Low-Level Radon Exposure and Lung Cancer Mortality

Robert Obenchain[1], Stan Young[2], Goran Krstic[3]

[1]Risk Benefit Statistics LLC, Indianapolis, IN, 46250, USA

[2]CGStat LLC, Raleigh, NC, 27607, USA

[3]Fraser Health Authority, New Westminster, BC, Canada

Abstract

Background: It is agreed that high level radon exposure is harmful to humans. However, some published literature suggests that low levels of radon show no adverse effects or may even be protective. Claims made using traditional methods of analysis on observational data often fail to replicate. Here, we use a simple, alternative data-analytic strategy for examining effects of low-level indoor radon exposure on lung cancer mortality. Our objective will be to demonstrate that local population characteristics can alter expected effects.
Methods: Observational data on indoor radon exposure levels and lung cancer mortality for 2,881 US counties were obtained from federal and state governmental agencies. A new "statistical thinking" step-by-step analysis strategy called *Local Control* (LC) allows us to perform analyses of observational data that are more objective and "fair." LC analytical strategy makes as few and as realistic assumptions as possible. As a result, key LC inferences are nonparametric, and estimates of potentially heterogeneous treatment effect-sizes are more robust.
Results: Our LC analyses suggest that lung cancer mortality usually tends to decrease as low-level radon exposure increases. Local Rank Correlation (*LRC*) effect-sizes are shown to be predictable from confounding local characteristics like % residents over 65, % residents who currently smoke and % obese residents.
Conclusions: At low indoor radon exposure levels, reverse (negative) *LRC*s between radon exposure level and lung cancer mortality predominate. The strengths of these associations vary with local demographics.

Keywords
Local Control Strategy, Observational Data, Fair Comparisons, Causal Inference

Corresponding author: Stan Young (stan.young@omicsoft.com)

Author roles: Obenchain R: LC Strategy Development, R-package Development, R Graphics, Writing – Second Draft, Writing - Review & Editing, Validation; **Young S**: Conceptualization - Regression within Clusters, Writing – Original Draft, Writing - Review & Editing, Project Administration, Validation; **Krstic G**: Data Curation, Investigation, Validation, Writing – Review & Editing.

Competing interests: No competing interests were disclosed.

Grant information: Most of the work on this paper was performed pro bono. Work by Obenchain and Young on LC Methods and Software was partially funded by grants to Christophe G. Lambert (PI), University of New Mexico, from PCORI (CER-1507-31607) and NIH (1R21-LM012389). All work by Krstic on this paper was unfunded. Views expressed in this paper represent those of the authors alone and not of their respective organizations or sponsors.

How to cite this article: Obenchain RL, Young SS, Krstic G. **Low-Level Radon Exposure and Lung Cancer Mortality.**

First published:

Introduction

There is little controversy about whether high radon exposure levels cause lung cancer. In support of their conservative indoor radon standards, the U.S. Environmental Protection Agency (EPA) cites a pair of 2006 residential radon meta-analysis papers[1,2] based on case-control studies in Europe and North America, respectively. In sharp contrast, a 2018 meta-analysis[3] finds protection at low indoor radon levels. Between 1989 and 2007, multiple papers[4-10] were published on both sides of the "indoor radon causes lung cancer" question.

Published findings are potentially confusing because interactions are involved. For example, Darby et al.[1] find very low lung cancer rates for non-smokers at all radon levels but, for smokers, lung cancer rates do increase with radon exposure. Thus smoking appears to be a so-called "lurking" variable that can either emphasize or obscure causal effects.

Our indoor radon analyses are based on data amassed from U.S. federal and state sources[9-16]. Table I below gives names and brief descriptions for 11 characteristics of 2,881 U.S. counties or parishes, which represent 91.7% of the 3,142 county-like entities contained within the United

States. Unfortunately, comparable data from Alaska, Hawaii, New Hampshire, Nevada and the District of Columbia were not available.

We focus here on the possibility that variation in county average level of indoor radon exposure is the *primary cause* of variation in local lung cancer mortality outcomes. However, we also investigate the extent to which county characteristics, such as % residents over 65, % residents who currently smoke and % obese residents, have clear-cut *interaction effects* on exposure-mortality relationships.

An initial glance at our lung cancer mortality and indoor radon exposure data, depicted in Figure 1, suggests that mortality may indeed decrease as radon exposure level increases.

The (vertical) ***y***-outcome variable plotted in Figure 1 is the county lung cancer mortality rate (deaths per 100,000 person-years) while the (horizontal) treatment-exposure measure is the natural logarithm of the county average indoor radon level in pCi/L (picocuries per liter). Because county average indoor radon levels are reported only to the nearest 0.1 pCi/L in the raw data, the 10 counties with exposures reported as 0.0 are Winsorized in Figure 1 to log(0.05), which is roughly –3. Figure 1 also shows the **ordinary least squares line** and a **cubic spline fit** from **R**-functions `lm()` and `smooth.spline()`, respectively.

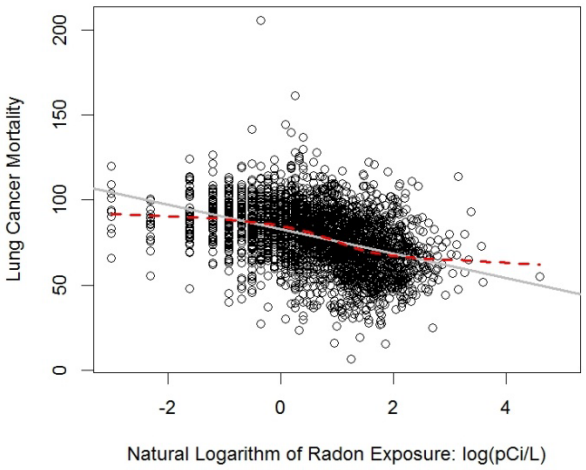

Figure 1. An Initial "Unadjusted" View

Our main objective here will be to conduct what is call a *Local Control* [17-21] analysis, LC, of the radon data set. LC strategy controls for county ***x***-characteristics (potential confounder variables) and displays visuals that help researchers locate and quantify interactions. LC first clusters counties with most-similar ***x***-characteristics together then measures mortality-exposure relationships locally, within each cluster. Each local measure is scalar-valued ...either a local average treatment effect (a binary difference between outcomes from two treatments) or else a local measure of strength of association (goodness-of-fit in linear regression). Because the indoor radon exposure measure is continuous (rather than binary) here, we will definitely need to use Local Rank Correlations, *LRC*s, to quantify exposure-mortality associations.

The ultimate objective of LC strategy can be to determine whether observed variation in *LRC* estimates across clusters can be reliably predicted using county demographic characteristics. A key intermediate LC step is to "confirm" that county demographic characteristics are *not ignorable* variables. Specifically, the observed *LRC* distribution across clusters of counties that are well-matched needs to be clearly different from the "artificial" *LRC* distribution resulting from *purely random clusters* of counties.

Quantitative prediction of *LRC* estimates from clusters of well-matched counties is illustrated here using recursive partitioning, a standard data mining method that reveals interactions. The overall stability of our LC analyses can be examined using sensitivity analyses that vary LC parameter settings, but that final topic is explored only within our Supplemental Materials.

In summary, Local Control analytical tactics are chosen to be as simple as possible, to make as few and as realistic assumptions as possible and, thus, to be nonparametric and/or robust in their estimation of potentially heterogeneous treatment effect-sizes. Our intension here is to illustrate this innovative and comprehensive "statistical thinking" strategy, which can be effectively applied to any sufficiently large data set.

Methods and Data

The presented case-study includes demographic and environmental data for 2,881 counties or parishes within the continental United States - amassed from various public sources[9-16]. The most

current version of our radon data set is in the public domain; see either **https://datadryad.org/** or **`demo(radon)`** using the **`LocalControlStrategy`**[21] **`R`**-package.

1) FIPS Code	Federal Information Processing Standard code (4 or 5 digits)
2) State	Two Character US State ID
3) County	County Name (character string)
4) Lung Cancer Mortality	Mortality Rate (Deaths per 100,000 Person-Years)
5) Radon	County Average Indoor Radon Exposure Level (pCi/L, rounded to a single decimal place.)
6) log(Radon)	Natural Logarithm of Radon (10 Counties with Indoor Radon rounded to 0.0 are Winsorized here to log(0.05) ≈ -2.996.)
7) Obesity	Percentage of County Residents who are Obese
8) Age Over 65	Percentage of County Residents who are Over 65
9) Currently Smoke	Percentage of County Residents who Currently Smoke
10) Ever Smoke	Percentage of County Residents who Ever Smoked
11) Median HH Income	Median Household Income in $1,000 (Contains a missing value for Shannon County, SD, FIPS = 46113.)

Table I. Eleven Characteristics of 2,881 U.S. Counties or Parishes

Local Control Strategy

LC analysis strategy[17-21] for cross-sectional observational data is easily explained. Non-technical audiences with basic understanding of clustering, linear regression, correlation and histograms are already familiar with its basic building blocks. LC starts by matching or clustering counties on their most important ***x***-characteristics ...while ignoring all information about county mortality and indoor radon exposure levels. The point is to assure that experimental units within a cluster are as alike as possible on their important baseline ***x***-characteristics (and as different as possible from counties within other clusters.) A simple two-variable (mortality vs exposure) analysis is then conducted within each ***x***-space cluster.

To apply LC strategy, we compute a *LRC* coefficient within each cluster. These local statistics enable "fair treatment comparisons" across clusters because all counties within the same cluster are relatively well-matched in ***x***-space. Next, we display the across-cluster distribution of *LRC* estimates in a simple histogram. Really small clusters (containing only 1 or 2 counties) fail to provide meaningful measures of exposure-mortality association and have to be discarded. This initial calculation of local associations can be thought of as a form of "nonparametric

preprocessing" of observational data. Viewing clusters as "Blocks" of similar counties, the overall LC model is suggestive of an unbalanced Nested ANOVA (*LRC* estimates within Blocks), where Blocks typically vary in size.

Initial LC inferences are nonparametric because they use permutation theory (resampling without replacement) to test whether the ***x*-**characteristics used to form clusters are truly *ignorable*. Specifically, one would compare the Observed distribution of *LRC* estimates computed from K clusters (of sizes N_1, N_2, ..., N_K) containing counties relatively well-matched in ***x*-**space with the corresponding "Random NULL *LRC*" distribution formed using many replications, M, where each resample (without replacement) forms K purely random clusters of the same given sizes (N_1, N_2, ..., N_K) as the clusters of well-matched counties. Inferences based upon random assignment of counties to clusters deliberately disregard the numerical values of all 11 county characteristics listed in Table I.

The primary LC analysis that we illustrate below in our **results** section uses **R**-package **`LocalControlStrategy`**[21] to form 50 "Ward" clusters of counties most similar on the three most important ***x*-**characteristics listed in Table I: Obesity, Age Over 65 and Currently Smoke. *LRC* associations between Lung Cancer Mortality and log(Radon) exposure variables are then estimated within these 50 "Blocks".

When the ***x*-**characteristics used to form clusters are *truly ignorable*, the Observed and Random NULL distributions of *LRC*s would be expected to be identical. Thus, whenever the Observed across-cluster *LRC* distribution is found to be clearly different from the Random NULL *LRC* distribution, this provides clear evidence that the assumption that county ***x*-**characteristics are *ignorable* is FALSE. Furthermore, if the total number of replications, M, is taken to be large enough, the Random NULL distribution of *LRC*s can usually be computed to any desired level of numerical precision. Nonparametric inferences based on M=1,000 random replications are presented in our **results** section.

Once *LRC* estimates from 50 clusters of well-matched counties are observed, they can be added, as a new variable (column), to the original data. Research attention can then (optionally) shift to focus on (supervised) prediction of across-cluster variation in these *LRC*s ...again using county

***x*-characteristics.** While traditional multiple regression techniques can be used to make such predictions, we favor use of the popular data mining method called *Recursive Partitioning*[22-23], also known as decision trees[24]. These partitioning methods create "tree models" by recursively selecting a "best" cut-point on one of the given ***x*-covariate** to divide a subgroup of counties into two parts. This splitting process continues until some "stopping rule" terminates each evolving tree-branch with a final "leaf" node.

Results

The initial phase of LC Strategy is clustering. A "sensitivity" analysis of Variance-Bias trade-offs in estimation of *LRC* distributions convinced us to use K = 50 "Ward" clusters. Figure 2 displays the resulting *LRC* distribution in a histogram with 14 non-empty "bins." Each bin has width 0.05 and height that "counts" the number of U.S. counties with an *LRC* estimate falling within that bin. While correlations can range from –1.0 to +1.0, we see that our 50 observed *LRC* estimates range here only between –0.70 and +0.10. In fact, more than half (1,624) of the N = 2,881 U.S. counties in the available data are members of clusters with *LRC*s in the five histogram bins between –0.45 and –0.20.

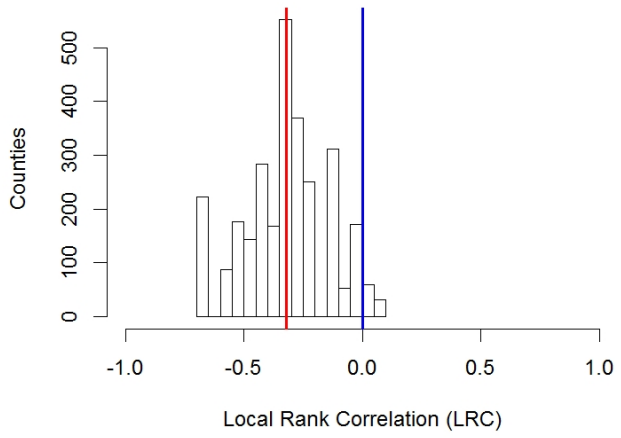

Figure 2. This Histogram shows the Observed *LRC* Distribution across 50 clusters. The overall mean *LRC* = -0.322 is denoted by the red vertical line within the modal bin, (-0.35, -0.30]. The blue vertical line at *LRC* = 0 shows that only two bins (containing 90 of 2,881 counties) have positive *LRC* estimates.

Note that observed *LRC*s are positive, but ***not*** *significantly greater than zero*, within only the two right-most bins of Figure 2. The (0.00, +0.05] bin contains a cluster of 59 counties, while (+0.05, +0.10] contains a cluster of 31 counties.

It is also instructive to examine scatter plots (radon exposure vs. mortality) for the counties within an individual cluster. Figures 3-5 illustrate such plots for three different clusters. Note that all 3 plots cover the very same exposure-mortality range as Figure 1.

Figure 3 plots shows exposure-mortality outcomes for the cluster of 59 counties that falls within the (0.0, +0.05] bin of Figure 2. Note that the **R** `smooth.spline()` **fit** shown in Figure 3 suggests why the local Pearson correlation is negative even though the corresponding *LRC* estimate is positive (+0.035).

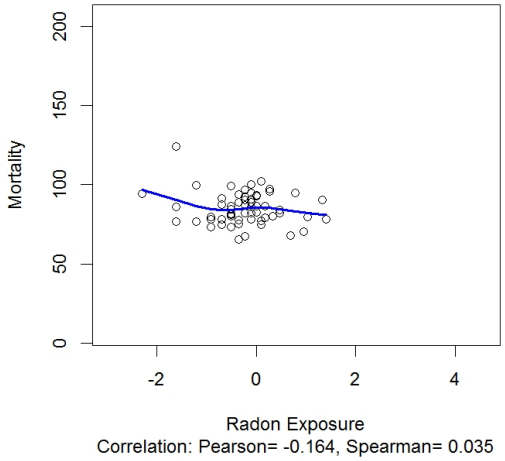

Figure 3. An observed *LRC* of +0.035 comes from a cluster of 59 counties. The corresponding local Pearson correlation is negative (-0.164) but not significant.

Figure 4 shows the exposure-mortality scatter within the cluster of 73 counties that has the most negative *LRC* = –0.687 ($p < 0.0001$). This cluster is one of three (totaling 222 counties) that fall within the extreme left bin, (–0.70, –0.65], of Figure 2.

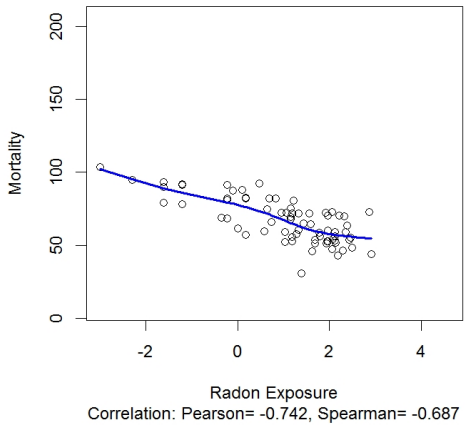

Figure 4. The most negative *LRC* = -0.687 estimate for a cluster of 73 Counties.

Finally, Figure 5 shows the scatter within the largest of 50 clusters (153 counties) with *LRC* = –0.3177 ($p < 0.0001$). This cluster is one of 7 (totaling 552 counties) that fall into the modal bin of Figure 2: (–0.35, –0.30].

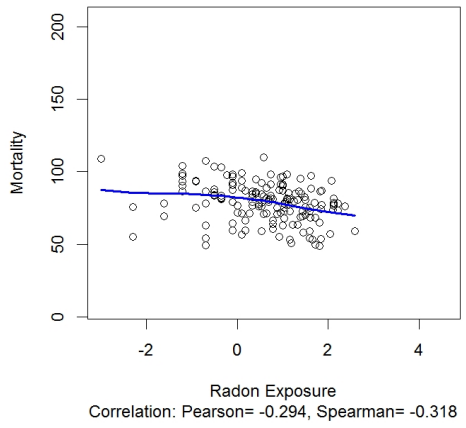

Figure 5. The single largest cluster (153 counties) has estimated *LRC* = -0.318.

In summary, the observed *LRC* distribution (Figure 2) that results from micro aggregation of US counties using their three primary ***x***-confounder characteristics to form 50 clusters of relatively well-matched counties is instructive in several ways. First, it shows that *Higher levels of Low indoor radon exposure are much more likely to be protective against lung cancer mortality (negative association) than to possibly cause it (positive association)*.

We also see a wide range of numerical *LRC* estimates. Is this *LRC* variation greater than what would be expected due to chance? Could this variation be attributable to corresponding variation in county ***x***-characteristics?

We will address both of the above questions in two distinct ways. First, we will infer that the county ***x***-characteristics used to form clusters are *not ignorable*. Then we will show that these same ***x***-characteristics are useful in actually predicting *LRC* variation.

County *x*-characteristics are *NOT Ignorable*!

Statistical inference compares an observed *LRC* distribution to its NULL distribution under the falsifiable hypothesis that the given ***x***-characteristics are actually *ignorable*. This NULL distribution is constructed by merging together 2,881 *LRC* estimates from each of M=1,000 replications. In each replication, (a) all 2,881 counties are *randomly* assigned to 1 of 50 pseudo-clusters of the same sizes, (N_1, N_2, ..., N_{50}), as the 50 observed clusters of well-matched counties, and (b) 2,881 NULL *LRC* estimates are calculated across each resulting set of 50 *random* pseudo-clusters.

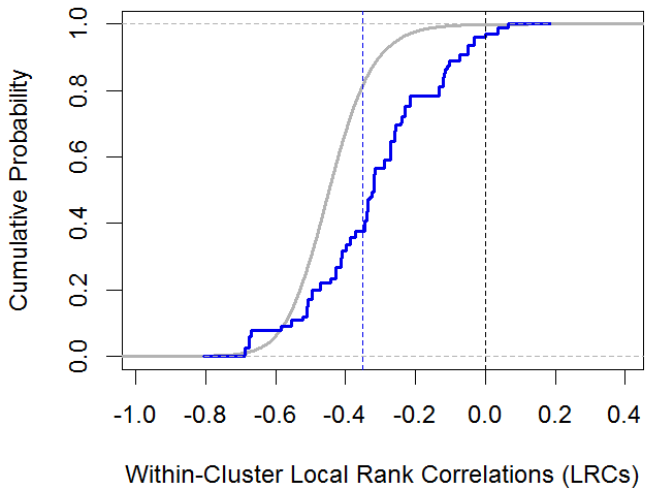

Figure 6. LC Confirm Phase: Empirical CDF Comparison of the Observed *LRC* Distribution with its Random NULL Distribution from M=1,000 replications.

It is visually clear from Figure 6 that the observed and random permutation *LRC* distributions have quite different Cumulative Distribution Functions (CDFs). The **`confirm()`** function[21] applies a Kolmogorov-Smirnov two-sample test that yields a D-statistic of 0.4539 at roughly *LRC* = –0.35 (**dashed vertical line**) in Figure 6.

An additional M=1,000 independent, random replications were then generated using the **`KSperm()`** function[21] to compute 1,000 NULL D-statistics ...all of which turned out to be less than 0.2147, i.e. much smaller than 0.4539. Thus the true p-value associated with the observed D = 0.4539 is estimated to be strictly less (and probably *much less*) than 0.001. Thus the hypothesis that the given x-covariates are *ignorable* is easily rejected (falsified) here.

This leaves only the final (optional stretch-goal) phase of LC strategy. This final objective is to reveal the extent to which *LRC* estimates within clusters are ***Heterogeneous*** (***predictable fixed-effects***) rather than ***Homogeneous (unpredictable random-effects)***.

Because clusters commonly vary considerably in size, it is essential to attach ***weights*** to individual *LRC* (or *LTD*) estimates when fitting across-cluster models. Our experience is that simply using weights directly proportional to cluster sizes is both realistic and robust. All of the predictor variables, including radon exposure level itself, can then be used in attempts to predict the Observed distribution of *LRC* associations.

LC strategy imposes no restrictions on choice of the supervised learning method used for predictive modeling during this (optional) final LC Reveal phase. Again, we find recursive partitioning[22-24] particularly helpful in detecting and "displaying" interaction effects.

Our favorite tree model, depicted in Figure 7, is based on (nonparametric) permutation theory using the **`party`** **`R`**-package[22]. Like other recursive partitioning methods, **`party`** searches across potential predictor variables to find a best “cut point” for separating data subsets into parts, usually two. Each resulting subset of counties is then split using a "stopping rule." In Figure 7,

the tree was restricted to have binary splits at only 3-levels, yielding $2^3 = 8$ final "leaf" nodes.

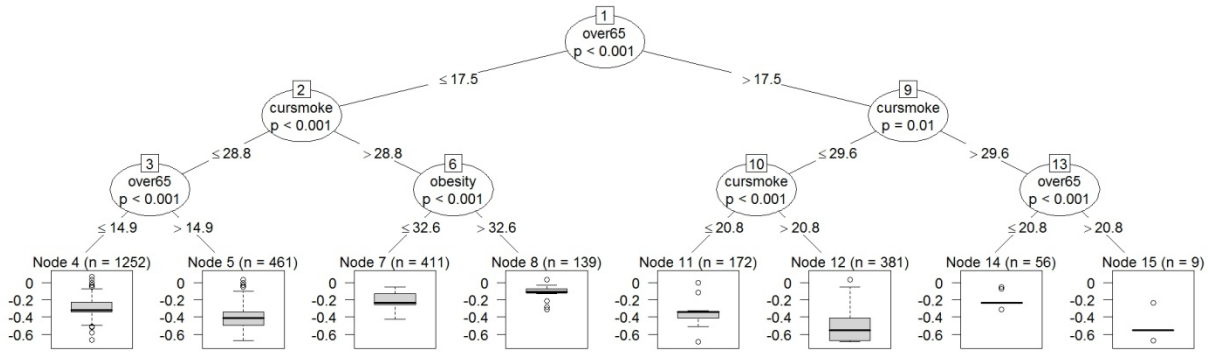

Figure 7. `party` R-package Tree Model for predicting LRC estimates (supervised learning).

Note that Node #4 is quite large (1,252 counties), and its *LRC* distribution (displayed by a "box-and-whisker" diagram) is similar to the full *LRC* distribution for all 2,881 counties. Next, note that Node #8 (139 counties) has the *LRC* sub-distribution with the lowest proportion of significantly negative mortality-exposure *LRC*s. Meanwhile, Node #5 (461 counties) and Node #7 (411 counties) have *LRC* distributions that are only a little less negative than "typical" (Node #4). But three of the final four nodes (#11, #12 and #15) have *LRC* sub-distributions even ***more*** negative than "typical." Table II summarizes these three major sub-groupings of *LRC* sub-distributions.

Counties with *LRC* Distributions Somewhat Less Negative than Typical	Counties with Typical (Mostly Negative) *LRC* Distributions	Counties with *LRC* Distributions Even More Negative than Typical
606 Counties (21.0%)	1,252 Counties (43.5%)	1,023 Counties (35.5%)

Table II. LC Reveal Phase comparison of *LRC* Sub-Distributions

The **`party`** tree of Figure 7 is rather "small" in the sense that it uses only 7 splits (defining only 8 leaf nodes), but it appears to do a remarkably good job of predicting nonparametric *LRC* estimates using only *three* ***x*****-**confounders. This "predictability" claim is, perhaps, better illustrated using a more conventional RP method[23] that characterizes nodes using their *LRC*

mean values ...with focus upon the splits that are *most significant* in an ANOVA-like sense. These traditional sorts of RP trees rarely correspond to "full" trees like Figure 7, where every intermediate node is split into two nodes. RP "unbalanced" trees can maximize overall goodness-of-fit (R^2) for any given total number of splits (7 here.)

On the other hand, we deliberately created Figure 8 using JMP®[24] by requesting the *very same splits* displayed in the **`party` R**-package tree of Figure 8. The LogWorth statistics displayed in the seven intermediate nodes of Figure 8 are defined as the negative of the base 10 logarithm of the p-value for the split below that node. These statistics further confirm that 6 of the 7 splits are indeed highly significant; the split of Node 10 on % elderly residents at 17.5% has the largest (least significant) traditional p-value of 0.00014. Furthermore, the overall goodness-of-fit is R^2= 0.472; this quite *simple* RP Tree model explains just slightly less than half of the total across-cluster variation in *LRC* estimates.

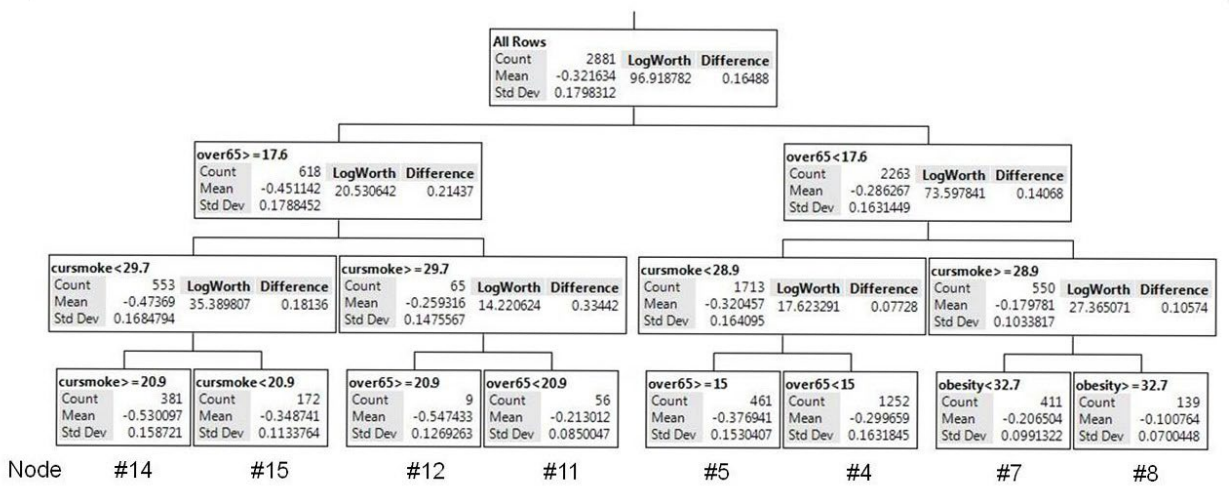

Figure 8. SAS / JMP® representation of the `party` tree in Figure 7.

All three ***x***-confounders used in the prediction tree shown in Figures 7 and 8 make common sense. The denominator of each Lung Cancer Mortality rate (deaths per 100,000 person-years)

includes county residents of *all ages*. Since cancer deaths are more likely to occur in elderly residents, it's no wonder that % of residents over 65 is used to make 3 of the 7-splits depicted in Figures 7 and 8. However, the distinct surprise here may well be that *LRC*s are consistently predicted to be *smaller* (more negative) in counties where elderly residents are *more prevalent*!

Only one highly significant split on % obese residents was among the "best" 7 splits. Note that the 139 counties in Final Node #8 with obesity at least 32.7% has a higher (less negative) *LRC* prediction of -0.101 than the *LRC* prediction of -0.207 for 411 counties in Final Node #7 with lower obesity rates.

Given the well-known and strongly-positive association between lung cancer mortality and smoking, it could be considered surprising that % residents who currently smoke is used in ***only three*** of the seven binary splits needed to create the "full" tree with 3 levels (8 final nodes). On the other hand, it is quite unfortunate that *separate* lung cancer mortality statistics for smokers and non-smokers were not available for U.S. counties. This unfortunate aggregation of lung cancer mortality rates essentially prevents effective use of LC strategy to address the traditional question: "Is there evidence that *smoking* is a primary cause of lung cancer mortality in the U.S. county data?"

Finally, note that radon exposure level [specifically, the log(radon) measure of Table I] was *not selected* for use in any of the 7 most predictive splits. In fact, since our simple tree model failed to select any measure of radon exposure level as a predictor of *LRC*s, we conclude that indoor radon exposure levels are relatively poor predictors of *LRC* associations between indoor radon exposure and lung cancer mortality.

Thus, roughly half of all across-cluster variation in *LRC*s appears to be purely random, while the other half appears to be predictable simply by the age and life-style characteristics of local residents. The effects of indoor radon exposure on lung cancer mortality in the US thus appear to be at least partially heterogeneous (predictable). This suggests that % over 65, % obese and % current smokers are meaningful "modifiers" of radon exposure effects on lung cancer mortality.

In summary, we have provided both strong visual evidence and sound statistical inferences supporting our arguments that *low indoor radon exposures can be protective against lung cancer mortality* rather than be a potential cause of lung cancer mortality. Our LC analyses dividing 2,881 U.S. counties into 50 clusters (relatively well-matched subgroups) yield *covariate adjustments* with much more meaningful *policy implications* than the simplistic scatter-plot displayed in Figure 1. We have both confirmed that ***x*-**matching *truly matters* in estimation of *LRC* distributions and also revealed that county ***x*-**characteristics can literally help predict observed variation in *LRC* estimates.

Discussion/Summary

What happens at low indoor radon exposure levels is difficult to evaluate empirically given usual sample sizes, variability and analysis perspectives that differ not only across subject-matter areas but also with the psychology of individual researchers. Several researchers have noticed low-dose, nonlinear relationships commonly described as U- or J-shaped. A variety of names have been given to this phenomenon: "autoprotection, heteroprotection, adaptive response, preconditioning, hormesis, xenohormesis, paradoxical…"[25]. Thus, the early observations of Cohen[5-8] appear to fit well into a much larger context whereby stress elicits protective effects, Parsons [26]. In fact, Parsons asserts that "…hormesis for ionizing radiation becomes an evolutionary expectation at exposures substantially exceeding background." Feinendegen[27] comments on radon hormesis as follows: "It develops with a delay of hours, may last for days to months, decreases steadily at doses above about 100 mGy to 200 mGy and is not observed any more after acute exposures of more than about 500 mGy." It is reasonable to consider our LC indoor radon exposure findings in this context.

The reliability of a claim coming from observational data is important. Local Control strategy does several things to support reliability. Covariates are controlled via clustering. The single question at issue is examined within each cluster. Often the answer to research questions on local effects is that "they depend."

A unique feature of LC strategy is its initial emphasis on *unsupervised, nonparametric* inference (permutation testing) to determine whether ***x*-**characteristics of experimental units (counties) are ignorable. One-size-fits-all radon mitigation policies can be fully justified only when all

available ***x*-**characteristics are indeed ignorable. Otherwise, rigid enforcement of the current low EPA threshold (> 4 pCi/L) for *requiring* radon mitigation might even *increase* expected lung cancer mortality.

Another unique feature of LC strategy is that within-cluster estimation of *LRC* (or *LTD*) distributions essentially moves the Exposure (or Treatment choice) variable to the left-hand-side of the *supervised* and possibly *parametric* models commonly used within the final (optional) LC "Reveal" phase for prediction of local outcomes. This left-ward "shift" typically results in models with much better fit to the dependent variable (*LRC* or *LTD*) than traditional models with only the y-outcome (lung cancer mortality) on the left-hand-side. After all, when the Exposure measure or Treatment choice indicator is on the right-hand-side of the model equation, it must then literally *compete* with all available ***x*-**covariates as potential predictors. Furthermore, right-hand-side variables are typically somewhat inter-correlated and may even be ill-conditioned (nearly multicollinear.)

The negative *LRC*s between low levels of indoor radon exposure and lung cancer mortality within U.S. counties observed here agree with the results of a cohort study in Ontario uranium miners by Navaranjan et al.[28] Our Figure 9 is based on these Ontario uranium miner estimates. When the Ontario Relative Risk (RR) measure of lung cancer incidence is plotted against cumulative exposure to radon, expressed in Working Level Months (WLM), a hormetic J-shaped relationship results. Specifically, note that this relationship appears inverse only at low levels of occupational exposure to radon (i.e., at radon levels below approximately 10 WLM). Figures 8 and 9 of Navaranjan et al.[28] also show consistently inverse relationships at low levels of cumulative WLM exposure for lung cancer mortality and incidence, respectively.

The Health Physics Society[29] provides a conversion from WLM to Bq/m^3, assuming 7,000 hours spent indoors per year, where 10 WLM would be equivalent to ~2,273 Bq/m^3 of indoor radon. Our LC analysis results indicate that an increase of county-level lung cancer mortality is not observable at indoor radon concentrations below ~16 pCi/L (i.e. ~592 Bq/m^3), which would be equivalent to ~2.6 WLM exposure (i.e., well within the range of the inverse relationship segment of the J-shaped curve in our Figure 9.)

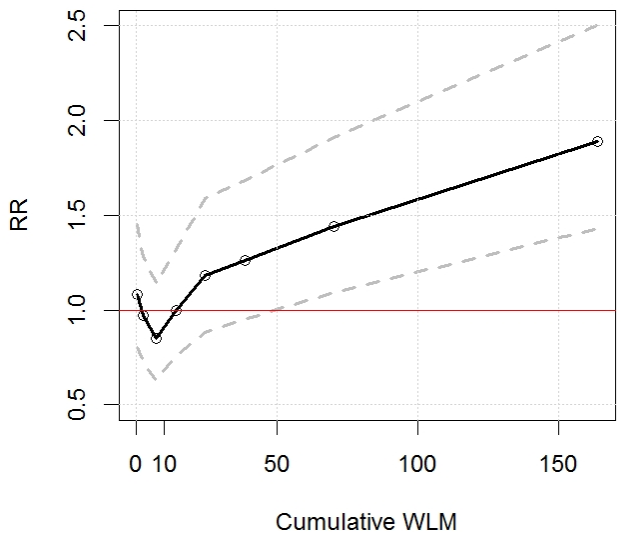

Figure 9. Reproduction of Ontario Uranium Miner results[28] using R software.

In addition, a linear regression analysis of available data from 26 countries of the Organization for Economic Co-operation and Development (OECD), including North America, shows a weak inverse (negative) correlation for age-standardized lung cancer mortality vs. mean indoor radon concentration, compared to a statistically significant positive correlation of lung cancer with smoking prevalence[30]. These findings are also in agreement with the LC results presented here.

Our Local Control analyses support the claim that lung cancer mortality decreases as low-level indoor radon exposure increases, with effect-sizes being largely predictable from confounding county characteristics like % residents over 65, % residents who currently smoke and % obese residents.

References

1. Darby S, Hill D, Deo H, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, Falk R, Farchi S, Figueiras A, Hakama M, Heid I, Hunter N, Kreienbrock L, Kreuzer M, Lagarde F, Mäkeläinen I, Muirhead C, Oberaigner W, Pershagen G, Ruosteenoja E, Rosario AS, Tirmarche M, Tomásek L, Whitley E, Wichmann HE, Doll R. Residential radon and lung cancer: detailed results of a collaborative analysis of individual data on 7148 persons with lung cancer and 14,208 persons without lung cancer from 13 epidemiologic studies in Europe. *Scandinavian journal of work, environment & health 2006*; 32 Suppl 1: 1-84.

2. Krewski D, Lubin JH, Zielinski JM, Alavanja M, Catalan VS, Field RW, Klotz JB, Létourneau EG, Lynch CF, Lyon JL, Sandler DP, Schoenberg JB, Steck DJ, Stolwijk JA, Weinberg C, Wilcox HB. A combined analysis of North American case-control studies of residential radon and lung cancer. *Journal of Toxicology and Environmental Health*, 2006; 69: 533-597.

3. Dobrzyński L, Fornalski KW, Reszczyńska J. Meta-analysis of thirty-two case–control and two ecological radon studies of lung cancer. *Journal of Radiation Research*. 2018; 59: 149-163.

4. Appleton JD. Radon: sources, health risks and hazard mapping. *AMBIO: A Journal of the Human Environment*, 2007; 36: 85-89. doi: http://dx.doi.org/10.1579/0044-7447(2007)36[85:RSHRAH]2.0.CO;2.

5. Cohen BL. Expected indoor 222 Rn levels in counties with very high and very low lung cancer rates. *Health Physics*, 1989; 57: 897-907.

6. Cohen BL. Test of the linear-no threshold theory of radiation carcinogenesis for inhaled radon decay products. *Health Physics*, 1995; 68: 157-174.

7. Cohen BL. Lung cancer rate vs. mean radon level in U.S. counties of various characteristics. *Health Physics*, 1997; 72: 114-119.

8. Cohen BL. The linear no-threshold theory of radiation carcinogenesis should be rejected. *J. Amer. Physicians and Surgeons*, 2008; 13: 70-76.

9. National Research Council (NRC). Committee on Health Risks of Exposure to Radon: BEIR VI. Health Effects of Exposure to Radon. Washington, DC: National Academy Press, 1999.

10. International Agency for Research on Cancer (IARC). Man-made Mineral Fibres and Radon. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Volume 43, 1998. World Health Organization (WHO), Lyon, France.

11. National Cancer Institute (NCI). Cancer Mortality Maps - U.S. National Institutes of Health (NIH), 2015a. Available at: http://ratecalc.cancer.gov/ratecalc/ (accessed in July 2015).

12. National Cancer Institute (NCI). Small Area Estimates for Cancer Risk Factors and Screening Behaviors - Ever Smoking Prevalence (Age 18+). U.S. National Institutes of Health (NIH), 2015b. Available at: http://sae.cancer.gov/estimates/lifetime.html (accessed in July 2015).

13. U.S. Census Bureau - American Fact Finder: 2000 Census of Population and Housing. http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_00_SF1_DP1 (accessed in July 2015)

14. U. S. Census Bureau. Small Area Income and Poverty Estimates. U.S. Department of Commerce, 2015. Available at: http://www.census.gov/did/www/saipe/data/statecounty/data/ (accessed in July 2015).

15. Masnick, M. 2011: U.S. 2008 obesity rates at the county level. Available at: http://www.maxmasnick.com/2011/11/15/obesity_by_county/ (Accessed in July 2015).

16. U.S. Environmental Protection Agency (EPA). Screening indoor radon data from the State Residential Radon Survey (SRRS), 2014 (Obtained through personal communication with the U.S. EPA - Radiation & Indoor Environments Division).

17. Obenchain, RL. The local control approach using JMP. *Analysis of Observational Health Care Data using SAS*, ed. D. E. Faries, A. C. Leon, J. M. Haro, and R. L. Obenchain, 2010; 151–192. Cary, NC; SAS Press.

18. Obenchain RL, Young SS. Advancing statistical thinking in observational health care research. *Journal of Statistical Theory and Practice*, 2013; 7: 456-469.

19. Wolfinger RD and Obenchain RL. JMP® Add-Ins Module for **Local Control**. https://community.jmp.com/docs/DOC-7453. *SAS Institute Inc.,* Cary, NC, 2015.

20. Obenchain RL, Young SS. Local Control Strategy: Simple Analyses of Air Pollution Data can reveal Heterogeneity in Longevity Outcomes. *Risk Analysis,* 2017; 37:1742-1753. (OnLine DOI: 10.1111/risa.12749)

21. Obenchain RL. ***LocalControlStrategy***: **R**-package for Robust Analysis of Cross-Sectional Data. Version 1.3.2; Posted 2019-01-07. https://CRAN.R-project.org/package=LocalControlStrategy

22. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Grap. Stat.*, 2006; 15(3), 651-674.

23. SAS JMP® Software. Analyze > Predictive Modeling > **Partition**. Version *13.1.0.* SAS Institute Inc., Cary, NC, 2016.

24. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, JaKa M, French S. Decision trees in epidemiological research. *Emerg Themes Epidemiol.* 2017; 14: 11. DOI: 10.1186/s12982-017-0064-4

25. Calabrese EJ, et al. Biological stress response terminology: Integrating the concepts of adaptive response and preconditioning stress within a hermetic dose – response framework. *Toxicology and Applied Pharmacology*, 2017; 222: 122-128.

26. Parsons PA. Radiation hormesis: Challenging LNT theory via ecological and evolutionary considerations. *Health Physics*, 2002; 82: 513–516.

27. Feinendegen LE. Evidence for beneficial low level radiation effects and radiation hormesis. Br J Radiol. 2015; 78: 3-7. DOI: 10.1259/bjr/63353075

28. Navaranjan N, Berriault C, Demers PA, Do M, and Villeneuve P. Ontario Uranium Miners Cohort Study Report. The Occupational Cancer Research Centre (OCRC), Cancer Care Ontario, Canada; 2015. https://tspace.library.utoronto.ca/bitstream/1807/74748/1/RSP-0308.pdf

29. Health Physics Society (HPS). Environmental and Background Radiation — Radon; 2012. https://hps.org/publicinformation/ate/q10245.html

30. Krstic G. Radon versus other lung cancer risk factors: How accurate are the attribution estimates? *Journal of the Air & Waste Management Association*, 2017; 67(3): 261-266. DOI: 10.1080/10962247.2016.1240725

Our Key Messages

- Local Control strategy starts by clustering counties that are most similar on resident age, smoking and obesity percentages. Within each cluster, LC then measures the Local Rank Correlation (*LRC*) between lung cancer mortality and indoor radon exposure level.

- LC strategy provides strong control over the covariates that identify meaningful subgroups of similar counties (experimental units), allowing detection of local effects.

- The across-cluster distribution of *LRC* estimates can be predicted using recursive partitioning or other model-fitting methods.

- Across 2,881 U.S. counties, *LRC*s between radon exposure and lung cancer mortality tend to be predominantly negative, the degree of which depends on county characteristics.

- By applying LC analysis strategy to U.S. federal and state data for 2,881 counties, we have demonstrated that higher levels of low indoor radon exposure are much more likely to be protective against lung cancer mortality than to cause it.