Annals of Work Exposures and Health, 2022, Vol. 66, No. 8, 1010–1021
https://doi.org/10.1093/annweh/wxac040
Advance Access publication 18 June 2022
Original Article

Original Article

Monitoring Worker Exposure to Respirable Crystalline Silica: Application for Data-driven Predictive Modeling for End-of-Shift Exposure Assessment

Cody Wolfe[1,*], Lauren Chubb[1], Rachel Walker[1], Milan Yekich[1] and Emanuele Cauda[1]

[1]Centers for Disease Control and Prevention (CDC), National Institute for Occupational Safety and Health (NIOSH), Pittsburgh Mining Research Division (PMRD), 626 Cochran Mill Rd Pittsburgh, PA, 15236 USA.

*Cody Wolfe. | Tel: 412-386-5307; email: CWolfe2@cdc.gov

Submitted 16 December 2021; revised 20 May 2022; editorial decision 20 May 2022; revised version accepted 26 May 2022.

Abstract:

In the ever-expanding complexities of the modern-day mining workplace, the continual monitoring of a safe and healthy work environment is a growing challenge. One specific workplace exposure concern is the inhalation of dust containing respirable crystalline silica (RCS) which can lead to silicosis, a potentially fatal lung disease. This is a recognized and regulated health hazard, commonly found in mining. The current methodologies to monitor this type of exposure involve distributed sample collection followed by costly and relatively lengthy follow-up laboratory analysis. To address this concern, we have investigated a data-driven predictive modeling pipeline to predict the amount of silica deposition quickly and accurately on a filter within minutes of sample collection completion. This field-based silica monitoring technique involves the use of small, and easily deployable, Fourier transform infrared (FTIR) spectrometers used for data collection followed by multivariate regression methodologies including Principal Component Analysis (PCA) and Partial Least Squares (PLS). Given the complex nature of respirable dust mixtures, there is an increasing need to account for multiple variables quickly and efficiently during analysis. This analysis consists of several quality control steps including data normalization, PCA and PLS outlier detection, as well as applying correction factors based on the sampler and cassette used for sample collection. While outside the scope of this article to test, these quality control steps will allow for the acceptance of data from many different FTIR instruments and sampling types, thus increasing the overall useability of this method. Additionally, any sample analyzed through the model and validated using a secondary method can be incorporated into the training dataset creating an ever-growing, more robust predictive model. Multivariant predictive modeling has far-reaching implications given its speed, cost, and scalability compared to conventional approaches. This contribution presents the application of PCA and PLS as part of a

What's Important About This Paper?

Respirable crystalline silica is a recognized and regulated health hazard commonly found in mining. Faster and cheaper ways to examine silica exposure need to be developed to protect worker overexposure and assess the need for corrective actions to reduce exposure. This study leverages the power of high-level computational modeling to analyze end-of-shift silica monitoring data obtained through Fourier transform infrared (FTIR) spectrometry.

computational pipeline approach to predict the amount of a deposited mineral of interest using FTIR data. For this specific application, we have developed the model to analyze RCS, although this process can be implemented in the analysis of any IR-active mineral, and this pipeline applied to any FTIR data.

Keywords: dust sampling; Fourier-transform infrared; infrared analysis; Quartz analysis; respirable dust; silica analysis; silica; silica exposure

1. Introduction

The inhalation of respirable crystalline silica (RCS) can lead to a potentially fatal disease called silicosis. Silicosis is a fibrotic lung disease characterized by the phagocytosis of crystalline silica within the lung resulting in lysosomal damage and the impairment of normal lung function. The creation of fibrotic tissue continues to trigger the inflammatory cascade which creates a feed-forward cycle and can impair lung function even after an individual is no longer exposed to silica (Leung *et al.*, 2012). In the United States, there are approximately two million workers in the construction industry and another three hundred thousand in other industries who are exposed to RCS. In the United States, the National Institute for Occupational Safety and Health (NIOSH) has a recommended exposure limit of 50 µg/m^3, and the Occupational Safety and Health Administration (OSHA), which sets the regulatory limits, requires sampling for RCS with a recommended exposure limit of 50 µg/m^3 for up to a 10-hour workday and 40-hour workweek. Additionally, the Mine Safety and Health Administration (MSHA) has an RCS permissible exposure limit that equates to 100 µg/m^3. In the mining environment, the amount of airborne RCS can be determined using portable, wearable sampling instruments, which unfortunately do not currently produce data in real-time but allow for end-of-shift analysis of the average quartz concentration by FTIR to assess a worker's exposure. Previous research has outlined the methodology for field-based detection of RCS utilizing portable Fourier transform infrared (FTIR) spectrometers for direct-on-filter (DOF) analysis of respirable dust samples (Miller *et al.*, 2012 , 2013; Cauda *et al.*, 2016, 2018; Hart *et al.*, 2018; Ashley *et al.*, 2020).

Current silica monitoring methods for the laboratory analysis of respirable dust samples is considered accurate, but it comes with the potential downsides of mineral interference, high cost considering numerous sample analysis, multiple days to weeks of waiting to receive results, and this analysis cannot traditionally be done on-site. For the field-based method, all the downsides have the potential to be addressed using portable FTIR in conjunction with predictive modeling. While not as accurate as a controlled lab-based data collection environment (Hart *et al.*, 2018), the implementation of field-based methods (portable FTIR) brings fast, high-quality data collection to the field and provides the potential for further downstream data analysis methods like predictive modeling. Portable FTIR does come with the cost of potential accuracy issues in data interpretation when compared to current silica monitoring methods and lacks the presence of an expert analyst who is familiar with this analytical technique like a traditional silica analysis would have. Recently, researchers from around the world have turned to the use of chemometrics to help increase RCS analysis accuracy using portable FTIR data (Weakley *et al.*, 2014; Miller *et al.*, 2017; Stach *et al.*, 2020; Wei *et al.*, 2020; Salehi *et al.*, 2021; Stach *et al.*, 2021).

The most used peak for silica determination is 800 cm^{-1} which results from the absorption of those frequencies by the stretching of the Silicon–Oxygen (Si–O–Si) bond. However, this same stretching absorbs similar energy in other forms of silica and silicates (Ojima, 2003), where the same bond is present in different configurations. To complement the 800 cm^{-1} peak, many analysis methods utilize the quartz doublet that occurs at approximately 767–816 cm^{-1}, which accounts for the

major bands at 780 and 800 cm^{-1} (Foster and Walker, 1984). Other minerals are known to have bands close to 800 cm^{-1} such as cristobalite, tridymite, and amorphous silica and are included in analytical methods such as NIOSH 7603/MSHA P7 as known confounder minerals. Given the complex nature of respirable dust mixtures in mining (Walker *et al.*, 2021), there is an increasing need to account for multiple variables quickly and efficiently during analysis. As RCS regulatory limits continue to decrease the ability to measure accurately and reproducibly for more complex samples with lower RCS is needed to make sure lower standards in the future are able to be monitored and analyzed correctly. Multivariate regression methodologies like Principal Component Analysis (PCA), Principal Component Regression (PCR), and Partial Least Squares (PLS) are employed in a variety of analytical fields and are commonly used for predicting measurement values (Pottel, 1995; Rahman *et al.*, 2014; Camacho *et al.*, 2016). One of the main reasons that these models have been developed is to confront the ever-growing probability that there are many more predictor variables than there are samples. The use of predictive modeling is a huge benefit when the variables that are being predicted are costly, time-consuming, or laborious to produce in other ways.

Combining the benefits of portable FTIR with chemometric modeling techniques can solve the problem of accuracy and still maintain the cost and time benefits awarded by portable FTIR methods. One of the main goals for the development of this computational framework or pipeline is to make the process and interface user-friendly and be universally applicable to all samples. In computing, a pipeline is a computational or data processing framework that streamlines the process of moving data from a raw format to a more analyzed format through a series of analytical pathways that are laid out in a reproducible and simplified way. Utilizing open-source products and publicly archiving the code allows for full transparency of the methods employed and for further development from sources outside the authors of this article. The implementation of this method should be easily deployable for any user regardless of their experience with multivariate modeling, computer science, or RCS analysis. Additionally, it is intended for this pipeline to be used as a backbone for future modeling efforts which can be implemented for minerals other than silica of a new type of data is generated.

With the creation of this analytical pipeline, we intend to create and explore the practicality of a pipeline that can predict the amount of RCS deposited on a filter (Figure 4). These modeling efforts rely heavily on the previous research into the field of RCS and portable field-based FTIR sample collection and aim to progress the field further down the path of near real-time hazard monitoring. In this article, we employ both PCA for outlier detection and PLS for predictive modeling all done in the MATLAB environment using a user-friendly GUI interface called PLS_toolbox. The data was preprocessed via mean-centering and cross-validated using venetian blinds with 10 splits. The methods and final models put forward here do not solve all issues related to the development and implementation of a multivariate RCS prediction model, but the need to document the development of the pipeline as the methods progress is understood.

2. Methods

Respirable dust samples were collected during laboratory studies at the NIOSH Pittsburgh Mining Research Division (PMRD) facilities and were used to investigate the development of multivariate chemometric predictive models. The samples were the product of three distinct sampling protocols, ultimately creating three distinct datasets. Characteristics of the samples from each group are summarized in Table 1.

Dust investigated

Respirable samples of crystalline silica were prepared from Minusil 5 and Minusil 10 ground silica (U.S. Silica, Frederick, MD), which is reported by the manufacturer to be 95%–99.9% crystalline. The mass median particle diameter is 1.6 μm and 3.4 μm for the two powders when aerosolized, respectively, as measured with an Aerodynamic Particle Sizer (3321, TSI, Sant Paul, MN). To create samples as mixtures of two minerals, minusil 5 was used as reliable and high-purity quartz material, and 13 other mineral mixtures (previously identified in

Table 1 – *Summary of the three sampling protocols used to create the modeling datasets in the study.*

Protocol	Type of dust samples	Number of samples	RCS (or quartz) range	Sampler used
1	Minusil 5 or Minusil 10	263	0 – 300 μg	Dorr Oliver cyclone
2	Minusil 5 + another mineral phase	320	0 – 1500 μg	Dorr Oliver cyclone
3	Minusil 5	480	0 – 600 μg	Multiple samplers

mine dust samples) were used as confounders: Albite, Amorphous silica, Anorthite, Chlorite, Cristobalite, Dolomite, Kaolinite, Magnetite, Muscovite, Oligoclase, Pyrite, Talc, and Tridymite. Additional information regarding the purity, source, processing, and other relevant details for the confounding minerals can be found in Supplemental Table 1. The confounding minerals did not contain crystalline silica which was confirmed using X-ray Diffraction (XRD). Each mineral was independently layered during separate sampling on top of the minusil 5 at varying ratios and then analyzed in the same way as the minusil samples. Multiple samples for each mineral were taken ranging from 0 – 1700 µg to closely mimic the distribution seen in the quartz samples (0 – 1500 µg) and evenly spread the confounding mineral data throughout the model.

Collection of samples

The samples for the three sampling protocols were collected in an aerosol chamber with the size of 71 cm x 66 cm x 56 cm. For protocols 1 and 3, the dust was aerosolized using a fluidized bed aerosol generator (model 3400 A, TSI, Shoreview, MN). For protocol 2, due to the limited amount of the mineral materials, the dusts were introduced into the aerosol chamber using a 250-mL flask closed by a rubber stopper. The stopper had two ports into which two steel tubes were connected; one tube let airflow into the flask to aerosolize the dust, and the other tube sucked aerosol from the flask into the aerosol chamber. For most of the sampling testing, the aerosol passed through a Kr-85 aerosol neutralizer (Model 3012 A, TSI, Shoreview, MN) before being introduced through the top of the chamber. A stable dust concentration within the chamber was maintained through each test by balancing the inlet airflow rate and the negative force induced by a fan at the bottom of the chamber. A high efficiency particulate air filter was present at the bottom of the chamber in front of the exhaust fan. A slight negative pressure was maintained in the chamber, and the respirable dust concentration was monitored by a tapered element oscillating microbalance (TEOM) (model 1400 A, Thermo Scientific, Waltham, MA).

Respirable dust samples were collected using 10-mm nylon Dorr-Oliver cyclones at a flow rate of 1.7 lpm for protocols 1 and 2. For protocol 3, samples were collected using six different samplers (Dorr-Oliver, Sensidyne Dorr-Oliver, conductive Dorr-Oliver, GS3, GK2.69, or Aluminum). The standard flow rate for each sampler was used (Baron, 2016). Dust samples were always collected on 37-mm, 5-µm pore PVC filters (SKC Inc., Eighty Four, PA). The sampling filters in protocol 1 were housed in four-piece conductive cassettes with a metal support ring (Chubb and Cauda, 2021). During protocol 2, the sampling filter was housed in a three-piece styrene cassette with a cellulose support pad (SKC Inc., Eighty Four, PA). Finally, for protocol 3, to also examine the effects the cassette type may have on predictive modeling, samples were collected using either a four-piece conductive cassette with a metal support ring or a three-piece styrene cassette with a cellulose support pad. Each sampler was placed inside the chamber and connected to a sampling pump (SKC AirChek TOUCH pump) placed outside the chamber. Sampling ports on the side of the chamber allowed the connection. Sampling pumps were shut off when a target loading was reached. The collection of the desired sample was carried out in one single sampling event for protocols 1 and 3. For protocol 2, quartz was first sampled, and then the second mineral phase was sampled as an additional layer on the same filter after gravimetric and spectroscopic analysis of the quartz samples were conducted.

Analysis of samples

Gravimetric analysis was done for the blank filters and loaded samples to determine the gravimetric mass of the sample. In the case of protocols 1 and 3, this is considered the reference quartz value and is used as the Y variable when training the predictive models. For protocol 2, the gravimetric analysis of only quartz is used as the Y variable. All filter samples were pre- and post-weighed in a temperature- and humidity-controlled weighing room using a model XP2U microbalance (Mettler Toledo, Columbus, OH). The sample sets included approximately 5% lab blanks to verify accurate measurement from the microbalance. Analysis of blanks indicated an average gain of 1 ± 2 µg. The standard deviation (SD) of the blanks gives a good indication of both the limit of detection (LOD) and limit of quantification (LOQ). Analytical best practices indicate that LOD = SD(blanks)*3 and LOQ=SD(blanks)*5 which equates to a LOQ of around 15 µg. Following gravimetric analysis, samples were analyzed on a portable transmission infrared spectroscopy instrument (ALPHA model, Bruker Optics, Billerica, MA); each sample spectrum consisted of 16 co-added scans (total analysis time of approximately 20 seconds) collected at a resolution of 4 cm^{-1} using Blackman Harris apodization. The resulting spectra for each dust sample is used as the X variable for the chemometric modeling. Each sample was inserted in the instrument so that the instrument's IR beam passed through the center portion of each sample. Following

FTIR analysis, the data was opened in Essential FTIR (Operant LLC, Monona, WI), then exported into Excel (Microsoft Inc.), and the background spectrum (clean filter before sample collection) was subtracted from the final spectrum. The collection of the background spectrum, necessary for testing the preprocessing of each sample, was carried out in a variety of ways. For protocol 1, the same filter media before sampling was used as the background to allow for the background removal of the FTIR spectrum unique to each filter for the samples from each dataset. For protocols 2 and 3, a blank PVC filter (from the same lot as the sample filters) was used as the background. For protocol 2, both gravimetric and FTIR analyses were performed multiple times to generate data from the quartz only as well as the quartz with the addition of the confounding mineral.

Modeling of the data

The primary objective of all modeling in this study is the prediction of the amount of quartz in each sample. As stated above, the gravimetric mass of quartz in the sample is used as the Y variable when modeling, and the FTIR spectrum values are used as the X variable. For the entirety of the data presented in this article, all principal component analysis (PCA) and partial least square (PLS) modeling was done in MATLAB using PLS_toolbox (v8.9.1; https://eigenvector.com/software/pls_toolbox-and-solo-interfaces/). Secondary validation for the PLS models was run using the Partial Least Squares and Principal Component Regression package (v2.7.3; https://github.com/bhmevik/pls) in the programming language of R which resulted in very similar results, and for the sake of brevity are not discussed. The first step in the modeling is to center the spectral data to its mean value and derive a novel set of coordinates (eigenvectors) which are used for downstream analysis. Next, the model discerns features within the spectrum that are the most variable between samples in the dataset and isolates them into components. These unique features, or components, are reduced to the smallest possible number while only retaining a select number of components that have the largest influence on the inter-sample variance and are defined as principal components (PCs) when running PCA or latent variables (LVs) when running PLS. As a first step in the modeling process, PCA was implemented which looks for a few linear combinations of variables that can be used to summarize the data without sacrificing too much information in the process. The data used to build the models found in this manuscript were generated in a laboratory environment using the same machine every time which makes it unlikely for a sample generated in house using the same testing conditions to be flagged as an outlier using PCA. This , however, remains a critical step in any prediction modeling endeavor and should not be overlooked as a sample that is flagged as an outlier using PCA will give an erroneous prediction for that sample and should not be allowed to progress down the pipeline to the prediction steps. As PCA is an unsupervised method of dimensionality reduction, it is used first to identify broad abnormalities within the dataset and detect any outliers that may be present. PCA assigns no importance to the relationship between the X and Y variables, which is what makes it unsupervised, while PLS relies heavily on the relationship between the X and Y variables. The main application of our chemometric modeling is using our known X and Y variables to predict unknown samples, and in this application, PLS is much better suited. PLS builds the LVs using the raw X variables as well as the relationship between the X and Y variables to relate a select number of underlying factors to the Y variable.

A normalization step was created wherein the raw data received from the FTIR is forced to a fixed number of wavenumber variables to account for data coming from various sources, all of which have different data densities regarding wavenumber. To accomplish this, a spline is fit to the data using the inbuilt spline function in R, and a fixed number of data points is extrapolated and used as wavenumbers for all further analysis. The density of the data did not impact the prediction accuracy of the overall model when the data was reduced to 500 variables, or when the data was increased to 10,000 variables. Increasing the data density does, however, impact the computation time required for the PLS prediction. As all of the data used to train the models at this stage was generated using the same FTIR instrument the normalization step was not necessary, and all data is at the default output for the Bruker ALPHA model of 1,765 variables for all models discussed.

3. Results

The first evaluation of the data was aimed to test out potential preprocessing methodologies and to accomplish this using the minusil 5 dataset. A small model was created using these spectra and used to compare various blank subtractions and baseline correction methods (Figure 1). The baseline for this comparison was the raw spectrum for each sample, which is the spectrum generated by the FTIR analyzer, and based on the analysis setup it still includes the effect of the filter matrix. The first preprocessing approach tested was subtracting the spectrum of the blank filter used for sample collection. This is performed to remove the matrix effects

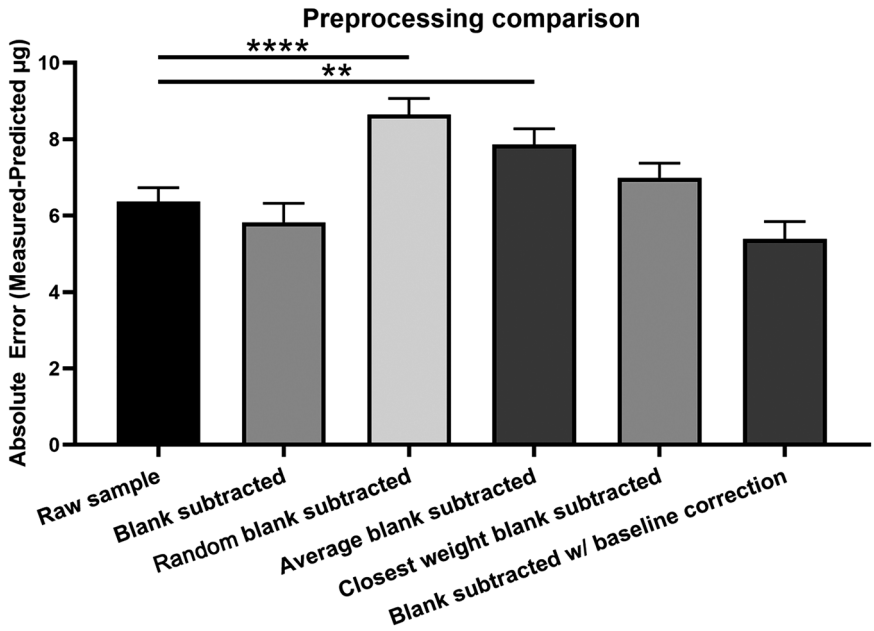

Figure 1. Preprocessing Comparison. Bar chart comparison between six potential preprocessing methods. All comparisons are by T test against the raw sample column and are not significant unless listed. Note the best performing data is with the blank subtracted and baseline corrected, but not significantly better than the raw data. ****; P<0.0001, **; P<0.01.

caused by the IR absorption of the filter media, leaving just the spectrum of the collected dust to be analyzed. This approach improved the absolute error (|measured quartz – predicted quartz| µg) compared to the raw spectra samples, but not significantly. Baseline correction was then implemented to the raw sample and the blank subtracted sample, both of which reduced the prediction error, but the beneficial effects were not significant. Since in practice the spectral profile of the filter pre-collection is not always available, the effects of alternative methods of filter subtraction were tested. All the filter blank spectra were used to average the spectra together at each wavenumber creating a spectrum that was representative of all the blanks, which was used as an average blank. Subtracting a random blank as well as subtracting the average blank both significantly diminished the predictive capacity of the model compared to the raw data, while subtracting out the closest blank by filter weight marginally increased the predictive capacity. With the best average absolute errors (blank subtracted with baseline correction) and the worst average absolute error (random blank subtraction), both having relatively good absolute error values of 8.63 µg and 5.39 µg respectively, it was thought to be best to keep all spectra in the raw form and apply no preprocessing before the modeling investigation.

The first series of data that were considered for the development of predictive models for quartz was comprised of two sets of minusil 5 data and one minusil 10 dataset (protocol 1). Initially, each of the three smaller datasets was modeled independently but were quickly pooled together as we found the root mean squared error of calibration (RMSEC) was much lower than the larger the training dataset. Additionally, there were no major differences when comparing a sample against the minusil 5 or minusil 10 models independently. The first model implemented is called Model 1 (M1) and is comprised of the full spectrum 400–4000 cm^{-1} from the combination of the minusil 5 and minusil 10 data (n263). We next thought that we could improve the RMSEC of the model if we cut out all the variables of the IR spectra that are not providing information relevant to the quartz prediction and thus created our second model (M2). This area narrows in on the quartz doublet feature that occurs at approximately 767–816 cm^{-1}, which results from stretching of the Si–O bond. Using the same samples and simply narrowing in on the quartz doublet region of the spectrum marginally improved the RMSEC as well as the average absolute error (Figure 2A-B). With an average absolute error of around 6.5 µg, M1 and M2 predict well, the goal was to increase the training data as well as samples with a higher loading of quartz.

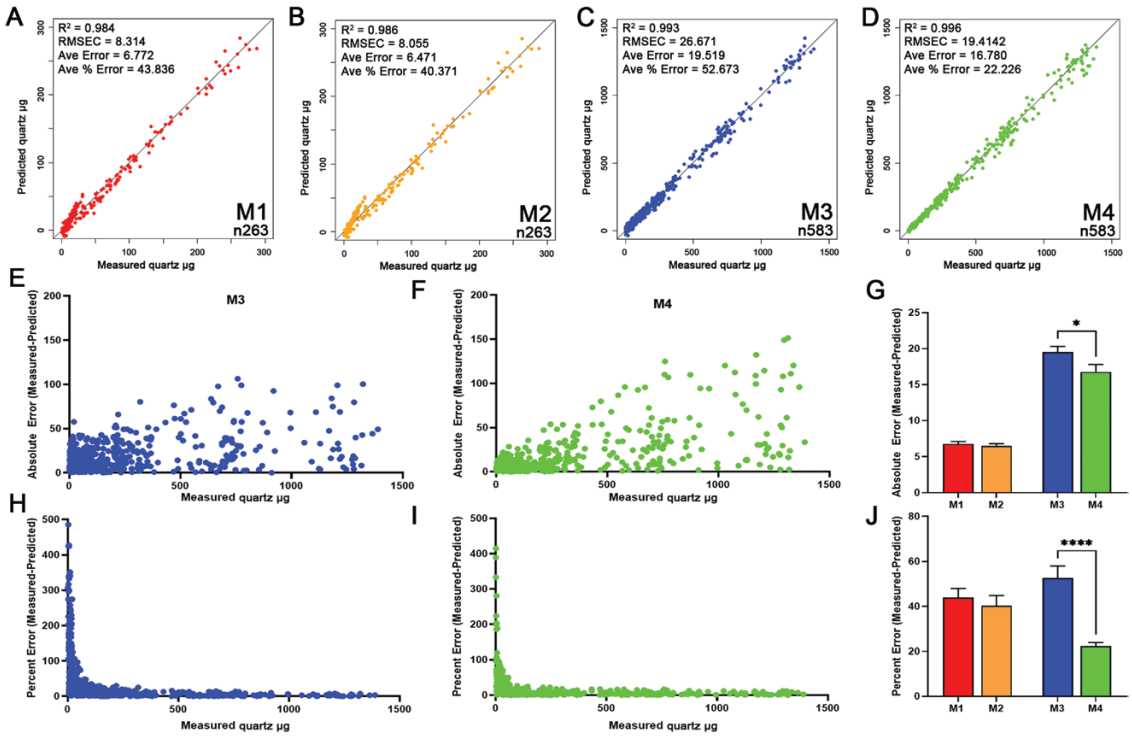

Figure 2. M1-M4 Predicted vs. Measured. A-D.The gravimetrically measured quartz µg vs the PLS predicted quartz µg for models M1–M4 respectively. Note the higher loading samples found in M3 and M4. Evaluation of the absolute error difference between the measured and predicted quartz values vs the measured quartz values for M3 (E), M4 (F). Bar plot depicting the average absolute error from M1–M4 (G). Evaluation of the percent error ((absolute error / measured quartz µg)*100) vs the measured quartz values for M3 (H), M4 (I). Bar plot depicting the average percent error from M1-M4 (J). ****; P<0.0001, *; P<0.05.

The way the data was generated from protocol 2 allowed for the acquisition of initial FTIR spectra when only quartz was present on the filter followed by quartz and a confounding mineral. The third and fourth models (M3 and M4) use the quartz only spectra to further increase the training data up to 583 samples and add samples with a much higher quartz loading (from protocol 1-2). These additions ultimately increase the RMSEC as well as the average error compared to M1 and M2, but this effect is predominantly driven by the addition of the higher loading samples (Figure 2C–D). Higher loading samples are classified as any sample with > 300 µg quartz and by nature skew the absolute error higher and the percent error lower than samples at a much lower loading (1000 µg sample with 10% error is ± 100 µg, while a 100 µg sample ± 10% is only 10 µg). Interestingly, M4 had the lowest percent error of all four models being significantly lower than both M3 and M2, which is mostly due to more accurate predictions in the lower quartz loading samples (<100 µg, Figure 2H-J). As expected M4, which is the quartz doublet region, predicted better than its full spectrum counterpart M3 with significant reductions in both absolute error and percent error (Figure 2E-J). As the training data within any given model begins to account for higher loading samples, the RMSEC and average error both increase, while the percent error falls. Additionally, the opposite is true wherein if the models were capped to only include samples less than 200 µg the RMSEC and average error both would decrease, while the percent error would increase. Supplemental Figure 1A-D shows the PCA scores across LV 1 and LV 2 along with the Q residual (Qres) and Hotelling's T^2 (T^2) scores for each model. Both M2 and M4 are limited to only the quartz doublet and as such have much higher variance captured by LV1 since the data is truncated to the area that has the most inter-sample variance compared to M1 and M3.

The second series of data implemented into the modeling was to include data generated from combining minusil 5 with a second mineral. These additional 237 samples, generated from unique combinations of minusil 5 with one of the 13 confounding minerals, were added to the 583 from M3 and M4 before and used to create a full spectral model (M5) as well as a quartz doublet model

(M6). As with each of the models before, M6 outperforms M5 in all metrics further confirming that the analysis of the quartz doublet yields a more accurate prediction than the analysis of the full spectrum (Figure 3A-D). The confounder models have much higher Qres and T^2 values compared to the quartz-only models (Supplemental Figure 1E-F) due to the inclusion of different minerals modifying the shape and intensity of the quartz doublet. When comparing M6 to M4, the RMSEC, average error, and percent error all go up with the inclusion of the confounder data, but not significantly. This indicates that the inclusion of a more complex dust mixture allows for the model to be trained on a more diverse dataset without sacrificing the integrity of the quartz prediction.

4. Discussion

The goal of the data presented in this article is to create and train a multivariate model capable of predicting the amount of quartz deposition on a filter without compromising on laboratory analysis accuracy. As we continued to train and test our models, the need to document and create a systematic pipeline of steps that allow for this to happen in a streamlined and user-friendly way. Figure 4 is the proposed pipeline that is built around the use of parts of the above-discussed models along with data normalization and quality control steps. One of the main goals of the study was to simplify the process as much as possible and with no immediate significant improvement to the prediction when doing either baseline correction or blank filter subtraction the simplest path does not include any data preprocessing steps. For the testing of these models, three metrics predominantly were relied upon: RMSEC, average error, and percent error to lay the groundwork for the pipeline, and all the samples in the training data are synthetic mixtures of known quartz amounts. Using these three metrics allows for the observation of the model as a whole and for determining the overall fit and predictive capacity. Within a model, the use of the PCA scores, Qres, and T^2 were all used to identify specific samples flagged as outliers compared to the model. Potential outliers could arise in this type of data at the time of sample collection, or FTIR analysis and often are hard to identify without the use of unsupervised learning methodologies like PCA, which can find samples that are outliers for reasons that are identifiable or unknown as well. Outliers might arise during sample collection through nonsymmetric deposition of respirable dust due to sampler malfunction or during FTIR analysis by incorrect placement or alignment of sample within the instrument. These quality control (QC) steps and parameters will ultimately become the first steps in the pipeline to assess that any sample put into the model fits well enough to have a degree of confidence in the resulting quartz prediction.

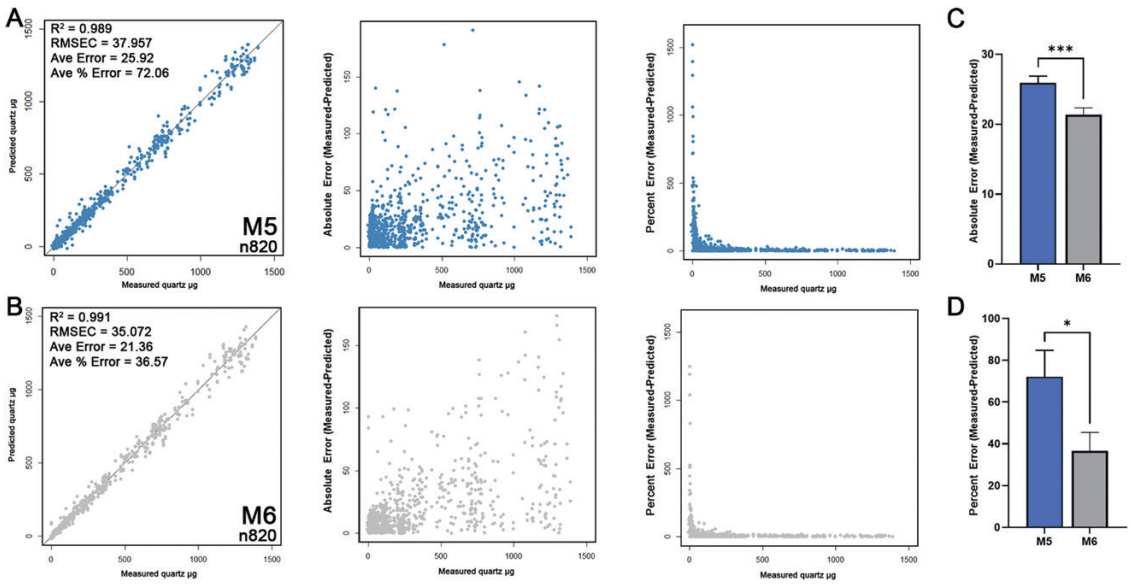

Figure 3. Summary statistics for M5 and M6. A-B. The gravimetrically measured quartz µg vs the PLS predicted quartz µg for models M5 and M6 respectively. Evaluation of the absolute error difference between the measured and predicted quartz values vs the measured quartz values and percent error ((absolute error / measured quartz µg)*100) vs the measured quartz values. Bar plot depicting the average absolute error from M5 and M6 (C). Bar plot depicting the average percent error from M5 and 64 (D). ***; P<0.001, *; P<0.05.

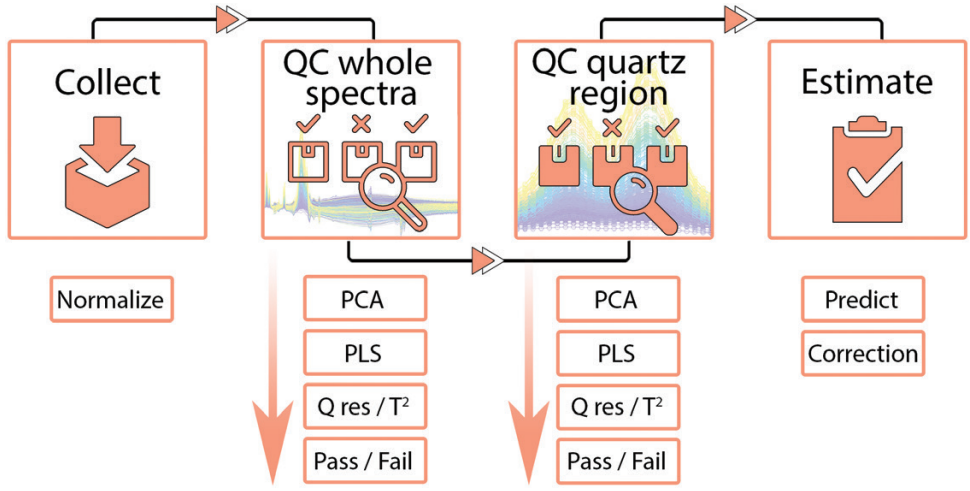

Figure 4. Proposed PLS pipeline. Schematic illustration of the progression of raw data through the four major steps of the proposed pipeline.

Following the normalization step laid out in the Methods section, the second step is to compare the sample in question to the full spectral model build using 583 samples containing either minusil 5 or minusil 10. This initial step is a broad comparison in which the model can easily determine if there were any problems in the acquisition or normalization of the data. First, a principal component analysis (PCA) step is performed to ensure that the sample falls within the allotted confidence 95% interval. If the sample passes the PCA step, it moves on to the partial least squares model where Qres and T^2 values are created for the sample. T^2 and Qres are summary statistics that are used to inform about how well a model fits a given sample and why that sample is assigned its scores. Qres indicates how well each sample conforms to the model and is a measure of the difference between the sample and its projection through the model. In contrast to Qres, which represents the amount of variation retained after projection through the model, the T^2 value indicates a measure of the variation within the model. These two scores represent a measure of goodness of fit between the sample and the model with lower scores indicating the data and model exhibit similar patterns across the PCs or latent variables retained in the model. In addition to exploratory analysis and dimensionality reduction, PCA is often used as a method of outlier detection in chemometrics. PCA analysis will assign a score for each sample to indicate how far the sample is from the centroid of the elliptical center that includes most of the data. This score in conjunction with the Qres and T^2 allows for robust quality control (QC) metrics generated for each sample before any modeling begins. A series of the above-discussed models and QC metrics were implemented into our current analysis pipeline as graphically depicted in Figure 4. Within the second step of the pipeline is an additional model (M5) which is comprised of the same 583 samples with an additional 237 samples (820 total) that were a mixture of minusil with one of the 13 confounding minerals. As can be seen in Figure 5B and D, the model containing the confounding minerals has much more spectral variability and represents a more complex dust environment.

If the sample passes the first QC step, it moves on to the third step in the pipeline which is the QC step specific to the quartz doublet region of the spectra. The sample progresses through the same PCA and PLS steps as before but is no longer looking for broad spectral anomalies but is specifically targeting the quartz signature region and assessing sample integrity compared to the model. Once again, the addition of the confounding minerals (M6) in the 820 model increased the variability found in the spectra compared to the 583 quartz-only model (Figure 5B, D). If the sample passes all the quality control steps, the final prediction is made using PLS on the 583 quartz-only quartz doublet region model. This gives the cleanest and most accurate prediction and can be compared to the predicted result generated from the 820 model to indicate the potential impact of confounders. If the two models predict very similar results, then it is less likely that there are confounding minerals in the sample, while if the two models differ significantly in their prediction, then that can be also used as an indication of the potential presence of confounders.

The study has known limitations. The current training of the models has all been done on samples generated

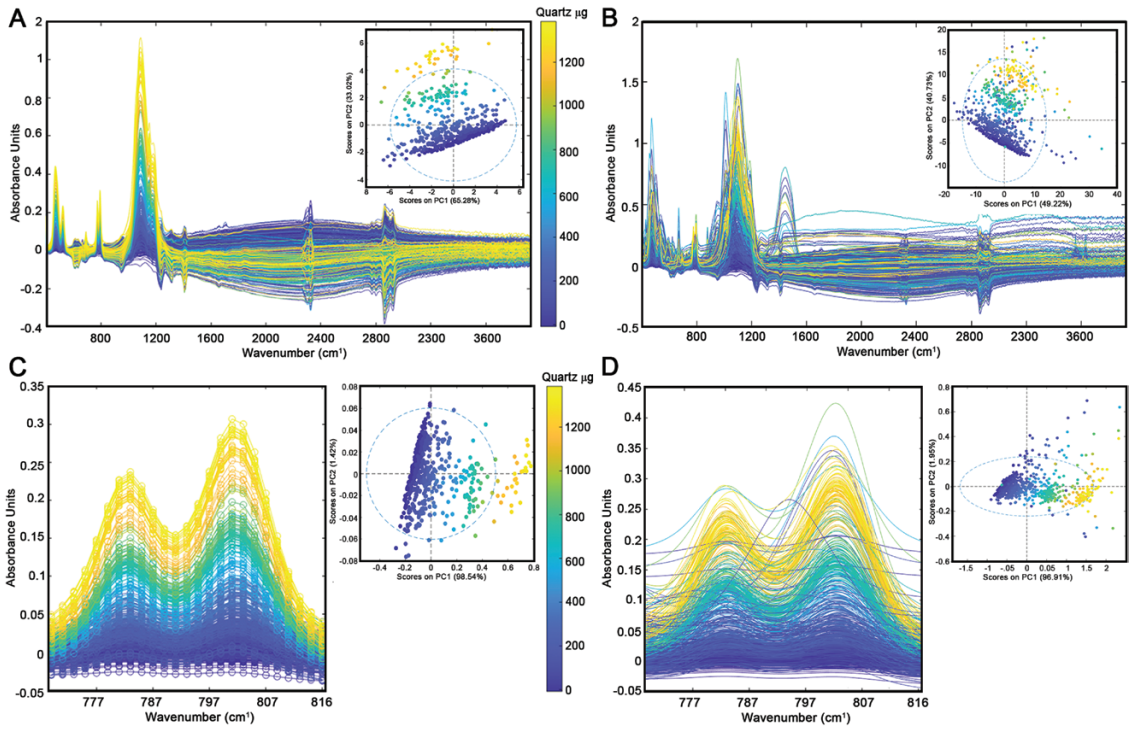

Figure 5. Quality control steps for the full FTIR spectra. A. 583 FTIR spectra overlaid which were generated using quartz-only samples. Inset PCA shows the most extreme loading samples (both high and low) are the most disparate compared to most of the model. B. 820 FTIR spectra which were generated using the 583 quartz-only samples as well as samples generated in combination with one of the 13 confounding minerals. The same samples used in the full spectral model were cut down to further investigate the 767-816 cm-1 region of the spectra which contains the primary quartz signature. C. 583 FTIR spectra overlaid which were generated using quartz-only samples. D. 820 FTIR spectra which were generated using the 583 quartz-only samples as well as samples generated in combination with one of the 13 confounding minerals. Note the relative spectral cleanliness of panel A and C compared to the addition of a single confounding mineral in B and D.

in the laboratory under very controlled conditions. This allows for greater certainty as to the true quartz gravimetric value for each sample but comes at the cost of reducing sample complexity. To combat that and to add some robustness into the model, the mixtures of quartz and one of the 13 confounding minerals were included. As can be seen in Figures 5 and 6, this increases the diversity of the shapes and intensity of the FTIR spectrum while still allowing high confidence in the true quartz amount in each sample. This artificial complexity pales in comparison to the complexities seen from real-world dust samples but is a way to gradually add complexity to the training data while still retaining a high degree of confidence in the reference quartz value (Walker *et al.*, 2021). Using a combination of FTIR, XRD, and standard addition techniques, the plan is to begin creating and validating a model which is comprised solely of real-world mine dust samples and be able to incorporate this model into the pipeline for another layer of predictive power. The FTIR analysis (performed on a respirable mine dust sample deposited on a PVC filter) will provide the FTIR test spectra (the X variables). The combination of XRD and standard addition calibration will produce a more accurate silica reference value (the Y variables) for the sample compared with other techniques, such as standardless XRD analysis. In all the models, most of the error is derived from the samples with higher quartz loadings and improved RMSEC and average error when limiting each of the models to only include data with less than 300 µg of quartz which would be more indicative of a real-world actionable exposure.

There are a nearly unlimited number of downstream modifications to make to the pipeline along with the addition of mine dust samples to improve prediction accuracy. The first and easiest to implement would be to continually increase the number of samples in the training data. The more data the model has access to the higher the confidence will be in the prediction. The model as it is now does not incorporate any form of baseline correction because it was felt to be best to keep the data as close to the raw FITR spectrum as possible. This is one area of potential improvement which can be coded into the normalization

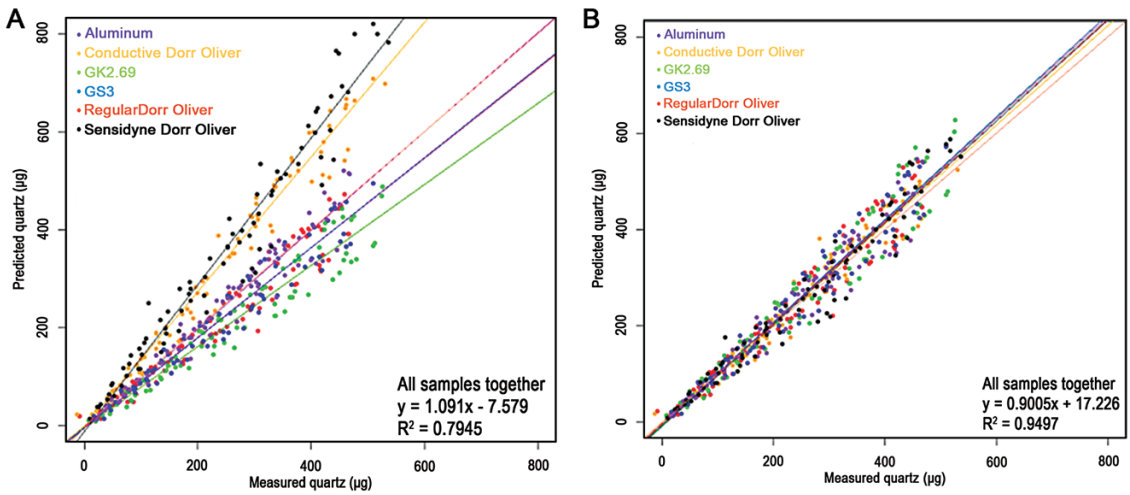

Figure 6. Application of correction factor for sampler and cassette used. Scatterplot of the predicted vs measured quartz content before correction (A) and following correction for sampler type (B). Note the increase in R2 following the sampler specific correction (0.7945 to 0.9497).

steps of the pipeline and allow for the potential for a recalibration of samples to match the settings and type of FTIR instrument used to generate the calibration data. Additionally, by using the difference between the prediction of the quartz only and the quartz + confounders models (M4 and M6) and the Qres and T^2 from the quartz-only model, the general level of confounders can be determined. If a sample is classified as having a high level of confounders in the quartz doublet region, the model could potentially fall back onto secondary or tertiary spectral regions to make a more accurate prediction.

Previous research (Cauda *et al.*, 2014) lead us to additionally generate data on and compare various types of samplers and cassettes used in sampling to assess if there was an impact on the PLS predictive capacity. In this regard, 480 samples taken from six sampler types and two cassette types were compared creating 12 unique sampler/cassette combinations to determine if there was any error introduced into the model based solely on sampling technique. All the data used to create the models M1-M6 described here were collected using a Dorr Oliver sampler. It was determined that the samples generated on the same sampler/cassette combination as the training data had the lowest prediction error, indicating that the sampler and cassette have a substantial effect on the prediction, and the effect can be accounted for through modeling. The effect of this can be seen in Figure 6A where the Dorr Oliver samples (red dots) have the most linear relationship between the actual value and the predicted value. Some samplers lead the model to over predict the quartz on the filter (Sensidyne Dorr Oliver and the conductive Dorr Oliver) while other samplers lead the model to underpredict like the aluminum and GK2.69 samplers. By calculating correction factors and applying them to each sample based on the sampler/cassette used, a significant increase in predictive power (R^2 = 0.7945 before correction (A) to R^2 = 0.9497 after (B), Figure 6) were observed, thus, further emphasizing the need for ongoing enhancement of any computation pipeline such as suggested here. This observed difference to over or under predict when the model and test data are collected on different samplers is most likely due to the differences in the dispersion patterns of the airborne material when deposited on a filter as previously studied (Miller *et al.*, 2013).

5. Conclusions

This study set about addressing the feasibility of implementing a combination of principal component analysis and partial least squares on FTIR spectrometer data to accurately predict the amount of crystalline silica deposited on a filter using lab-made mixtures of silica and various other minerals commonly found in the mining environment. The key driver for this research is to be able to create automatic multivariate predictive models to be included in a monitoring approach for on-site silica exposure predictions that are user-friendly and able to be continually trained and improved without sacrificing accuracy. The impact of this research will lead to a better understanding of the amount, types, and quality of real-world mine dust samples that need to be collected to train a computational model to predict RCS deposition on a filter to be used for accurate end-of-shift silica exposure predictions.

SUPPLEMENTARY DATA

Supplementary data are available at *Annals of Work Exposures and Health* online.

Conflict of Interest

The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the National Institute for Occupational Safety and Health, or the Centers for Disease Control and Prevention. Mention of any company or product does not constitute an endorsement by NIOSH or CDC. The authors declare no other conflict of interest relating to the material presented in this paper.

Funding

The publication was funded by the National Institute for Occupational Safety and Health and the Centers for Disease Control and Prevention.

Data Availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

Ashley EL, Cauda E, Chubb LG *et al*. (2020) Performance Comparison of Four Portable FTIR Instruments for Direct-on-Filter Measurement of Respirable Crystalline Silica. *Ann Work Expo Health*; **64**: 536–46.

Baron PA. (2016) Factors affecting aerosol sampling. *NIOSH Manual of Analytical Methods: Department of Health and Human Services*. Centers for Disease Control and Prevention. National Institute for Occupational Safety and Health. https://www.cdc.gov/niosh/docs/2014-151/pdfs/chapters/chapter-ae.pdf

Camacho J, Pérez-Villegas A, García-Teodoro P *et al*. (2016) PCA-based multivariate statistical network monitoring for anomaly detection. *Computers & Security*; **59**: 118–37.

Cauda E, Chubb L, Miller A. (2016) Silica adds to respirable dust concerns: what if you could know the silica dust levels in a coal mine after every shift? *Coal Age*; **121**: 31–3.

Cauda E, Chubb L, Reed R *et al*. (2018) Evaluating the use of a field-based silica monitoring approach with dust from copper mines. *J Occup Environ Hyg*; **15**: 732–42.

Cauda EG, Drake P, Lee T, Pretorius C. (2014) High-volume samplers for the assessment of respirable silica content in metal mine dust via direct-on-filter analysis. *10th International Mine Ventilation Congress, IMVC2014 South Africa*.

Chubb LG, Cauda EG. (2021) A novel sampling cassette for field-based analysis of respirable crystalline silica. *J Occup Environ Hyg*; **18**: 103–9.

Foster RD, Walker RF. (1984) Quantitative determination of crystalline silica in respirable-size dust samples by infrared spectrophotometry. *Analyst*; **109**: 1117–27.

Hart JF, Autenrieth DA, Cauda E *et al*. (2018) A comparison of respirable crystalline silica concentration measurements using a direct-on-filter Fourier transform infrared (FT-IR) transmission method vs. a traditional laboratory X-ray diffraction method. *J Occup Environ Hyg*; **15**: 743–54.

Leung CC, Yu IT, Chen W. (2012) Silicosis. *Lancet*; **379**: 2008–18.

Miller AL, Drake PL, Murphy NC *et al*. (2012) Evaluating portable infrared spectrometers for measuring the silica content of coal dust. *J Environ Monit*; **14**: 48–55.

Miller AL, Drake PL, Murphy NC *et al*. (2013) Deposition Uniformity of Coal Dust on Filters and Its Effect on the Accuracy of FTIR Analyses for Silica. *Aerosol Sci Technol*; **47**: 724–33.

Miller AL, Weakley AT, Griffiths PR *et al*. (2017) Direct-on-Filter α-Quartz Estimation in Respirable Coal Mine Dust Using Transmission Fourier Transform Infrared Spectrometry and Partial Least Squares Regression. *Appl Spectrosc*; **71**: 1014–24.

Ojima J. (2003) Determining of crystalline silica in respirable dust samples by infrared spectrophotometry in the presence of interferences. *J Occup Health*; **45**: 94–103.

Pottel H. (1995) The use of partial least squares (PLS) in quantitative FTIR: determination of gas concentrations in smoke gases of burning textiles. *Fire Mater*; **19**: 221–31.

Rahman Z, Siddiqui A, Bykadi S *et al*. (2014) Near-infrared and fourier transform infrared chemometric methods for the quantification of crystalline tacrolimus from sustained-release amorphous solid dispersion. *J Pharm Sci*; **103**: 2376–85.

Salehi M, Zare A, Taheri A. (2021) Artificial Neural Networks (ANNs) and Partial Least Squares (PLS) Regression in the Quantitative Analysis of Respirable Crystalline Silica by Fourier-Transform Infrared Spectroscopy (FTIR). *Ann Work Expo Health*; **65**: 346–57.

Stach R, Barone T, Cauda E *et al*. (2020) Direct infrared spectroscopy for the size-independent identification and quantification of respirable particles relative mass in mine dusts. *Anal Bioanal Chem*; **412**: 3499–508.

Stach R, Barone T, Cauda E *et al*. (2021) A Novel Calibration Method for the Quantification of Respirable Particles in Mining Scenarios Using Fourier Transform Infrared Spectroscopy. *Appl Spectrosc*; **75**: 307–16.

Walker R, Cauda E, Chubb L *et al*. (2021) Complexity of Respirable Dust Found in Mining Operations as Characterized by X-ray Diffraction and FTIR Analysis. *Minerals*; **11**: 383.

Weakley AT, Miller AL, Griffiths PR *et al*. (2014) Quantifying silica in filter-deposited mine dusts using infrared spectra and partial least squares regression. *Anal Bioanal Chem*; **406**: 4715–24.

Wei S, Kulkarni P, Zheng L *et al*. (2020) Aerosol analysis using quantum cascade laser infrared spectroscopy: application to crystalline silica measurement. *J Aerosol Sci*; **150**: 105643.