) and in the Prediction group ($n=466$, 10 missing latency). Claimants with silicosis had higher cumulative exposure to respirable dust, more years of life lost among the deceased, were more likely to be deceased by 31 December 2019, and were more likely to have been diagnosed with tuberculosis.

Table 4: Characteristics of claimants in the Modelling (by silicosis diagnosis) and Prediction groups

	Silicosis ($n=1897$)	No silicosis ($n=1236$)	Prediction group ($n=466$)
Predictor variables	**Median (range)**		
Cumulative exposure (mg/m^3 years)	5.8 (0.11, 24.4)	3.3 (0.05, 20.8)	4.9 (0.2, 15.8)
Age on 31 December 2019	66.6 (47.8, 105.6)	63.0 (34.6, 93.0)	68.5 (48.2, 97.1)
Years of life lost if deceased (relative to 31 December 2019)[a]	3.7 (0.1, 20.0)	2.2 (1.5, 2.9)	4.9 (1.0, 15.3)
Latency (years)[b]	38.0 (10, 63)	37.0 (9, 62)	38.9 (21, 69)
	***n* (column %)**		
Deceased by 31 December 2019	348 (18.3 %)	13 (1.05%)	466 (100%)
Tuberculosis	1414 (74.5%)	586 (47.4%)	199 (42.7%)

[a] *Modelling group:* n=*348 and* n=*13 with and without silicosis, respectively.*

[b] *Modelling group:* n=*1886 and* n=*1225 with and without silicosis, respectively (22 missing); Prediction group:* n=*456 (10 missing).*

Of the 3133 claimants with complete data in the Modelling group who were alive at 31 December 2019, 55.9% had been diagnosed with silicosis, while, of the claimants deceased at that date, 96.4% had been diagnosed with silicosis. It is evident that deceased status is strongly related to presence of silicosis. Of the 361 deceased miners in the Modelling group with complete records, only 13 had not been diagnosed with silicosis.

Model development

Table 5 shows the AUC and AIC for each model (M1, M2, M3) as well as the *p*-values for the likelihood ratio test comparing M2 to M1 and M3 to M2.

Table 5: Comparison of models for predicting risk of silicosis

Model	Area under the curve	Akaike Information Criterion	Likelihood ratio test *p*-value
M1	0.772	3437.7	–
M2	0.774	3425.5	0.0003[a] (M2 cf M1)
M3	0.774	3423.6	0.725[b] (M3 cf M2)

[a] *Tests the null hypothesis that the quadratic terms for cumulative exposure and age are zero.*

[b] *Tests the null hypothesis that the coefficient for latency is zero.*

Characteristics of all three models considered were similar, but our examination of competing models indicated that latency did not contribute significantly to model fit and that the model was improved with inclusion of quadratic terms for cumulative exposure and age. On this basis, model M3 was chosen as the final model for prediction and refitted, using the 3133 records with complete data (excluding latency). The refitted model was associated with estimated AUC=0.78 or, when restricted to deceased claimants, AUC=0.82. The coefficients for this final model are shown in Table 6 and the associated ROC curves are shown in Supplementary figure 1.

Table 6: Final model to predict risk of silicosis[a]

Predictor variable	Estimated coefficient	Odds ratio	*p*-value
Linear cumulative exposure (mg-years /m^3)	0.1810	1.20	<0.001
Quadratic cumulative exposure	-0.0079	0.99	0.002
Linear age (years)	0.0438	1.04	<0.001
Quadratic age	-0.0014	0.99	0.003
Years of life lost	0.4646	1.59	0.014
Vital status	1.4983	4.47	0.004
Tuberculosis	1.1152	3.05	<0.001
Intercept	-0.3113	–	<0.001

[a] *The coefficients in the table can be used to calculate probability estimates for any given predictor values, using the logistic equation. Cumulative exposure is centred at 5 mg-years/m^3 and age is centred at 65 years. The odds ratios are provided additionally for ease of interpretation.*

The odds of silicosis are estimated to be 4.5 times higher in individuals who are deceased than those still alive at 31 December 2019 (among individuals with the same exposure, age, tuberculosis status and years of life lost). For every additional year of life lost, odds of silicosis are estimated to increase by a multiple of 1.59. The odds of silicosis are estimated to be 3.05 times higher in individuals with tuberculosis than those without (among individuals with the same exposure, age, vital status and years of life lost). (The odds ratios for cumulative exposure and age cannot be directly interpreted, because of the quadratic terms.)

Table 7 summarises the predicted risks of silicosis in the Modelling group, stratified by vital status, and in the Prediction group. It is notable that the predicted risks of silicosis in both deceased groups are high (all above 70%) and that they have very similar ranges.

Table 7: Predicted risk of silicosis

Claimant group	Predicted risk of silicosis [Mean (minimum, maximum)]
Modelling: Alive	0.5588 (0.0620, 0.8973)
Modelling: Deceased	0.9640 (0.7024, 1.0000)
Prediction	0.9616 (0.7055, 0.9996)

Model evaluation

A goodness-of-fit comparison of observed and expected (from the M3 prediction model) numbers of silicosis cases in predicted risk deciles is shown in Table 8 and graphically in Supplementary figure 2. Both indicate that the model fits the data well. The *p*-values for the Hosmer–Lemeshow and Stukel goodness-of-fit tests for this model (M3) were $p=0.46$ and $p=0.87$, respectively; both also indicate good fit.

Table 8: Comparison of observed and expected (under M3) counts with and without silicosis by predicted risk deciles

Decile	Predicted risk	Observed with silicosis	Expected with silicosis	Observed without silicosis	Expected without silicosis	Total
1	[0, 0.273]	69	67.2	245	246.8	314
2	(0.273, 0.381]	93	101.7	220	211.3	313
3	(0.381, 0.475]	143	134.7	170	178.3	313
4	(0.475, 0.552]	162	160.6	152	153.4	314
5	(0.552, 0.621]	193	183.5	120	129.5	313
6	(0.621, 0.684]	189	204.5	124	108.5	313
7	(0.684, 0.749]	223	225.5	91	88.5	314
8	(0.749, 0.818]	251	245.8	62	67.2	313
9	(0.818, 0.924]	269	267.1	44	45.9	313
10	(0.924, 1]	305	306.5	8	6.5	313

Supplementary figure 3 shows side-by-side box plots, by true silicosis status, of predicted risk of silicosis for all claimants in the Modelling group and for deceased claimants in the Modelling group. The greater the separation between the distributions of predicted risk in those with and without 'disease' (here, silicosis), the better the predictive performance of a model. Note that all deceased claimants in the Modelling group have predicted risk of silicosis above 70%.

Silicosis risk classification thresholds

For the purposes of Qhubeka Trust claims, having estimated the likelihood of silicosis for each miner in the Prediction group, the question remains as to what level of predicted risk should be regarded as sufficient to classify a claimant as 'silicosis positive'. By 'risk classification threshold', we mean a classification rule such that all claimants with a predicted risk at or above the threshold would be classified as positive for 'disease' (here, silicosis). Typically, in deciding on a risk threshold, considerations are weighed as to the consequences of false positives (FPs) and false negatives (FNs), and the benefits of true positives (TPs) and true negatives (TNs). In some circumstances, it is a high priority to reduce FNs (missed cases), e.g. for a disease that has poor prognosis if untreated, or, as in the case of the Qhubeka Trust, to avoid qualifying claimants being denied compensation. In some circumstances, it may be a high priority to reduce FPs, e.g. when the next steps for those classified as positive are medically invasive. If avoiding FNs and FPs are of equal priority, one might consider selecting a risk threshold that maximises both sensitivity and specificity. However, this approach ignores the impact of the prevalence of the outcome of interest. If, for instance, a disease is common (i.e. has a high prevalence, as with silicosis in the deceased claimants), FNs will dominate the misclassified individuals compared with FPs. Sometimes it is possible to assign values to costs and benefits of false and true positives and negatives, and then choose a threshold which maximises the net benefit of a classification rule. In the absence of quantifiable costs and benefits, one approach, which does acknowledge prevalence to some extent, is to consider a risk threshold with the greatest percentage of correct classifications.

Our evaluation of silicosis risk prediction thresholds is based on the $n=361$ Modelling group claimants who are deceased, as this group most closely parallels the Prediction group, where all claimants were deceased. Note again, referring to Table 7, that the distributions of predicted risk in these two groups are very similar. Only 13 (out of 361) or 3.6% of the deceased in the Modelling group are true silicosis negative cases. This points again to the strong relationship between premature death and silicosis.

For the deceased claimants in the Modelling group, Supplementary figure 4 shows the estimated sensitivity and specificity at each potential risk classification threshold and Table 9 shows estimates of prediction accuracy for the following risk classification thresholds: 0.51–0.7, 0.75, 0.8, 0.85, 0.9, 0.95.

Table 9: Estimated model accuracy by risk threshold (Modelling group, 361 deceased claimants)

Risk threshold	Percentage correct	Sensitivity	Specificity	Positive predictive value	Negative predictive value	False negative (FN; of 348 with silicosis) False positive (FP; of 13 without silicosis)
0.51–0.7	96.4%	100%	0.00%	96.4%	–	FN=0, FP=13
0.75	95.3%	98.8%	0.00%	96.4%	0.00%	FN=4, FP=13
0.80	94.7%	98.3%	0.00%	96.3%	0.00%	FN=6, FP=13
0.85	92.8%	96.0%	7.7%	96.5%	6.7%	FN=14, FP=12
0.90	89.5%	92.2%	15.4%	96.7%	6.9%	FN=27, FP=11
0.95	79.2%	79.9%	61.5%	98.2%	10.3%	FN=70, FP=5

In interpreting Table 9 and using it, and Supplementary figure 4, to guide consideration of an appropriate risk threshold to apply to the deceased claimants in the Prediction group, it must be kept in mind that:

- As the risk threshold to classify claimants as having silicosis increases, sensitivity will decrease and specificity will increase.
- Of the 361 deceased claimants with complete records in the Modelling group, only 13 were not diagnosed with silicosis, i.e. this group has 96.4% true silicosis positive cases (TP), and 3.6% true silicosis negative cases (TN).
- The lowest model-predicted risk of silicosis in these deceased claimants is 70.3%.
- Hence, risk thresholds at 70% or below would classify all Modelling group deceased claimants as positive for silicosis (0 FNs, 13 FPs).
- The lowest predicted risk of silicosis among the 13 deceased subjects without silicosis is 0.84 (see Supplementary figure 3).
- Hence, increasing risk thresholds up to 0.84 will simply increase FNs without reducing FPs, i.e. the impact of moving the risk threshold in Table 9 from 0.7 to 0.75 to 0.8 is simply to increase the number of FNs, while the number of FPs remains unchanged.
- For all thresholds considered here, the NPVs are low, i.e. if a claimant is classified as silicosis negative, there is a low probability of them actually being a true silicosis negative.
- By contrast, PPVs are high, i.e. if a claimant is classified as silicosis positive, there is a high probability of them actually being a true silicosis case.
- Because silicosis is so common in deceased claimants, a threshold with equal sensitivity and specificity (seen from Supplementary figure 4 to be above 0.95), would result in FNs far outweighing FP classifications.
- Of thresholds in the range 0.7–0.95, the highest percentage of correctly classified deceased claimants is 96.4% at a classification threshold of 0.7 or 70%.

Assessment of optimism

Because estimates of model performance that use the same data to develop and evaluate models may be biased upward, i.e. optimistic, we carried out a bootstrap analysis to estimate the extent of optimism in the estimates of the above model characteristics (AUC, % correct, sensitivity, specificity). The estimated optimism for all characteristics was very small and hence no adjustment to the above estimates was required.

Discussion

We have developed a model to predict silicosis risk based on a claimant's cumulative exposure to respirable dust, age, years of life lost prematurely, vital status and diagnosis of tuberculosis. We have demonstrated that the model fits the Modelling group data well and has excellent discriminating ability between those with and without silicosis. Based on this model, we can estimate the risk of silicosis for each claimant in the Prediction group, i.e. the deceased claimants without medical evaluation.

Estimates for the deceased claimants in the Modelling group lead us to anticipate that the fraction of non-silicosis claimants in the Prediction group will be small (3.6% of the deceased in the Modelling group). This also means that estimation of accuracy characteristics relating to true negatives, in particular, specificity, is constrained by the small number of deceased claimants without silicosis. However, examination of Table 9 makes it clear that any threshold which is chosen to improve specificity (which would reduce false positive classifications), will be at the cost of a disproportionate increase in false negative classifications (which would reduce sensitivity).

Silicosis risk classification thresholds were evaluated starting with 0.51 (51% and representing the balance of probabilities of having silicosis) and ranging to 0.95 (95%). A threshold of 0.7 (70%) was associated with the highest per cent of correct classifications (96.4%) and the highest estimated sensitivity (100%: no true cases of silicosis missed), while the very small number of false positives (those without silicosis misclassified as having silicosis) remained more or less constant up to a threshold of 0.9 (90%).

Consideration of risk classification thresholds includes weighing the consequences for the Qhubeka Trust claimants of false negative and false positive classifications. The dependants of FN claimants would be denied compensation to which they are entitled. On the other hand, the consequences of FPs among the deceased Qhubeka Trust claimants are benefits in the form of compensation, i.e. there is no disadvantage to these individuals. Furthermore, we project that the number of true negatives is very small and hence the compensation benefit disadvantage to the entire group of claimants of a risk classification threshold that classifies all undiagnosed claimants as silicosis positive, is also very small, particularly as this is spread over a large number of beneficiaries, and consequently minimally impactful.

A silicosis risk threshold of 70% accords with the highest percentage of claimants being correctly classified, and constitutes the classification rule with the highest sensitivity, with little improvement in specificity at higher risk thresholds up to 90%. All 466 members of the Prediction group have a predicted silicosis risk greater than 70%. If a risk classification threshold of 0.7 (70%) is chosen, then all claimants in this group will be classified as having silicosis at the time of their death.

We note the poor state of employee record-keeping by the mining employers, even for such basic information as individual miners' ages, as well as jobs worked, length of service in these jobs over their careers, their exposures in these jobs, and their health records over many years. This has consequences for claimants with silica-related diseases who may be unable to show their eligibility for benefit. This situation is despite the mining industry having been enjoined as far back as the 1960s21(p.232) to institute a system of integrated individual mining medical records for black miners that encompasses their individual measures of cumulative exposure. The models developed and applied here might have had even greater accuracy with better record-keeping and monitoring by employers. Indeed, the models would have been unnecessary if the silicosis status of the deceased in the Prediction group had been known.

Conclusions

The situation faced by the Qhubeka Trust, where medical records of deceased claimants are not sufficient to provide clear assessment of silicosis, is one that will be encountered in other settings. For instance, a 2018 study22 estimated that, at that time, more than 100 000 miner compensation claims in southern Africa were still unpaid. It is very likely that the administration of such claims will face the same challenges of miners who will have died before medical evaluation. The approach we have described here gives hope that it is still possible, in a scientifically robust manner, to address the eligibility of such claimants.

Competing interests

Both authors were paid as consultants by the Qhubeka Trust to undertake the research presented in this article.

Authors' contributions

Both authors were involved in every aspect of the study: conceptualisation; methodology; data acquisition and cleaning; data analysis; validation; writing.

References

1. US National Institute for Occupational Safety and Health (NIOSH). Criteria for a recommended standard: Occupational exposure to crystalline silica. DHHS (NIOSH) Publication Number 75-120. Washington DC: NIOSH, Centers for Disease Control and Prevention; 1974. Available from: https://www.cdc.gov/niosh/docs/75-120/default.html

2. Rees JP. A note on the exposure-response curve of South African gold miners. J Mine Vent Soc S Afr. 1960:145–148.

3. Beadle DG. An epidemiological study of the relationship between the amount of dust breathed and the incidence of silicosis in South African gold miners. In: Davies CN, editor. Inhaled particles and vapours. Proceedings of an International Symposium organized by the British Occupational Hygiene Society; 1965 September 28 – October 01; Cambridge, UK. Oxford: Pergamon; 1967, p. 479–492. https://doi.org/10.1016/B978-1-4832-1329-3.50045-7

4. Beadle DG, Bradley AA. The composition of airborne dust in South African gold mines. In: Shapiro HA, editor. Pneumoconiosis (Proceedings of the International Conference). Cape Town: Oxford University Press; 1969. p 462–466.

5. Beadle DG. An epidemiological study of the relationship between the amount of dust breathed and the incidence of silicosis in South African gold miners. In: Davies CN, editor. Inhaled particles and vapours II. Oxford: Pergamon Press; 1967. p. 470–490.

6. Beadle DG. The relationship between the amount of dust breathed and the development of radiological signs of silicosis: An epidemiological study in South African gold miners. In: Walton WH, editor. Inhaled particles III. Old Woking: Unwin Brothers; 1971. p. 953–966.

7. Page-Shipp RJ, Harris E. A study of the dust exposure of South African white gold miners. J South Afr Inst Min Metall. 1972;73(1):10–24. https://hdl.handle.net/10520/AJA0038223X_238

8. Verma DK, Sebesteyn A, Julian JA, Muir DCF, Schmidt H, Bernholtz CD, et al. Silica exposure and silicosis among Ontario Hardrock Miners: II Exposure estimates. Am J Ind Med. 1989;16(1):13–28. https://doi.org/10.1002/ajim.4700160104

9. Hnizdo E, Sluis-Cremer GK: Risk of silicosis in a cohort of white South African gold miners. Am J Ind Med. 1993;24:447–457. https://doi.org/10.1002/ajim.4700240409

10. US National Institute for Occupational Safety and Health (NIOSH). NIOSH Hazard Review: Health effects of occupational exposure to respirable crystalline silica. DHHS (NIOSH) Publication Number 2002-129. Washington DC: NIOSH, Centers for Disease Control and Prevention; 2002. Available from: https://www.cdc.gov/niosh/docs/2002-129/default.html

11. Churchyard G, Dekker K, Ehrlich R, te Water Naude J, Myers J. SIMRAC report 606: Silicosis prevalence and risk factors in long service black miners on the South African goldmines. Safety in Mines Research Advisory Committee; 2003.

12. Churchyard GJ, Ehrlich R, te Water Naude JM, Pemba L, Dekker K, Vermeijs M, N White N, Myers J. Silicosis prevalence and exposure-response relations in South African goldminers. Occup Environ Med. 2004;61:811–816. https://doi.org/10.1136/oem.2003.010967

13. Churchyard GJ, Kleinschmidt I, Corbett EL, Murray J, Smit J, De Cock KM. Factors associated with an increased case fatality rate in HIV-infected and non-infected South African gold miners with pulmonary tuberculosis. Int J Tuberc Lung Dis. 2000;4(8):705–712.

14. Bloch K, Johnson LF, Nkosi M, Ehrlich R. Precarious transition: A mortality study of South African ex-miners. BMC Public Health. 2018;18:862. https://doi.org/10.1186/s12889-018-5749-2

15. Ehrlich R, Akugizibwe P, Siegfried N, Rees D. The association between silica exposure, silicosis and tuberculosis: A systematic review and meta-analysis. BMC Public Health. 2021;21:953. https://doi.org/10.1186/s12889-021-10711-1

16. Hosmer DW, Lemeshow S, Sturdivant R. Applied logistic regression. 3rd ed. New York: Wiley; 2013. https://doi.org/10.1002/9781118548387

17. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr.1974;19:716–723. https://doi.org/10.1109/tac.1974.1100705

18. Stukel TA. Generalized logistic models. J Am Stat Assoc. 1988;83:426–431. https://doi.org/10.1080/01621459.1988.10478613

19. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med. 1997;16:965–980. https://doi.org/10.1002/(sici)1097-0258(19970515)16:9<965::aid-sim509>3.0.co;2-o

20. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–387. https://doi.org/10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4

21. Du Toit RSJ. Mine ventilation officials and the quality of mine air. Presidential address: J Mine Vent Soc S Afr. 1962:225-236.

22. Kistnasamy B, Yassi A, Yu J, Spiegel SJ, Fourie A, Barker S, et al. Tackling injustices of occupational lung disease acquired in South African mines: Recent developments and ongoing challenges. Glob Health. 2018;14(1):60. https://doi.org/10.1186/s12992-018-0376-3