Nontuberculous Mycobacterial Infection and Environmental Molybdenum in Persons with Cystic Fibrosis: A Case-Control Study in Colorado.

Ettie M. Lipner, Ph.D., M.P.H.[a,b], James L. Crooks, Ph.D, M.S.[b,c], Joshua French, Ph.D[d],

Michael Strong, Ph.D[a], Jerry A. Nick M.D.[e*] and D. Rebecca. Prevots, Ph.D, M.P.H.[f*]

[a]Center for Genes, Environment and Health, National Jewish Health, Denver, CO USA, Email: StrongM@NJHealth.org; [b]Department of Epidemiology, Colorado School of Public Health, Aurora, CO USA; [c]Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, CO USA, Email: CrooksJ@NJHealth.org; [d]Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO USA, Email: JOSHUA.FRENCH@UCDENVER.EDU; [e]Department of Medicine, National Jewish Health, Denver, CO USA, Email: NickJ@NJHealth.org; [f]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD USA, Email: rprevots@niaid.nih.gov.

*These authors contributed equally to this work

Address for correspondence:

Ettie M. Lipner, Ph.D., M.P.H.; Center for Genes, Environment and Health, National Jewish Health, 1400 Jackson Street. Denver, CO 80602. Tel: 303-398-1861, Email: LipnerE@NJHealth.org

FUNDING SOURCES: EML was supported by the Cystic Fibrosis Foundation, Clinical Pilot and Feasibility Award. DRP was supported by the Division of Intramural Research, NIAID. JF was supported by NSF awards 1463642 and 1915277. MS and JLC were supported by NSF award 1743597. JAN was supported by NIH award R01HL146228 and CFF award NICK19G0.

Running title: NTM and molybdenum in persons with cystic fibrosis

Keywords: Nontuberculous Mycobacteria; Molybdenum; Case-Control Study; Cystic Fibrosis; Water Quality; Geospatial

Total word count: 3,290

Abstract word count: 253

ABSTRACT

Rationale: Nontuberculous mycobacteria (NTM) are ubiquitous environmental bacteria that may cause chronic lung disease and are one of the most difficult to treat infections among persons with cystic fibrosis (pwCF). Environmental factors likely contribute to increased NTM densities, with higher potential for exposure and infection.

Objective: To identify water-quality constituents that influence odds of NTM infection among pwCF in Colorado.

Methods: We conducted a population-based nested case-control study using patient data from the Colorado CF Center NTM database. We associated data from pwCF and water-quality data extracted from the Water Quality Portal to estimate odds of NTM infection. Using Bayesian generalized linear models with binomial-distributed discrete responses, we modeled three separate outcomes; any NTM infection, infections due to *Mycobacterium avium* complex species, and infections due to *Mycobacterium abscessus* group species.

Results: We observed a consistent association with molybdenum in the source water and *Mycobacterium abscessus* group species infection among pwCF in all models. For every 1-unit increase in the log concentration of molybdenum in surface water, the odds of infection for those with *Mycobacterium abscessus* group species compared to those who were NTM culture-negative increased by 79%. The odds of *Mycobacterium abscessus* group infection varied by county; the counties with the highest probability of infection are located along the major rivers.

Conclusions: We have identified molybdenum in the source water as the most predictive factor of *Mycobacterium abscessus* group infection among pwCF in Colorado. This finding will help inform patients at risk for NTM of their relative risks in residing within specific regions.

1. INTRODUCTION

Pulmonary nontuberculous mycobacterial (NTM) disease among persons with cystic fibrosis (pwCF) is challenging to treat, requiring prolonged treatment courses (1). Over a recent 5-year interval, nearly 20% of children and adults with CF in the United States who were tested had positive cultures for NTM, of whom 39% had infections with *Mycobacterium abscessus* (2), which is one of the most difficult to treat NTM species (3). Distinct geographic variability of NTM disease has been demonstrated in both general and CF populations (2, 4, 5). Environmental determinants of NTM infection and disease include factors related to moisture in the environment, as well as soil (6) and soil components (4, 7, 8). However, the sources of NTM infection and exposure risks are poorly understood. Environmental conditions related to soil properties, natural water, and engineered water system characteristics, including biofilm formation in premise plumbing, likely contribute to increased NTM densities with higher potential for NTM exposure and infection. Prevention of infections with NTM among pwCF is a critical clinical need (9).

In two previous studies, we explored the role of water exposure in NTM risk. We identified three high-risk watersheds in Colorado (CO) (10), and further used source water data (11) to identify factors potentially influencing the higher risk in these watershed regions. Molybdenum in surface water was a significant contributor to the risk of NTM infection; a 1-unit increase in the log concentration of molybdenum in surface water was associated with a 17% increased risk of NTM infection. Research to date suggests a physiological connection linking molybdenum and essential metabolism of *Mycobacterium tuberculosis*, a phylogenetically related organism to NTM, potentially impacting survival, pathogenesis and persistence (12-14). Given the genetic relatedness of *M. tuberculosis* and NTM, we hypothesize that higher concentrations of specific water-quality constituents, potentially molybdenum, which the bacteria may require for metabolism and growth, result in higher densities of NTM in surface water sources in certain regions. Thus, infection rates would be higher in regions with a water supply from sources with high densities of NTM. In our current study, we hypothesize that specific water-quality constituents in surface water in Colorado influence the odds of having NTM infection among pwCF. To test this hypothesis, we conducted a nested case-control study using water-quality data from the Water Quality Portal, sponsored by the U.S. Geological Survey, U.S. Environmental Protection Agency, and National Water Quality Monitoring Council, together with CF patient data extracted from the Colorado CF Center NTM database.

2. METHODS

2.1 Data Collection

2.1.1 *Study Design and Subjects*

This study was a nested case-control study using demographic and clinical data from the Colorado CF Center NTM database. The Colorado CF Center comprises the Pediatric CF Program at The Children's Hospital Colorado in Aurora, Colorado, and an Adult CF Program at National Jewish Health in Denver, Colorado. The Colorado CF Center is the only CF Center in the state and has nearly complete capture of all CF patients in Colorado. This study therefore can be described as a population-based CF study.

The Colorado CF Center NTM database contained data on pwCF resident in Colorado from January 2007 through January 2019. We extracted patient ZIP code, NTM species, and demographic information. Because we did not have patient address information and our data were too sparse at the ZIP code level, we aggregated all patient ZIP codes to the county-level. Cases were defined as CF patients who had at least one positive NTM culture and were resident in Colorado at the time of their first positive culture, as determined by chart review. We excluded CF patients who had cultured positive only for *M. gordonae* infection. Controls were defined as patients with CF who had at least three negative cultures within a single county over a period of at least three years ("NTM-negative"). Our study population comprised 388 CF patients; 193 cases and 195 controls. This study was approved by the NJH Institutional Review Board (HS-1683).

2.1.2 NTM species

Frequencies of NTM species from patient isolates are listed in Supplementary Table 1. Molecular assays by Line Probe Assay analysis or targeted gene sequencing were used to differentiate *Mycobacterium* species. NTM identification was performed by the Advanced Diagnostics laboratory at NJH, a National Reference Laboratory for NTM.

2.1.3 Water-Quality Data Compilation:

We obtained water-quality data from the Water Quality Portal (WQP) (15), a water quality database collected or hosted by the U.S. Geological Survey, the U.S. Environmental Protection Agency and the National Water Quality Monitoring Council. Our water-quality dataset has been described previously (11). Supplementary Table 2 presents the median and standard deviation values of the water-quality constituents obtained from the WQP that were used in our analyses.

2.2 Statistical Analysis

All water-sample sites were aggregated by county. Subsequently, we calculated the median value of each water-quality constituent for each county. Apparent concentration-unit reporting errors were corrected (for example, three orders of magnitude deviations for individual values were multiplied by 1,000 to align them with the range of the remaining source-specific data). Water-quality constituents were eliminated if data were not available for more than 50 percent of counties. Following these curation steps, seventeen remaining water-quality constituents remained for analysis (Supplementary Table 2). We used a natural log transformation of all county-median variables (17 variables). We standardized all the water-quality constituents' log concentrations to have a mean of 0 and standard deviation of 1. For counties with missing data, we imputed the median value of all water-quality constituents. We also calculated drive time between county centroids and NJH. For patients with any NTM infection, thirty-one counties were dropped from the analysis because there was not at least one case or one control resident in those counties, with thirty-three remaining counties (51.6%) available for analysis. For patients with MAC and *M. abscessus* infection, thirty-one counties (48.4%) and twenty-nine counties (45.3%), respectively, were available for analysis. Each patient was assigned the water quality value for his or her respective county of residence. The counties with available data are shown with non-gray coloring in Figure 1.

2.2.1 Variable Reduction using Principal Component Analysis (PCA)

Principal component analysis (PCA) was used to reduce the number of predictors considered in our subsequent models. PCA is used to determine orthogonal "components" that explain the most variation in the data, where each component is a weighted combination of the predictor variables. For the components explaining the most variation, the variables with the most weight in these components were identified for use in future models. PCA was performed on 17 water-quality constituents summarized at the county-level (after these values were natural log transformed, scaled, and imputed).

Principal components 1 and 2 explained 58.9% of the data variability. Any constituent in the first two components that had a greater contribution than what is expected under equal contribution were identified as important contributors (16, 17). This process is illustrated graphically in Supplementary Figure 1, where the dashed red line represents what is expected under equal contribution. This threshold captured 11 out of 17 constituents: cadmium, calcium, chloride, magnesium, molybdenum, manganese, potassium, selenium, sodium, sulfate, and zinc.

2.2.2 Parameters used in Bayesian Binomial Regression Models

We used Bayesian generalized linear models (GLMs) to model the relationship between NTM infection and demographic and water quality variables. In these models, the dependent variable is NTM infection status, and the predictors are demographic and water quality variables. Diagnostic tools were used to confirm that the fitted models adequately represented the observed pattern of the data. Because age, sex, and race\ethnicity are associated with the risk of NTM infection (2, 18, 19), and could also influence county of residence, we included these as confounders in our model. These relationships are depicted in a Directed Acyclic Graph (DAG) in Supplementary Figure 2.

For each subject, county-level median values of each water-quality constituent (standardized, imputed) were included. In addition, we included a binary variable indicating whether a county's centroid center was within a 1-hour drive to NJH. To control for a higher proportion of patients residing in counties located in the Front Range with greater access to treatment, we categorized counties based on whether their centroid center was within a 1.0-hour drive to NJH. We also performed sensitivity analyses to exclude the drive- time variable from our models (Supplementary Table 3).

2.2.3 Bayesian Binomial Regression Models with Individual Metals from Principal Components 1 & 2

We modeled three separate outcomes (any NTM infection, infections due to *Mycobacterium avium* complex (MAC) species, and infections due to *M. abscessus* group species as a function of water-quality constituents and demographic variables (Supplementary Table 4). Then, for each outcome, we constructed a subsequent model (Model 1) that included only those water-quality constituents whose variance inflation factor was less than 10 to mitigate the potential impact of collinear covariates. For the three models, we sequentially removed the constituent with the highest variance inflation factor. The constituents with variance inflation factors over 10 included magnesium, sodium, potassium, and sulfate, resulting in a final model (Model 1) with the following water-quality constituents: cadmium, calcium, chloride, manganese, molybdenum, selenium, zinc. The correlation matrix for water-quality constituents are shown in Supplementary Table 5. Finally, we constructed separate single-constituent Bayesian GLMs for the water-quality constituents which were significant in Model 1 (as assessed by having a 90% central credible interval (CI) which did not include 1) (Model 2). We estimated the odds of NTM infection among pwCF given exposure to water-quality constituents in surface water sources.

We present an odds ratio and 90% central CI for each model variable. CIs were used to assess the posterior probability of an association between each model variable and a change in the odds of NTM infection. 90% CIs were reported owing to greater computational stability than the 95% CIs in the **rstanarm** package (20).

We predicted the probability that an unobserved CF patient living in a county will have an NTM infection and displayed the results as a probability map across Colorado counties (Figure 1). The software used to perform the analysis are discussed in the Supplementary

Materials. Reproducible source code for the analyses is also provided in the Supplementary Materials.

3. RESULTS

3.1 Study Population Characteristics

Our study population comprised pwCF who received medical care at the Colorado CF Center, and included 195 CF NTM culture-negative patients and 193 pwCF who had at least one positive culture, of whom 147 (76.2%) had MAC infection (*M. avium, M. intracellulare, M. chimaera*) and 82 (42.3%) had *M. abscessus* complex infection (*M. abscessus/chelonae, M. massiliense, M. bolletti*). Forty-six (23.7%) patients had both MAC and *M. abscessus* infections at any time. Patients with both MAC and *M. abscessus* infections were included in both subsets of patients. Demographic characteristics of cases and controls are shown in Table 1. We observed a younger mean age and a higher proportion of males among pwCF with *M. abscessus* infection compared to those with MAC infection. Given well-understood growth rate differences (21), distinct ecological niches (22) and specialized medical treatments (23) for MAC and *M. abscessus* infections, we modeled three separate outcomes: Any NTM infection, infections due to MAC species, and infections due to *M. abscessus* species.

3.2 Bayesian Binomial Regression Models with Individual Metals from Principal Components 1 & 2

Molybdenum was the only constituent significantly associated with increased odds of infection (i.e., 90% CI failed to include 1) (Table 2; Model 1). The results of these models indicate that for every 1-log unit increase in molybdenum concentrations in surface water, the odds of having NTM infection is 1.7, 1.9, and 2.5 times higher for infections caused by all NTM species, MAC species, and *M. abscessus* species, respectively, after controlling for other water-quality constituents.

We then examined the 90% CI for exponentiated parameters of Model 1. The parameters whose 90% CI failed to include 1 were included in separate single-constituent models (Table 3; Model 2). For All NTM species and *M. abscessus* species, the credible intervals for molybdenum were entirely above 1, indicating a significantly higher odds of infection. Even more convincingly, the posterior probability that the molybdenum coefficient is positive (i.e., associated with increased odds of NTM for pwCF) is 96.96% for All NTM species, 94.15% for MAC species, and 99.96% for *M. abscessus* species (Supplementary Table 6). Our results indicate that for every 1-log unit increase in molybdenum concentration in surface water sources at the county-level, the odds of having NTM infection caused by *M. abscessus* species increased by 79% compared with pwCF who were NTM-negative. When modeling all NTM species, we observed a weaker association for molybdenum. We did not observe an association between molybdenum and MAC infections. We also estimated these associations without including drive-time in the models (Supplementary Table 3). The association that we observed between molybdenum and *M. abscessus* infections remained significant (OR = 1.60), although slightly attenuated compared to our main results (OR= 1.79). The association between molybdenum and all NTM infections did not retain significance without including drive-time. Therefore, our results indicate that increasing concentrations of molybdenum in surface water increases the odds of *M. abscessus* infection.

In Figure 1, we calculated the predicted probability that an unobserved pwCF living in a county will have a *M. abscessus* infection based on a model using molybdenum as an independent predictor while controlling for drive time, age race, and gender. The counties with the highest probability of *M. abscessus* infection are located along the major rivers; the South Platte River flowing through Denver, Logan, Sedgwick, and Weld counties, the Colorado River flowing through Mesa county, and the Arkansas River flowing through Pueblo county.

DISCUSSION

We found that molybdenum in surface water sources was associated with increased odds of NTM infection among pwCF, specifically for those with *M. abscessus* group infections. For every 1-log unit increase in molybdenum concentration in surface water among pwCF, the odds of NTM infection caused by *M. abscessus* species increased by 79% compared with those who were NTM-negative (Table 3; Model 2).

As discussed previously (11), molybdenum is involved in the essential metabolism of *Mycobacterium tuberculosis* (12-14), and, given the genetic relatedness of these organisms, it is biologically plausible that it may play a similar role in NTM metabolism (24). In this study, we replicated the molybdenum-NTM infection association in a CF population with water-quality constituent median values calculated for county line boundaries (instead of watershed boundaries (11)). This study also goes a step further to suggest that molybdenum in surface water may increase the odds of acquiring NTM, specifically for *M. abscessus* infection, rather than for MAC infection, in a CF population.

Molybdenum may promote NTM growth in surface water, thereby increasing the risk of exposure and infection. Because we did not have access to environmentally-measured NTM densities, we used infection prevalence as a proxy for NTM abundance, assuming that higher NTM abundance increases the risk of NTM exposure and infection. A recent study (25) demonstrated that NTM abundance from premise plumbing samples as measured by 16S rRNA gene sequencing approach was significantly correlated with higher disease prevalence in population-based epidemiological studies (4). This approach assumes that regions with high disease prevalence correlate with regions of high NTM densities (or more pathogenic species (25)), where certain regional environmental factors create a hospitable environment for NTM to persist. While previous literature has not identified molybdenum in soil or water as a risk factor for NTM, other surveyed metals have been identified as potential risk factors for NTM growth in the environment. In the coastal swamps of the southeastern U.S., high numbers of *Mycobacterium avium*, *Mycobacterium intracellulare*, and *Mycobacterium scrofulaceum* (MAIS) were correlated with high zinc concentrations in water samples (26). Although we did not observe an association between zinc concentrations in surface water and MAC infections, different NTM species may require specific environmental conditions for growth in different habitats, and thus discrepant findings are not unexpected. By analyzing water-quality data across diverse geographic regions, we hope to identify factors that are necessary in promoting NTM growth in water sources, as well as identifying whether these factors differ for MAC and *M. abscessus* species.

Figure 1 presents the predicted probability of *M. abscessus* infection by county. The more highly populated of these counties with the highest probability of *M. abscessus* infection, Denver, Mesa, Pueblo, and Weld, have public water supplies with centralized water distribution systems that come almost entirely from surface water sourced primarily from these rivers (27). Among the rural counties with high probabilities of infection located along the South Platte River, the public water supply for Logan county is primarily from surface water, while Sedgwick county relies heavily on groundwater (27). Many of the counties located along the major Colorado rivers also use water from these rivers for crop irrigation (27). These county-level probabilities of infection suggest that potential sources of NTM exposure may come through municipal water systems that take water from these rivers as well as possibly from crop irrigation. The results shown in this map reflect the same high-risk regions that we have reported previously (11).

This study reports an important finding for the CF population. MAC and *M. abscessus* are the two most clinically relevant NTM species, which comprise 95% of NTM infections among pwCF (2, 3, 9, 28, 29). Adjemian *et al*. observed significant increases for *M. abscessus* between 2010-2014 in the Mountain states among pwCF (2). Rendering a framework of the necessary environmental factors that predict NTM exposure and infection is crucial for the development of prevention strategies.

STRENGTHS AND LIMITATIONS

In our previous studies (10, 11), we did not have sufficient data to identify and exclude individuals who had moved to Colorado after their initial infection diagnosis. The data used in this study ensured that a subject's first positive culture occurred in Colorado, which prevented selection bias from influencing our results.

Only a subset of the water-quality constituent dataset for the state of Colorado was used in our analysis due to constraints in our study design. Counties were dropped from the analysis if no pwCF resided there. As a result, our findings were based on approximately half of Colorado's counties. While our patient population included nearly all pwCF in Colorado, our results may therefore be generalizable to all pwCF in Colorado but only to the counties included in the analysis. In addition, some limitations are inherent to our water-quality constituent dataset (11). Water sampling locations were not from random or systematically representative locations and the number of sites sampled across counties were variable. Additionally, data were imputed to some counties with missing information. Therefore, we do not know the degree of bias in the resulting median concentration values for each county. If exposure misclassification with respect to water-quality constituents were present, we would assume it to be nondifferential with respect to cases and controls. This type of misclassification would bias the odds ratio toward the null. Finally, since source water samples were used in these analyses rather than tap water, our findings may not be representative of the water that people are exposed to in their homes after filtration and treatment.

CONCLUSIONS

This study has identified molybdenum in surface water as the most predictive environmental factor of NTM infection among pwCF in Colorado, specifically for *M. abscessus* infection. We are too early in this discovery process to make specific recommendations, although if future studies confirm that molybdenum is in fact a necessary or sufficient factor for growth of *M. abscessus* species in water sources, these findings could inform patients at risk for NTM of their relative risks in residing within specific regions. Analyzing water-quality data across diverse geographic regions may render a framework of factors that are necessary for NTM growth, specifically factors that may differ for MAC and *M. abscessus* species. Investigating whether molybdenum metabolism in the (human) host affects NTM susceptibility will also have important implications for at-risk populations.

ACKNOWLEDGEMENTS: The authors thank Dr. Katherine Walton-Day and Dr. Carleton R. Bern for their expertise and consultation on the water-quality constituents in our dataset. The authors also thank Christine Barboa for assistance with chart review.

CONFLICT OF INTEREST

All authors report no conflict of interest.

References

1. Floto RA, Olivier KN, Saiman L, Daley CL, Herrmann JL, Nick JA, et al. US Cystic Fibrosis Foundation and European Cystic Fibrosis Society consensus recommendations for the management of non-tuberculous mycobacteria in individuals with cystic fibrosis. Thorax. 2016;71 Suppl 1:i1-22.
2. Adjemian J, Olivier KN, Prevots DR. Epidemiology of Pulmonary Nontuberculous Mycobacterial Sputum Positivity in Patients with Cystic Fibrosis in the United States, 2010-2014. Ann Am Thorac Soc. 2018;15(7):817-26.
3. Degiacomi G, Sammartino JC, Chiarelli LR, Riabova O, Makarov V, Pasca MR. Mycobacterium abscessus, an Emerging and Worrisome Pathogen among Cystic Fibrosis Patients. Int J Mol Sci. 2019;20(23).
4. Adjemian J, Olivier KN, Seitz AE, Falkinham JO, 3rd, Holland SM, Prevots DR. Spatial clusters of nontuberculous mycobacterial lung disease in the United States. Am J Respir Crit Care Med. 2012;186(6):553-8.
5. Spaulding AB, Lai YL, Zelazny AM, Olivier KN, Kadri SS, Prevots DR, et al. Geographic Distribution of Nontuberculous Mycobacterial Species Identified among Clinical Isolates in the United States, 2009-2013. Ann Am Thorac Soc. 2017;14(11):1655-61.
6. Reed C, von Reyn CF, Chamblee S, Ellerbrock TV, Johnson JW, Marsh BJ, et al. Environmental risk factors for infection with Mycobacterium avium complex. Am J Epidemiol. 2006;164(1):32-40.
7. Prevots DR, Marras TK. Epidemiology of human pulmonary infection with nontuberculous mycobacteria: a review. Clin Chest Med. 2015;36(1):13-34.
8. Adjemian J, Olivier KN, Prevots DR. Nontuberculous mycobacteria among patients with cystic fibrosis in the United States: screening practices and environmental risk. Am J Respir Crit Care Med. 2014;190(5):581-6.
9. Gross JE, Martiniano SL, Nick JA. Prevention of transmission of Mycobacterium abscessus among patients with cystic fibrosis. Curr Opin Pulm Med. 2019;25(6):646-53.
10. Lipner EM, Knox D, French J, Rudman J, Strong M, Crooks JL. A Geospatial Epidemiologic Analysis of Nontuberculous Mycobacterial Infection: An Ecological Study in Colorado. Annals of the American Thoracic Society. 2017.
11. Lipner EM, French J, Bern CR, Walton-Day K, Knox D, Strong M, et al. Nontuberculous Mycobacterial Disease and Molybdenum in Colorado Watersheds. Int J Environ Res Public Health. 2020;17(11).
12. Levillain F, Poquet Y, Mallet L, Mazeres S, Marceau M, Brosch R, et al. Horizontal acquisition of a hypoxia-responsive molybdenum cofactor biosynthesis pathway contributed to Mycobacterium tuberculosis pathoadaptation. PLoS Pathog. 2017;13(11):e1006752.
13. Williams MJ, Kana BD, Mizrahi V. Functional analysis of molybdopterin biosynthesis in mycobacteria identifies a fused molybdopterin synthase in Mycobacterium tuberculosis. J Bacteriol. 2011;193(1):98-106.
14. McGuire AM, Weiner B, Park ST, Wapinski I, Raman S, Dolganov G, et al. Comparative analysis of Mycobacterium and related Actinomycetes yields insight into the evolution of Mycobacterium tuberculosis pathogenesis. BMC Genomics. 2012;13:120.
15. US Geological Survey UDoA, National Water Quality Monitoring Council. Water Quality Portal. 2012.

16. A. K. Practical Guide To Principal Component Methods in R: PCA, M(CA), FAMD, MFA, HCPC, factoextra: STHDA (http://www.sthda.com); 2017.

17. analysis SStfh-tpd. Articles - Principal Component Methods in R: Practical Guide. CA - Correspondence Analysis in R: Essentials [Available from: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/113-ca-correspondence-analysis-in-r-essentials/.

18. Adjemian J, Olivier KN, Seitz AE, Holland SM, Prevots DR. Prevalence of nontuberculous mycobacterial lung disease in U.S. Medicare beneficiaries. Am J Respir Crit Care Med. 2012;185(8):881-6.

19. Olivier KN, Weber DJ, Wallace RJ, Jr., Faiz AR, Lee JH, Zhang Y, et al. Nontuberculous mycobacteria. I: multicenter prevalence study in cystic fibrosis. Am J Respir Crit Care Med. 2003;167(6):828-34.

20. Goodrich B GJ, Ali I, Brilleman S. rstanarm: Bayesian applied regression modeling via Stan. R package version 2.19.3. 2020.

21. Runyon EH. Typical Myobacteria: Their Classification. Am Rev Respir Dis. 1965;91:288-9.

22. Honda JR, Virdi R, Chan ED. Global Environmental Nontuberculous Mycobacteria and Their Contemporaneous Man-Made and Natural Niches. Front Microbiol. 2018;9:2029.

23. Henkle E, Winthrop KL. Nontuberculous mycobacteria infections in immunosuppressed hosts. Clin Chest Med. 2015;36(1):91-9.

24. Falkinham JO, 3rd. Surrounded by mycobacteria: nontuberculous mycobacteria in the human environment. J Appl Microbiol. 2009;107(2):356-67.

25. Gebert MJ, Delgado-Baquerizo M, Oliverio AM, Webster TM, Nichols LM, Honda JR, et al. Ecological Analyses of Mycobacteria in Showerhead Biofilms and Their Relevance to Human Health. mBio. 2018;9(5).

26. Kirschner RA, Jr., Parker BC, Falkinham JO, 3rd. Epidemiology of infection by nontuberculous mycobacteria. Mycobacterium avium, Mycobacterium intracellulare, and Mycobacterium scrofulaceum in acid, brown-water swamps of the southeastern United States and their association with environmental variables. Am Rev Respir Dis. 1992;145(2 Pt 1):271-5.

27. (www.cfwe.org) CFFWE. Citizen’s Guide to Where Your Water Comes From. 2005.

28. Viviani L, Harrison MJ, Zolin A, Haworth CS, Floto RA. Epidemiology of nontuberculous mycobacteria (NTM) amongst individuals with cystic fibrosis (CF). J Cyst Fibros. 2016;15(5):619-23.

29. Martiniano SL, Nick JA, Daley CL. Nontuberculous Mycobacterial Infections in Cystic Fibrosis. Thorac Surg Clin. 2019;29(1):95-108.

Figure Legend.

Figure 1. Predicted probability of *M. abscessus* infection for counties where pwCF resided. Gray lines represent county line boundaries in Colorado. County names are printed in *black*. *Blue areas* indicate lakes, reservoirs, and rivers.

Table 1. Descriptive statistics of cases (NTM culture positive) and controls among a Colorado CF patient population.

Characteristic	Controls
	(CF only)
	n = 195
		Patient infection
		from
		all NTM species
		n = 193
			Patient infection from
			MAC species
			n = 147
				Patient infection from
				MABSC species
				n = 82

Age, yr, mean±SD	35.66±11.90	37.30±13.37	37.66±13.93	35.20±10.80
Female sex, n (%)	95 (48.7)	109 (55.9)	86 (58.5)	36 (43.9)
White race, n (%)	187 (95.9)	187 (96.9)	143 (97.3)	80 (97.6)

Table 2. Model 1. Bayesian binomial regression model examining water-quality constituents (with VIF values less than 10 from Model 1) associated with odds of NTM infection among pwCF. Bolded estimates have 90% Cis that fail to include 1. CI = Credible Interval.

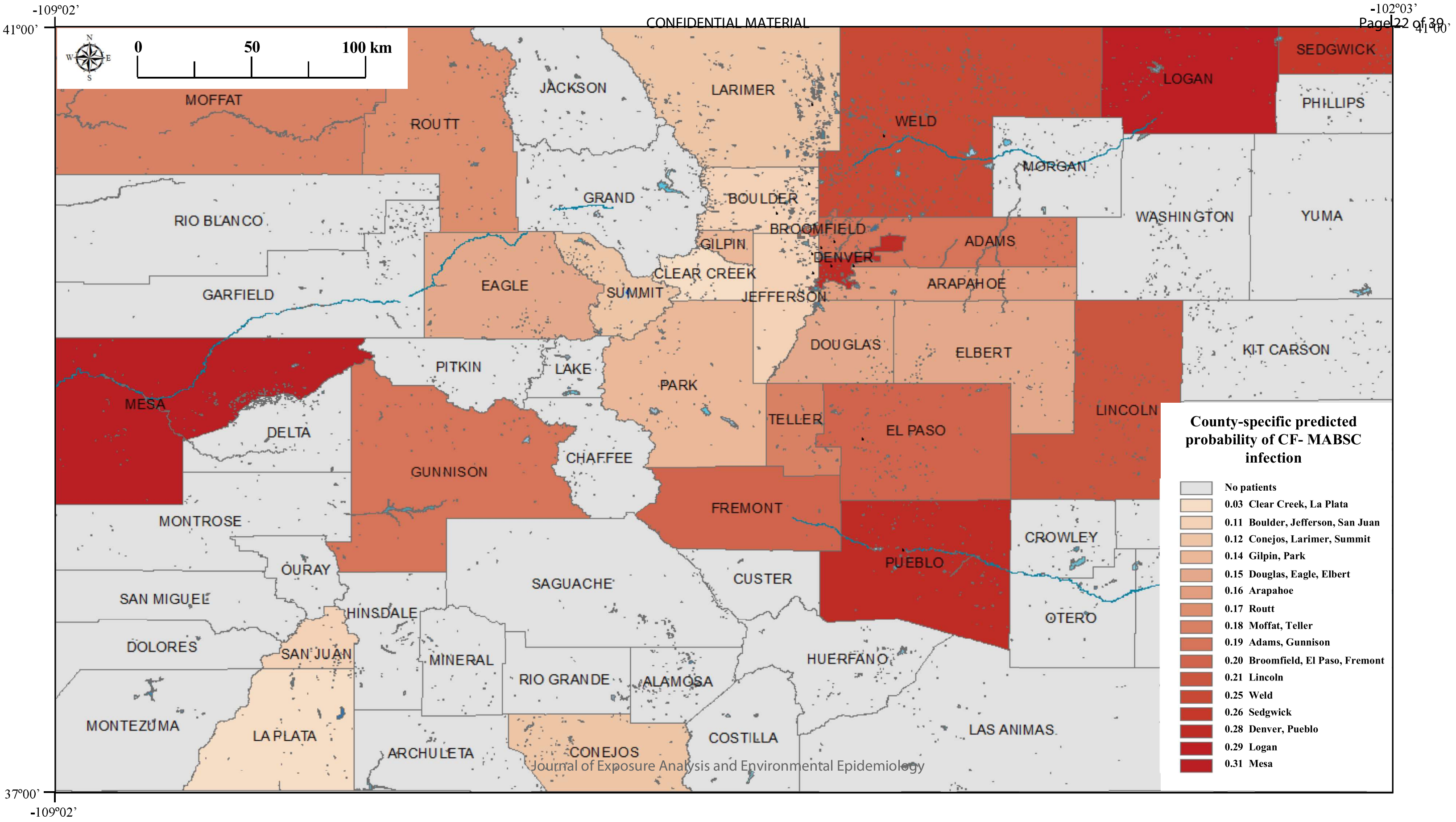
All NTM species		MAC species		MABSC species	
Variable	Odds Ratio (90% CI)	Variable	Odds Ratio (90% CI)	Variable	Odds Ratio (90% CI)
Age: (1 Year)	1.01 (1.00, 1.03)	Age: (1 Year)	1.01 (1.00, 1.03)	Age: (1 Year)	1.00 (0.98, 1.03)
Gender: Male	0.77 (0.54, 1.11)	Gender: Male	0.70 (0.48, 1.03)	Gender: Male	1.33 (0.84, 2.10)
Race: Non-White*	0.75 (0.29, 1.93)	Race: Non-White*	0.72 (0.21, 1.93)	Race: Non-White*	0.40 (0.07, 1.43)
Drive-time (>1.0 hours to NJH)	1.52 (0.90, 2.56)	Drive-time (>1.0 hours to NJH)	1.67 (0.96, 2.92)	Drive-time (>1.0 hours to NJH)	1.93 (0.93, 4.10)
Cadmium (1-log unit)	1.15 (0.84, 1.57)	Cadmium (1-log unit)	1.20 (0.86, 1.68)	Cadmium (1-log unit)	1.22 (0.80, 1.88)
Calcium (1-log unit)	0.89 (0.54, 1.45)	Calcium (1-log unit)	0.80 (0.47, 1.36)	Calcium (1-log unit)	0.54 (0.25, 1.13)
Chloride (1-log unit)	1.05 (0.73, 1.51)	Chloride (1-log unit)	1.06 (0.72, 1.60)	Chloride (1-log unit)	1.10 (0.67, 1.84)
Manganese (1-log unit)	0.88 (0.57, 1.30)	Manganese (1-log unit)	0.98 (0.64, 1.52)	Manganese (1-log unit)	0.84 (0.45, 1.49)
Molybdenum (1-log unit)	**1.69 (1.04, 2.80)**	Molybdenum (1-log unit)	**1.87 (1.09, 3.25)**	Molybdenum (1-log unit)	**2.47 (1.28, 4.90)**
Selenium (1-log unit)	0.85 (0.54, 1.32)	Selenium (1-log unit)	0.81 (0.51, 1.28)	Selenium (1-log unit)	1.16 (0.61, 2.25)
Zinc (1-log unit)	1.37 (0.82, 2.32)	Zinc (1-log unit)	1.00 (0.57, 1.75)	Zinc (1-log unit)	2.14 (0.99, 5.42)

*Reference group is White Alone

Table 3. Model 2. Single-exposure Bayesian binomial regression model examining significant metals from Model 1 associated with odds of NTM infection among pwCF. Bolded estimates have 90% CIs that fail to include 1. CI = Credible Interval.

All NTM species		MAC species		MABSC species	
Variable	Odds Ratio (90% CI)	Variable	Odds Ratio (90% CI)	Variable	Odds Ratio (90% CI)
Age: (1 Year)	1.01 (1.00, 1.02)	Age: (1 Year)	1.01 (1.00, 1.02)	Age: (1 Year)	1.01 (0.98, 1.02)
Gender: Male	0.76 (0.54, 1.06)	Gender: Male	**0.68 (0.47, 0.99)**	Gender: Male	1.21 (0.77, 1.92)
Race: Non-White*	0.76 (0.29, 1.95)	Race: Non-White*	0.67 (0.21, 1.88)	Race: Non-White*	0.76 (0.09, 1.49)
Drive-time (>1.0 hours to NJH)	1.28 (0.88, 1.92)	Drive-time (>1.0 hours to NJH)	1.32 (0.89, 1.99)	Drive-time (>1.0 hours to NJH)	**1.28 (1.05, 2.75)**
Molybdenum (1-log unit)	**1.29 (1.03, 1.62)**	Molybdenum (1-log unit)	1.26 (0.99, 1.61)	Molybdenum (1-log unit)	**1.79 (1.34, 2.44)**

*Reference group is White Alone


-109º02’
41º00’
-102º03’
41º00’
37º00’
-109º02’
N
W
E
S
0
50
100 km
MOFFAT
ROUTT
JACKSON
LARIMER
WELD
LOGAN
SEDGWICK
PHILLIPS
MORGAN
RIO BLANCO
GRAND
BOULDER
BROOMFIELD
ADAMS
WASHINGTON
YUMA
GILPIN
DENVER
EAGLE
CLEAR CREEK
SUMMIT
JEFFERSON
ARAPAHOE
GARFIELD
DOUGLAS
ELBERT
KIT CARSON
MESA
PITKIN
LAKE
PARK
DELTA
TELLER
EL PASO
LINCOLN
GUNNISON
CHAFFEE
FREMONT
MONTROSE
CROWLEY
OURAY
PUEBLO
SAGUACHE
CUSTER
SAN MIGUEL
HINSDALE
OTERO
DOLORES
SAN JUAN
MINERAL
HUERFANO
RIO GRANDE
ALAMOSA
MONTEZUMA
LA PLATA
COSTILLA
LAS ANIMAS
ARCHULETA
CONEJOS
County-specific predicted probability of CF- MABSC infection
No patients
0.03 Clear Creek, La Plata
0.11 Boulder, Jefferson, San Juan
0.12 Conejos, Larimer, Summit
0.14 Gilpin, Park
0.15 Douglas, Eagle, Elbert
0.16 Arapahoe
0.17 Routt
0.18 Moffat, Teller
0.19 Adams, Gunnison
0.20 Broomfield, El Paso, Fremont
0.21 Lincoln
0.25 Weld
0.26 Sedgwick
0.28 Denver, Pueblo
0.29 Logan
0.31 Mesa

Supplementary Table 1. Frequencies of NTM group species from patient isolates

Species groups diagnosed from patient isolates.	**Culture-positive CF patients n=193**
M. abscessus, M. bolletii, M. chelonae	2
M. abscessus, M. chelonae	32
M. abscessus, M. chelonae, M. chimaera	1
M. abscessus, M. chimaera, M. chelonae, M. xenopi	1
M. abscessus, M. massiliense, M. chelonae	2
M. avium_complex	65
M. abscessus, M. avium_complex, M. chelonae	30
M. abscessus, M. avium_complex, M. chelonae, M. chimaera, M. massiliense	1
M. abscessus, M. avium_complex, M. chelonae, M. fortuitum	1
M. abscessus, M. avium_complex, M. chelonae, M. massiliense	4
M. avium_complex, M. chimaera	4
M. avium_complex, M. gordonae	5
M. avium_complex, M. gordonae, M. intracellulare	1
M. avium_complex, M. intracellulare	12
M. abscessus, M. avium_complex, M. chelonae, M. intracellulare	3
M. abscessus, M. avium_complex, M. chelonae, M. intracellulare, M. yongonense	1
M. avium_complex, M. intracellulare, M. chimaera	1
M. avium_complex, M. intracellulare, M. yongonense	2
M. avium_complex, M. lentiflavum	2
M. avium_complex, M. simiae	1
M. avium_complex, M. thermoresistible	1
M. avium_complex, M. chimaera, M. yongonense	3
M. abscessus, M. avium_complex, M. chelonae, M. chimaera, M. gordonae, M. lentiflavum	1
M. chimaera	2
M. abscessus, M. chelonae, M. gordonae	1
M. intracellulare	3
M. intracellulare, M. yongonense	1
M. kansasii	4
M. lentiflavum	2
M. abscessus, M. chelonae, M. lentiflavum	1
M. mucogenicum	1
M. abscessus, M. avium_complex, M. chelonae, M. fortuitum, M. massiliense, M. simiae, M. szulgai	1
M. chimaera, M. yongonense	1

Supplementary Table 2. Median and standard deviation (SD) values of water-quality constituents[*] obtained from the Water Quality Portal (WQP) used in PCA.

Exposure Characteristics	**Median ± SD** (µg/L)
Aluminum	18 ± 4371.6
Arsenic	<0.5 ± 49.9
Cadmium	0.1 ± 50.6
Calcium	32110 ± 70745.7
Chloride	2230 ± 219285.6
Copper	1.6 ± 440.8
Iron	38 ± 26245.6
Lead	<0.5 ± 326.4
Magnesium	6691 ± 40822.9
Manganese	22.6 ± 7406.7
Molybdenum	4.3 ± 18.8
Nickel	1.2 ± 37.2
Potassium	1347 ± 6884.6
Selenium	0.06 ± 48.0
Sodium	6100 ± 123203.3
Sulfate	19000 ± 598707.4
Zinc	17 ± 5951.9

[*] The filtered portion (means the water was passed through a 0.45 micrometer filter) of the water-sample fractions were used.

Supplementary Table 3. Sensitivity analyses. Single-exposure Bayesian binomial regression model examining significant metals from Model 1 associated with odds of NTM infection among pwCF, excluding drive time. Bolded estimates have 90% CIs that fail to include 1. CI = Credible Interval.

All NTM species		MAC species		MABSC species	
Variable	Odds Ratio (95% CI)	Variable	Odds Ratio (95% CI)	Variable	Odds Ratio (95% CI)
Age: (1 Year)	1.01 (1.00, 1.02)	Age: (1 Year)	1.01 (1.00, 1.03)	Age: (1 Year)	1.00 (0.98, 1.02)
Gender: Male	0.77 (0.54, 1.07)	Gender: Male	0.70 (0.48, 1.00)	Gender: Male	1.26 (0.81, 1.97)
Race: Non-White*	0.78 (0.30, 1.97)	Race: Non-White*	0.69 (0.22, 1.92)	Race: Non-White*	1.26 (0.09, 1.51)
Molybdenum (1-log unit)	1.22 (0.99, 1.52)	Molybdenum (1-log unit)	1.19 (0.95, 1.48)	Molybdenum (1-log unit)	**1.60 (1.22, 2.12)**

*Reference group is White Alone

Supplementary Table 4. Bayesian binomial regression model examining the 11-contributing water-quality constituents from principal components 1 and 2 and other covariates associated with odds of NTM infection among pwCF. Bolded estimates have 90% CIs that fail to include 1. CI = Credible Interval.

All NTM species		MAC species		MABSC species	
Variable	Odds Ratio (95% CI)	Variable	Odds Ratio (95% CI)	Variable	Odds Ratio (95% CI)
Age: (1 Year)	1.01 (1.00, 1.03)	Age: (1 Year)	1.01 (1.00, 1.03)	Age: (1 Year)	1.00 (0.98, 1.02)
Gender: Male	0.75 (0.53, 1.06)	Gender: Male	0.69 (0.47, 1.01)	Gender: Male	1.27 (0.79, 2.05)
Race: Non-White[a]	0.82 (0.30, 2.12)	Race: Non-White[a]	0.75 (0.23, 2.16)	Race: Non-White[a]	0.38 (0.08, 1.46)
Drive-time (>1.0 hours to NJH)	1.37 (0.69, 2.64)	Drive-time (>1.0 hours to NJH)	1.57 (0.77, 3.06)	Drive-time (>1.0 hours to NJH)	1.26 (0.48, 3.10)
Calcium (1-log unit)	**0.15 (0.02, 0.99)**	Calcium (1-log unit)	0.22 (0.02, 1.57)	Calcium (1-log unit)	**0.02 (0.001, 0.22)**
Cadmium (1-log unit)	0.88 (0.52, 1.46)	Cadmium (1-log unit)	0.91 (0.53, 1.54)	Cadmium (1-log unit)	0.93 (0.42, 2.12)
Chloride (1-log unit)	**0.42 (0.20, 0.81)**	Chloride (1-log unit)	**0.43 (0.20, 0.90)**	Chloride (1-log unit)	0.41 (0.16, 1.02)
Magnesium (1-log unit)	1.91 (0.21, 19.5)	Magnesium (1-log unit)	1.07 (0.10, 12.4)	Magnesium (1-log unit)	11.48 (0.65, 247.2)
Manganese (1-log unit)	0.68 (0.35, 1.27)	Manganese (1-log unit)	0.64 (0.31, 1.23)	Manganese (1-log unit)	0.97 (0.41, 2.36)
Molybdenum (1-log unit)	**2.89 (1.32, 6.89)**	Molybdenum (1-log unit)	**2.54 (1.08, 6.49)**	Molybdenum (1-log unit)	**7.11 (2.16, 25.5)**
Potassium (1-log unit)	**5.56 (1.20, 30.9)**	Potassium (1-log unit)	**5.75 (1.23, 33.1)**	Potassium (1-log unit)	7.69 (0.92, 76.7)

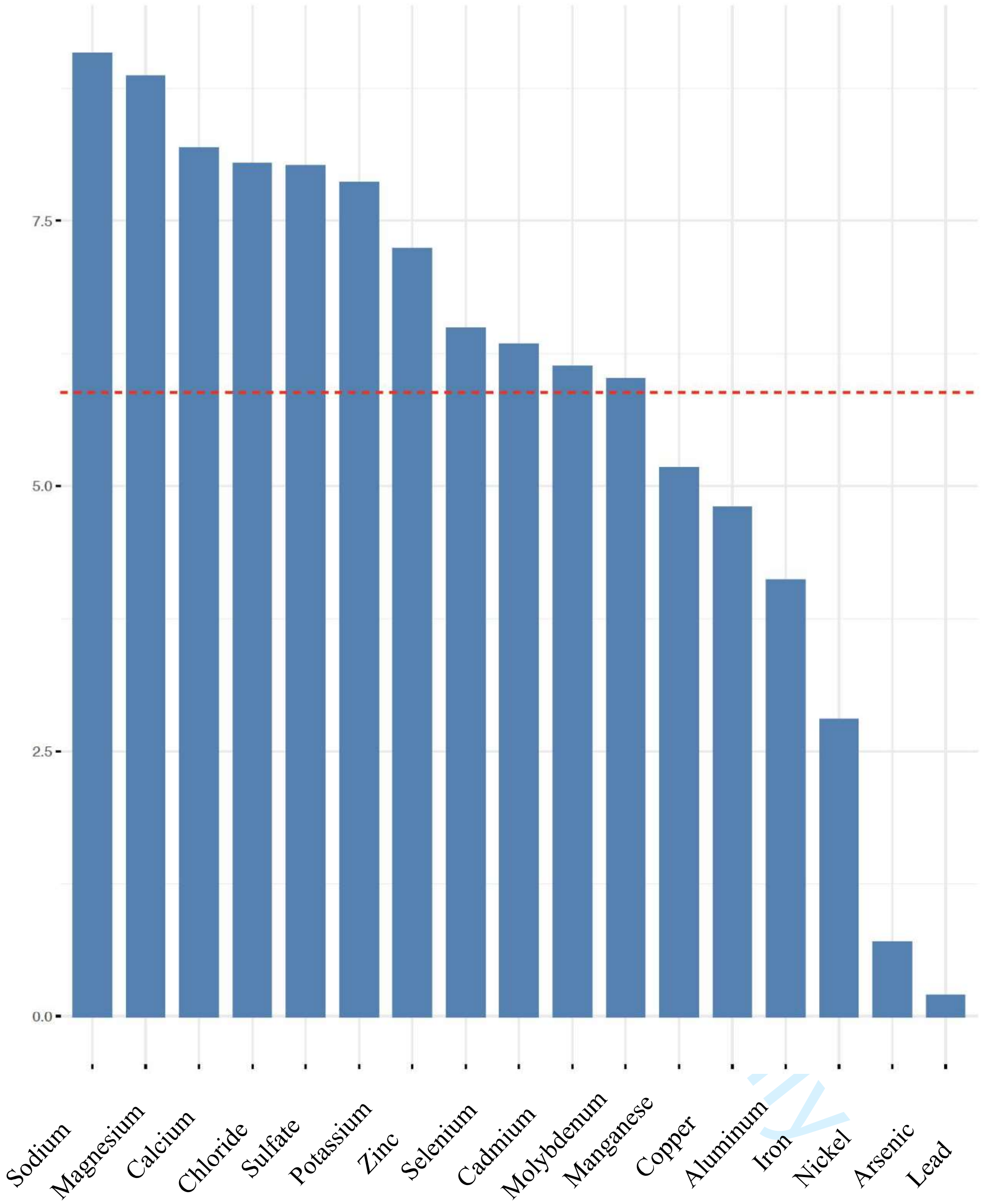
Selenium (1-log unit)	0.60 (0.33, 1.07)	Selenium (1-log unit)	**0.55** **(0.30, 0.98)**	Selenium (1-log unit)	0.96 (0.40, 2.32)
Sodium (1-log unit)	0.24 (0.03, 1.51)	Sodium (1-log unit)	0.43 (0.05, 3.16)	Sodium (1-log unit)	**0.04** **(0.002, 0.60)**
Sulfate (1-log unit)	**8.23** **(2.27, 37.3)**	Sulfate (1-log unit)	**6.17** **(1.54, 29.4)**	Sulfate (1-log unit)	**20.0** **(2.89 170.7)**
Zinc (1-log unit)	1.59 (0.79, 3.29)	Zinc (1-log unit)	1.36 (0.68, 2.83)	Zinc (1-log unit)	1.56 (0.54, 5.26)

Supplementary Table 5. Correlation matrix (Pearson's Correlation Coefficient, ρ) for the water-quality constituents contributing to Principal Components 1 & 2.

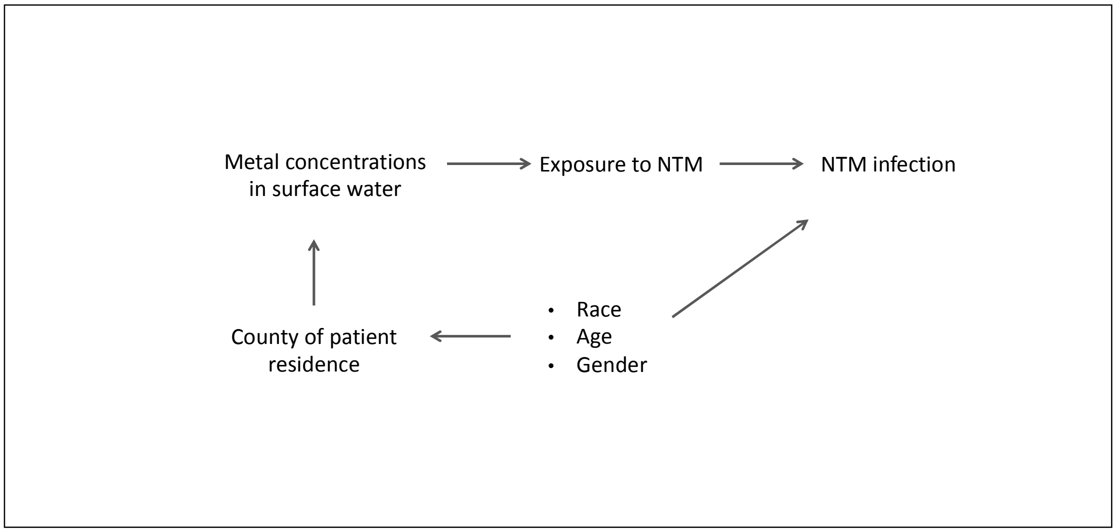
	Cd	Ca	Cl	Mg	Mn	Mo	K	Se	Na	SO_4^{2-}	Zn
Cadmium (Cd)	1.00										
Calcium (Ca)	0.22	1.00									
Chloride (Cl)	0.13	0.78	1.00								
Magnesium (Mg)	0.31	0.96	0.77	1.00							
Manganese (Mn)	0.50	0.59	0.46	0.53	1.00						
Molybdenum (Mo)	-0.06	0.70	0.69	0.69	0.38	1.00					
Potassium (K)	0.28	0.88	0.80	0.88	0.65	0.76	1.00				
Selenium (Se)	-0.17	0.72	0.68	0.70	0.32	0.78	0.75	1.00			
Sodium (Na)	0.25	0.88	0.82	0.91	0.59	0.82	0.95	0.76	1.00		
Sulfate (SO_4^{2-})	0.22	0.90	0.88	0.90	0.52	0.64	0.82	0.67	0.87	1.00	
Zinc (Zn)	0.58	0.27	0.26	0.26	0.60	0.14	0.25	0.07	0.26	0.28	1.00

Supplementary Table 6. Summary of posterior probability that the Molybdenum covariate is associated with increased odds of NTM infection among pwCF.

All NTM species	MAC species	MABSC species
96.96%	94.15%	99.96%



Supplementary Figure 1. Contribution of water-quality constituents to principal components 1 and 2.


Metal concentrations in surface water
Exposure to NTM
NTM infection
County of patient residence
Race
Age
Gender

Supplementary Figure 2. Directed Acyclic Graph (DAG) depicting the relationship between confounders, exposure and dependent variables.

Supplementary Methods

2.2 Statistical Analysis

Analysis of data was performed using the R packages: **rgdal** (1), **sp** (2), **rstanarm** (3), **dplyr** (4), **standardize** (5), **missMDA** (6),**gmapdistance** (7), **FactoMineR** (8), and **factoextra** (9). All water-sample sites were aggregated by county using the **sp** package. We calculated the median value of each water-quality constituent for each county using the **dplyr** package. The R source code that we created to calculate the county medians is available in the Supplementary Materials. Using the scale function from the **standardize** package, we standardized all the water-quality constituents' log concentrations to have a mean of 0 and standard deviation of 1. Using the scale function from the **standardize** package, we standardized all the water-quality constituents' log concentrations to have a mean of 0 and standard deviation of 1. For counties with missing data, we imputed the median value of all water-quality constituents using the imputePCA function in the **missMDA** package. Drive time between county centroids and NJH were calculated using the R **gmapsdistance** package.

2.2.1 <u>Variable Reduction using Principal Component Analysis (PCA)</u>

PCA was performed using the PCA function in the **FactoMineR** package on 17 water-quality constituents summarized at the county-level (after these values were natural log transformed, scaled, and imputed). We used the fviz_contrib function in the **factoextra** package to identify the most important variables in explaining variability of principal components 1 and 2.

<u>2.2.3 Parameters used in Bayesian Binomial Regression Models</u>

We used Bayesian generalized linear models (GLM) to model the response as Binomial, which links the logit of the probability of NTM occurrence to a weighted linear combination of the predictors via the **rstanarm** package (3).

For the prior distribution of the intercept, we used a Student's t distribution with a 1 degree of freedom, a location parameter of 0 and a scale parameter of 2.5 For the prior distributions of the remaining regression coefficients, we used independent and identically-distributed normal distributions with a mean of 0 and a standard deviation of 5. Our models assumed overdispersed, binomial-distributed discrete responses and used the logit link function; the posterior distributions were approximated using 10,000 Markov chain Monte Carlo (MCMC) iterations, which includes a default warmup period of 5,000 iterations.

2.2.4 Bayesian Binomial Regression Models with Individual Metals from Principal Components 1 & 2.

The posterior probabilities shown in Table 4 (Model 3) were calculated using the **rstanarm** and **rstan** packages (3, 10). We used the posterior_linpred function from the **rstanarm** R package to predict the probability that an unobserved CF patient living in a county will have an NTM infection and displayed the results as a probability map across Colorado counties (Figure 1).

References:

1. Bivand R, Keitt T, Rowlingson B. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.5-12. 2020. https://CRAN.R-project.org/package=rgdal.
2. Roger S, Bivand EP, Virgilio Gomez-Rubio. Applied spatial data analysis with R. Springer, NY2013.
3. Goodrich B Gabry J, Ali I, Brilleman S. rstanarm: Bayesian applied regression modeling via Stan. R package version 2.19.3. 2020.
4. Wickham H, Francois R, Henry L, Muller K. dplyr: A Grammar of Data Manipulation. R package version 083 2019.
5. Eager CD. standardize: Tools for Standardizing Variables fo Regression in R. R package version 0.2.1. 2017.
6. Josse J, Husson, F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. Journal of Statistical Software. 2016;70(1):1-31.
7. Azuero Melo R, Rodriguez D, Zarruk D. gmapsdistance: Distance and Travel Time Between Two Points from Google Maps. R package version 3.4. 2018.
8. Le S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. 2008;25(1):1-18.
9. Kassambara A, Mundt F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. 2020.
10. Team SD. RStan: the R interface to Stan. R package version 2.19.3. 2020.

Documentation from raw data download to cleaned water chemistry dataset

A. The key words entered into the Water Quality Portal (WQP):

For metals and nonmetals:

Country: “US”
State: “US:CO”
Site Type: “Aggregate surface-water-use”, “Lake, Reservoir, Impoundment”, “Stream”, “Wetland”.
Sample Media: “Water (NWIS, STORET)”
Characteristic Group: “Inorganics, Major, metals (NWIS, STORET)”, “Inorganics, Minor, metals (NWIS, STORET)”, “Inorganics, Major, Non-metals (NWIS, STORET)”, “Inorganics, Minor, Non-metals (NWIS, STORET)”
Date range – from: “01-01-2000” to: “12-31-2018”

B. We download the files “Sample results (narrow)” (downloaded as “narrowresult“) and “Site data only” (downloaded as “station”).
The spreadsheets were merged based on the “MonitoringLocationIdentifier”.

C. Cleaning procedures
The following steps were performed on the 4 datasets -- major metals, minor metals, major nonmetals, minor nonmetals, pH and total coliform:

1. All measurement values with less than sign (“ < ”) listed in the “ResultMeasureValue” column were eliminated from the dataset.
For the Monitoring Location Identifier "LEWWTP-BEAR CR", the longitude was entered as “-104.03298”, this was changed to the correct longitude, “-105.03298”.
2. We removed any entries with following unit codes: “%”, “lb/day”, “ueq/L”
3. We used only filtered (Dissolved) sample fractions. We excluded any sample fractions labeled as “Fixed”, “Suspended”, “Bed Sediment”, “Comb Available”, “Unfiltered”, “Acid Soluble”, “Recoverable”, “Total Recovrble”, “Total”, “Pot. Dissolved”, or missing.
4. We removed the following entries with the following Monitoring Location Identifiers from the dataset: “0801478-EME”, “0801478-EMET SHAFT”, “0801478-EMET-SP”, “0801478-EM-1”, “0801478-MAR-01”, “0801478-MARION”, “0801478-OG1TMW3”, “0801478-PRP-01”, “0801478-PRP-01 MS”, “0801478-PRP-01 MSD”, “0801478-YT”, “"0801478-YT-1", “0801478-YTBH”, “0801478-YTPD”, “0801478-SDDS”, “0801478-SDDS-1”, “0801478-SDDS-2”, “0801478-SD-1A”, “0801478-SD1A40”, “0801478-MRP-01”, “0801478-SHG-EMSP”, “USGS-410039105374401”, “USGS -40480010546000”, “USGS-390500106323000”, “USGS-372900106470000”, “UTEMTN-HAYFIELD_RESVR”, “UTEMTN-4000 BLOCK POND”, “UTEMTN-MBLWWELL 1”.
These locations are either in mine shafts, are snow collections sites, or groundwater mistakenly labeled as surface water.
5. We deleted the measurement taken at Monitoring Location Identifier “0800257-CC-26” on date “2000-08-17” because the values for all metals and nonmetals were suspiciously high.

Specific changes made to each dataset.

For the major metals dataset:
1. Units labeled as “mg/l CaCO3” were relabeled as “mg/l”
2. Calcium: Measures > “1000” and labeled “mg/l”, were relabeled as “ug/l” (18 measurements)
3. Sodium: The measure = “0.5” labeled “ug/l”, was relabeled as “mg/l”.
The measure = “17800” labeled as “mg/l”, was relabeled as “ug/l”.
4. Magnesium:The measure = “2500” labeled as “mg/l”, was relabeled as “ug/l”.
5. Potassium: No changes were necessary

For the minor metals dataset:
1. Units labeled as “mg/l Cr” were relabeled as “mg/l”
2. Units labeled as “ppb” were relabeled as “ug/l”
3. Copper: Measures < “0.19” labeled as “ug/l”, were relabeled as “mg/l”
Measures with the Monitoring Location Identifiers of “LEWWTP-BEAR CR”, “LEWWTP-DOWN”, “LEWWTP-UP”, where the values were > “0” labeled “mg/l”, were relabeled as “ug/l”.
4. Aluminum: Measures < “0.8” labeled as “ug/l”, were relabeled as “mg/l”
5. Nickel: Measures < “0.03” labeled as “ug/l”, were relabeled as “mg/l”
Measures > “1” labeled as “mg/l”, were relabeled as “ug/l”
6. Molybdenum: Measures > “0.8” labeled as “mg/l”, were relabeled as “ug/l”. This included only Monitoring Location Identifiers “LEWWTP-UP”, “LEWWTP-DOWN”, “LEWWTP-BEAR CR”.
7. Manganese: Measures > “19” labeled “mg/l” at Monitoring Location Identifier “SACWSD-MCKAY”, were relabeled “ug/l”.
For Monitoring Location Identifiers “LEWWTP-UP”, “LEWWTP-DOWN”, “LEWWTP-BEAR CR” where values > “0”, and for Monitoring Location Identifiers “ARR-SWSC-1”, “ARR-SWSC-2”, “ARR-SWSC-3”, “ARR-SWSC-4” where values are > “1” labeled as “mg/l”, were relabeled “ug/l”.
8. Cadmium: Measures < “0.008” labeled as “ug/l”, were relabeled as “mg/l”
The measure = “3730” labeled as “mg/l”, were relabeled as “ug/l”
9. Iron: Measures between “0.02” & “0.16 labeled as “ug/l”, were relabeled as “mg/l”
10. Lead: No changes were made.
11. Zinc: No changes were made.

For the major nonmetals dataset:
1. Units labeled as “mg/l CaCO3” were relabeled as “mg/l”
2. We omitted Silica from the analysis. Since Silica can be reported as SiO2 in water, but is often reported as Si. The conversion factor from SiO2 to Si is 0.467. Some people do not know about this issue and may have entered values incorrectly. As a result, the median values could be twice as high or half as large as they should be. This would be an artifact in the data and there is no systematic way to distinguish the correct entry.

3. Chloride: Measures < “10” labeled as “ug/l”, were relabeled as “mg/l”
4. Sulfate: “Sulfate as S_Dissolved” was relabeled as “Sulfate_Dissolved”.
5.

For the minor nonmetals dataset:
1. Units labeled as “ppb” were relabeled as “ug/l”
2. The Longitude entered as “-17.74332” was changed to “-107.74332”
3. Selenium: Measures < “0.01” labeled as “ug/l”, were relabeled as “mg/l”

For all datasets, measurements with units labeled as “mg/l” were multiplied by 1000 so all measurements are in ug/l.

```
## read county and zipcode information
zcta <- readOGR(dsn="Colorado_ZCTA/Colorado_ZCTA.shp")
counties <- readOGR(dsn="Colorado_County_Boundaries/Colorado_County_Boundaries.shp")

zcta <- readOGR(dsn="OregonZipcodes/ORE_zipcodes.shp")
counties <- readOGR(dsn="OregonCounties2015/orcntypoly.shp")

### Major metals
# read data of interest
dat1 = read.csv("X.csv", header = TRUE)

lon1 = dat1$LongitudeMeasure
lat1 = dat1$LatitudeMeasure

# convert coordinates to SpatialPoints object
# the first part of the coordinates
# the second part is the coordinate reference system
# and ensures sp_pts has the same CRS as zcta
coords_SpatialPoints = sp::SpatialPoints(cbind(lon1, lat1), CRS(projargs = proj4string(zcta)))

# determine which region each coordinates falls into
match_coords_to_zcta = over(coords_SpatialPoints, zcta)
# OBJECTID is the index of the ZCTA each coordinate falls into
# e.g., 336 means the 336th ZCTA
# ZCTA5CE10 and GEOID10 seem to both be the actual zip code
match_coords_to_counties = over(coords_SpatialPoints, counties)

# identify the coordinates not in a zcta
no_match_zcta = which(is.na(match_coords_to_zcta$OBJECTID))
length(no_match_zcta)
match_coords_to_zcta = apply(match_coords_to_zcta, 2, forcats::fct_explicit_na)
# identify the coordinates not in a county
no_match_counties = which(is.na(match_coords_to_counties$OBJECTID))
length(no_match_counties)

# plot zcta with coordinates that didn't match
#plot(zcta)
#points(coords_SpatialPoints[no_match_zcta,], pch = 20, col = "orange")

# update names of match* objects
names(match_coords_to_zcta)[1] = "zcta_idx"
names(match_coords_to_counties)[1] = "counties_idx"

# add zcta and countys ids to each observations in dat1
```

```
dat1 = cbind(match_coords_to_zcta, match_coords_to_counties, dat1[, -(1:2)])

# save for later use
save(dat1, file = "dat1_merged.rda", compress = "bzip2")
load("dat1_merged.rda")

## ElementRSFT3 is the variable name for the metals in my dataset
### zctas
r = dat1 %>% # on dat1
  group_by(ZCTA5CE10, ElementRSFT3) %>%
  summarize(median50 = median(Measure)) %>% # for each huc8id and ElementRSFT, compute
statistic in observed values
  gather(key = item, value = value, -c(ZCTA5CE10, ElementRSFT3)) %>% # Add Season here #
place each statistic in a separate row with appropriate measurement name
  arrange(ZCTA5CE10, ElementRSFT3) #Add season here # order th

r = tibble::add_column(r, ElemMeasure = paste0(r$ElementRSFT3, r$item)) %>% # Add
r$Season here # add a new column that combines ElementRSFT and statistic name
  dplyr::select(-c(ElementRSFT3, item)) %>% #Maybe add Season here. try without it. # then
remove ElementRSFT and item columns
  tidyr::spread(ElemMeasure, value)


# write this to file
write.csv(r, file = "datzcta_X.csv")

### counties
r = dat1 %>% # on dat1
  group_by(COUNTY, ElementRSFT3) %>% # #Add Season here #group data by huc8id and
ElementRSFT3
  summarize(median50 = median(Measure)) %>% # for each huc8id and ElementRSFT, compute
statistic in observed values
  gather(key = item, value = value, -c(COUNTY, ElementRSFT3)) %>% # Add Season here # place
each statistic in a separate row with appropriate measurement name
  arrange(COUNTY, ElementRSFT3) #Add season here # order them for convenience

r = tibble::add_column(r, ElemMeasure = paste0(r$ElementRSFT3, r$item)) %>% # Add
r$Season here # add a new column that combines ElementRSFT and statistic name
  dplyr::select(-c(ElementRSFT3, item)) %>% #Maybe add Season here. try without it. # then
remove ElementRSFT and item columns
  tidyr::spread(ElemMeasure, value) # then spread the values by ElemMeasure

write.csv(r, file = "datcounties_X.csv")
```