# Low-level Radon Exposure and Lung Cancer Mortality

Robert Obenchain[a], S. Stanley Young[b,*], Goran Krstic[c]

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

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

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

[*]Corresponding Author: genetree@bellsouth.net

## 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. One objective is 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 U.S. 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" than regression-like methods. 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 robust.

**Results:** Our LC analyses suggest that lung cancer mortality usually tends to decrease as background radon exposure increases. Local rank correlation (LRC) effect-sizes are shown to be predictable from confounding local characteristics like percentage of residents over 65, percentage of residents who currently smoke and percentage of obese residents.

**Conclusions:** At low indoor radon exposure levels, reverse (negative) LRCs 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

## Introduction

There is little controversy about whether high radon exposure levels cause lung cancer. In support of their conservative indoor radon mitigation standards, the U.S. Environmental

Protection Agency (EPA) cites a pair of residential radon meta-analyses based on case-control studies in Europe (Darby et al., 2006) and North America (Krewski et al., 2006). In sharp contrast, a recent meta-analysis (Dobrzyński, Fornalski aand Reszczyńska, 2018) finds protection at low indoor radon levels. Between 1989 and 2008, at least seven other publications added fuel to the "indoor radon causes lung cancer" debate (Cohen, 1989, 1995, 1997, 2008; International Agency for Research on Cancer, 1998; National Research Council, 1999; Appleton, 2007.)

Published findings are potentially confusing because interactions are involved. For example, Darby et al. (2006) found 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: a variable that can either emphasize or obscure potential effects of other factors.

Our indoor radon analyses are based on data amassed from U.S. federal and state archives (National Cancer Institute, 2015a, 2015b; U.S. Census Bureau - American Fact Finder, 2000; U. S. Census Bureau, 2015; Masnick, 2011; U.S. Environmental Protection Agency, 2014).

Table 1 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. Potential strengths and weaknesses of these data are summarized in an Appendix.

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

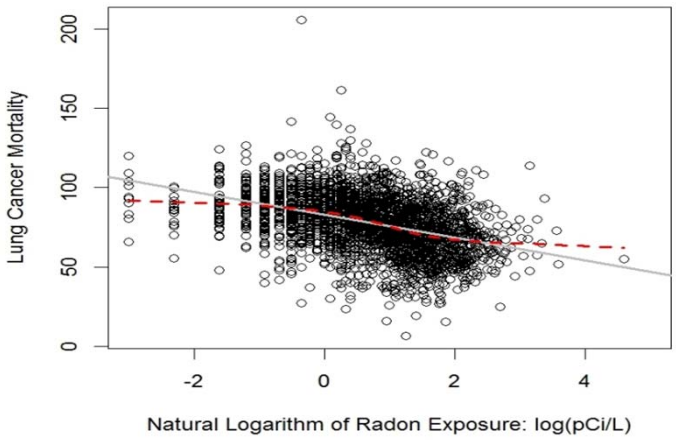


**Figure 1. An Initial "Unadjusted" View**

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 radon exposures reported as 0.0 are Winsorized in Figure 1 to log(0.05), which is roughly -3, at the left-hand extreme of Figure 1. This figure also shows both the ordinary least squares line and a cubic smoothing spline fit using default settings for the smooth.spline() R-function.

Our main objective will be to conduct what is call a Local Control (LC) analysis of the relationship between radon exposure and lung cancer mortality (Obenchain, 2010, 2019; Obenchain and Young, 2013, 2017; Wolfinger and Obenchain, 2015). LC methods control for county *x*-characteristics (potential confounder variables) and display visuals that help researchers locate and quantify effects of interactions. Application of LC strategy starts by clustering together U.S. counties with most-similar *x*-characteristics, then measures the strength of the exposure-mortality relationship locally, within each cluster. LC strategy requires use of a local effect-size measure that is scalar-valued, and two such measures have been implemented in R software (Obenchain, 2019). If our exposure variable were binary (an indicator for two "treatment" choices), our effect-size measure would be a Local Treatment Difference (LTD) between mortality averages ("new" minus "control"). Since radon exposure levels are continuous here (except for rounding), the effect-size measure of interest will be the Local Rank Correlation (LRC) between radon exposure and lung cancer mortality. Each LRC estimate can be viewed as the Slope of a best-fitting-line within its cluster of U.S. counties. Meanwhile, the absolute value (or square) of each LRC quantifies the local strength of exposure-mortality association (goodness-of-fit for a local linear regression based on ranks), while the numerical sign of each LRC signals whether mortality rate increases (+) or decreases (-) as radon exposure level increases.

The ultimate objective of LC strategy can be to determine whether observed variation in LRC estimates across clusters can be reliably predicted using county-level demographic ***x***-characteristics. A key intermediate LC step is to "Confirm" that county ***x***-characteristics are not Ignorable. Note that, when county ***x***-characteristics are Ignorable, clusters formed using them would be "meaningless" and purely "random". Thus, we will show that that "Ignorable Confounders" is a falsifiable (NULL) hypothesis for the radon exposure-morality data. Specifically, we will accurately simulate the LRC-like distribution resulting from forming clusters purely at random (ignoring ***x***-characteristics) and compare that distribution with the "true" LRC distribution resulting from clusters of U.S. counties relatively well-matched in ***x***-space.

Finally, quantitative prediction of LRC estimates will be 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 on the LocalControlStrategy R-package, Obenchain(2019).

In summary, LC 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 primary intension here is to illustrate this innovative and comprehensive "statistical thinking" strategy, which can be effectively applied to any sufficiently large set of cross-sectional data.

## Methods

## Data

The data analyzed here are described in Table 1. We have placed these data in the public domain in the sense that a "radon" data.frame is part of the LocalControlStrategy R-package. Anyone may freely download this package and all other statistical software needed to reproduce the LC analyses described here. For example, novice users of R can use "commands" like: demo(radon).

**Table 1. Eleven characteristics of 2,881 U.S. counties or parishes**

| | | |
|---|---|---|
| **1** | FIPS Code | Federal Info. Processing Standard (unique, 4 or 5 digit code) |
| **2** | State | Two Character U.S. State ID |
| **3** | County | County or Parish Name (character string) |
| **4** | Lung Cancer Mortality | Deaths per 100,000 Person-Years |
| **5** | Radon | Average Indoor Exposure (pCi/L, single decimal place) |
| **6** | Natural log(Radon) | Values reported as 0.0 are Winsorized to log(0.05) = -2.966 |
| **7** | Obesity | Percentage of County Residents considered Obese |
| **8** | Age Over 65 | Percentage of Residents Over 65 |
| **9** | Currently Smoke | Percentage of Residents who Currently Smoke |
| **10** | Ever Smoke | Percentage of Residents who Ever Smoked |
| **11** | Median HH Income | Household Income in $1,000; One missing value (FIPS = 46113) |

## LC strategy

LC strategy for analysis of 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 deliberately 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 analysis, using county ranks on mortality and exposure levels, is then conducted within each *x*-space cluster.

To apply LC strategy, we first compute a LRC coefficient for 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 must 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 fitting an unbalanced nested ANOVA: LRC estimates within Blocks that vary in size.

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 all but two (Lung Cancer Mortality & Radon) of the 11 county characteristics listed in Table 1.

The primary LC analysis that we illustrate below in our Results section uses LocalControlStrategy to form 50 "ward.D" clusters of counties most similar on the three most important *x*-characteristics listed in Table 1: Obesity, Age Over 65 and Currently Smoke. LRC associations between Lung Cancer Mortality and log(Radon) exposure variables are then estimated within these 50 design-like "Blocks".

When the *x*-characteristics used to form clusters are truly ignorable, the observed and random NULL distributions of LRCs 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 LRCs can usually be computed to any desired level of numerical precision. Nonparametric inferences based on 1,000 random replications are presented in our Results section.

Once LRC estimates from 50 clusters of well-matched counties have been computed, 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 LRCs, 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 (Hothorn, Hornik and Zeileis, 2006; SAS JMP® Software, 2016), also known as decision trees (Venkatasubramaniam et al., 2017). These partitioning methods create "tree models" by recursively selecting a "best" cut-point on one of the given *x*-covariates 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.D" 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 LRCs in the five histogram bins between –0.45 and –0.20.

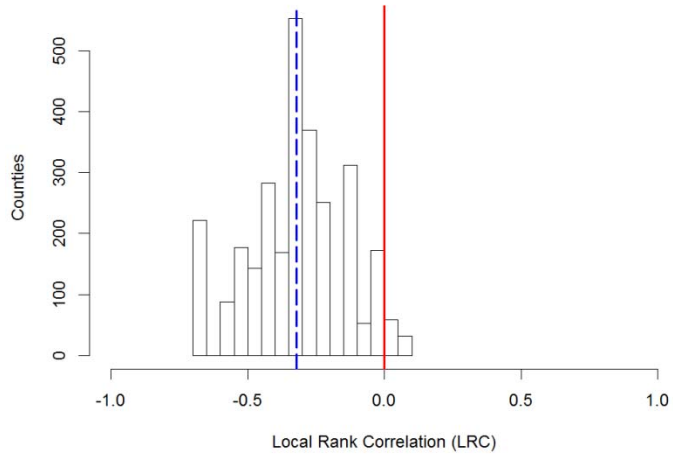


**Figure 2. This histogram shows the observed LRC Distribution across 50 clusters. The overall mean LRC = -0.322 is denoted by the dashed vertical line within the modal bin, (-0.35, -0.30]. The vertical line at LRC = 0 shows that only two bins (containing 90 of 2,881 counties) have positive LRC estimates.**

Note that these observed LRCs 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 three plots cover the very same exposure-mortality range as Figure 1.

Figure 3 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).

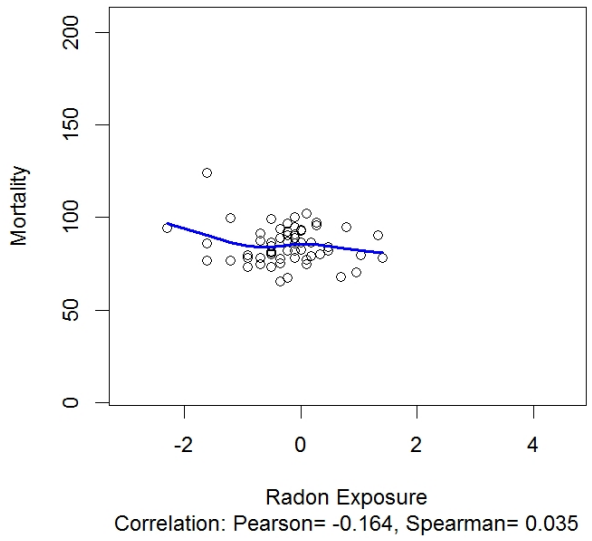


**Figure 3. An observed LRC of +0.035 comes from a cluster of 59 counties. The corresponding local Pearson correlation is negative (-0.164) but not significant.**

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.

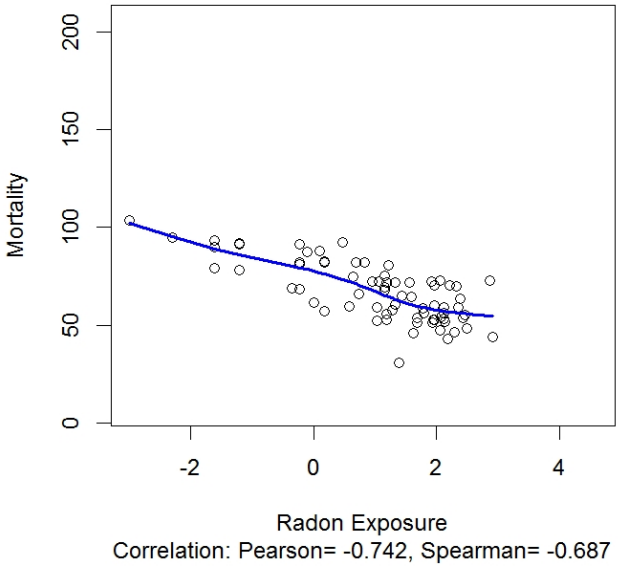


**Figure 4. The most negative LRC = -0.687 estimate for a cluster of 73 Counties.**

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].

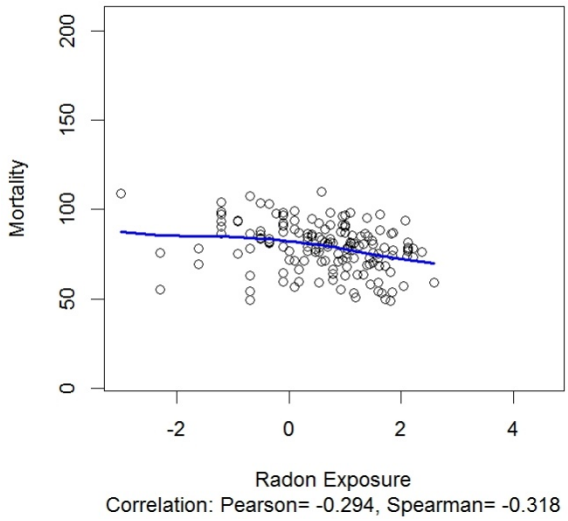


**Figure 5. The single largest cluster (153 counties) has estimated LRC = -0.318.**

In summary, the observed LRC distribution (Figure 2) is instructive in several ways. After all, it results from micro aggregation of 2,881 US counties, uses three primary *x*-confounder characteristics, and forms 50 clusters of relatively well-matched counties. First, Figure 2 shows a staggeringly high prevalence (in Counts) of US counties where lung cancer mortality tends to decrease as low-level indoor radon exposures increase (i.e. negative LRC associations) over the few counties (90 out of 2,881) in only 2 of 50 clusters where LRC estimates are positive but not significant at the 5% level. In other words, higher values of low-level indoor radon exposure are much more likely to be protective against lung cancer mortality than to possibly cause it.

We also see a wide range of sizes for 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 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 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 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.

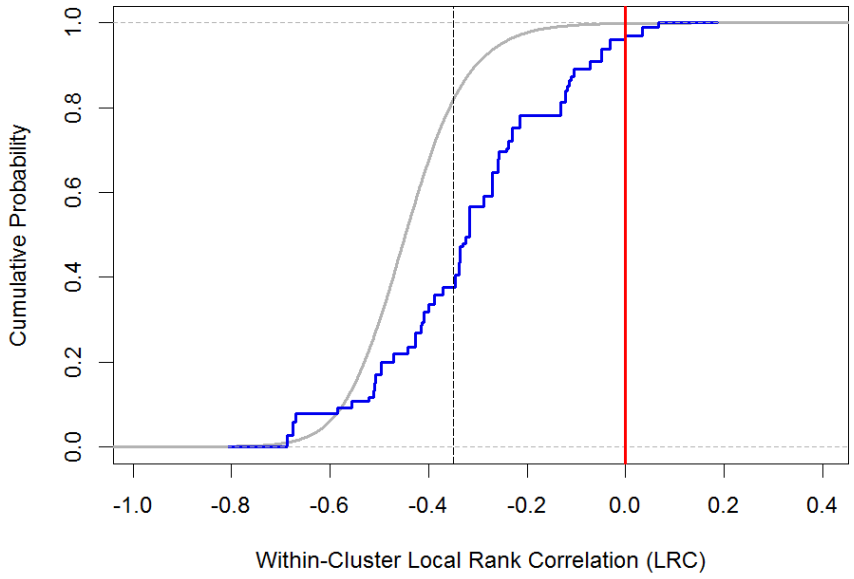


**Figure 6. LC Confirm Phase: Empirical CDF Comparison of the Observed LRC Distribution with its Random NULL Distribution from 1,000 replications.**

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 (Obenchain, 2019) 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 1,000 independent, random replications were then generated using the KSperm() function (Obenchain, 2019) 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 realistic and robust. All 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 particularly helpful in detecting and "displaying" interaction effects.

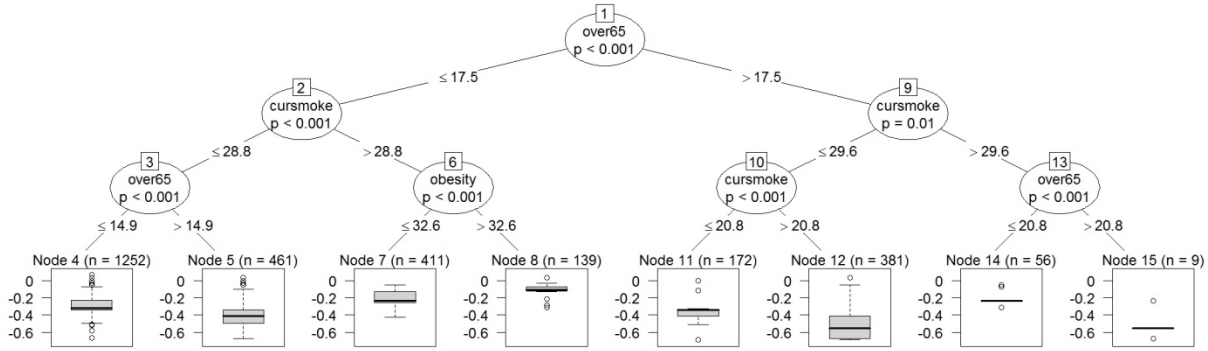


**Figure 7. party R-package tree model for predicting LRC estimates (supervised learning).**

A typical small-tree model, depicted in Figure 7, is based on (nonparametric) permutation theory using the party R-package (Hothorn, Hornik and Zeileis, 2006). 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 small-tree was restricted to have binary splits at 3 levels, yielding $2^3 = 8$ final "leaf" nodes.

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 LRCs. 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 2 summarizes these three major sub-groupings of LRC sub-distributions.

**Table 2. LC Reveal Phase comparison of LRC sub-distributions**

| Counties with LRC distributions less negative than typical | Counties with typical (mostly negative) LRC distributions | Counties with LRC distributions even more negative than typical |
|---|---|---|
| 606 (21.0%) | 1,252 (43.5%) | 1,023 (35.5%) |

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 conventional RP method (JMP®, 2016) 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" and "incomplete" trees can maximize overall goodness-of-fit ($R^2$) for any given total number of splits (seven here).

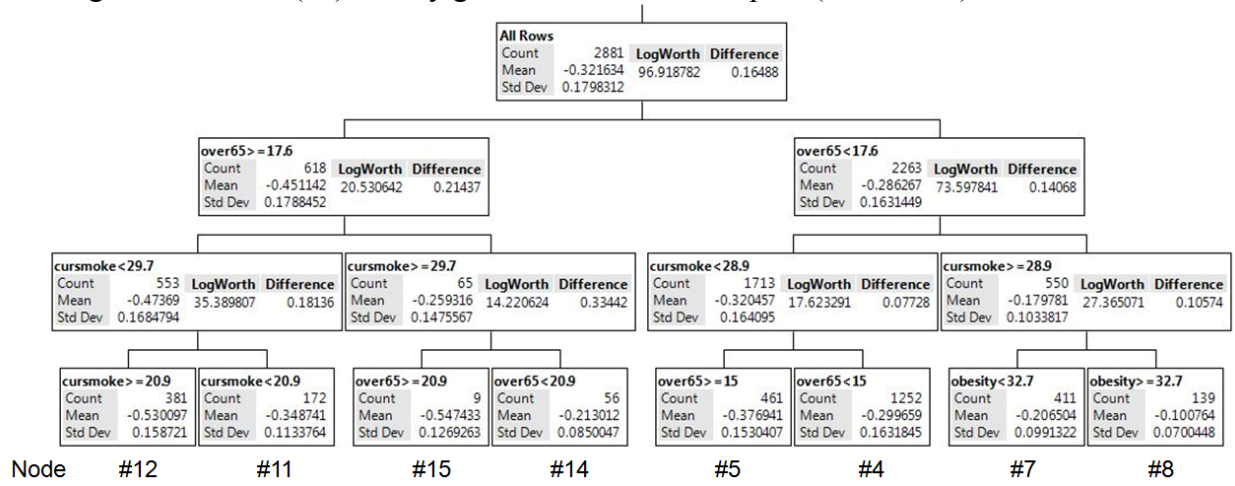


**Figure 8. SAS / JMP® representation of the party tree of Figure 7.**

On the other hand, we deliberately created Figure 8 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 six of the seven splits are indeed highly significant; the split of Node 10 on percentage of 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.

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 percentage of residents over 65 is used to make three of the seven splits depicted in Figures 7 and 8. However, the distinct surprise here may well be that LRCs are consistently predicted to be smaller (more negative) in counties where elderly residents are more prevalent.

Only one highly significant split on percentage of 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 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 1] was not selected for use in any of the seven most predictive splits. In fact, since our simple tree model failed to select any measure of radon exposure level as a predictor of LRCs, 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 LRCs 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 percentage of residents over 65, percentage of obese residents and percentage of 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 actually 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

Over the last 30 years, there has been little "true" consensus about the effects of low indoor radon exposure on mortality. Published studies have tended to either "embrace" or "question" the traditional Linear No Threshold (LNT) assumption for quantifying long-term "risk" from ionizing radiation. Several radiation researchers have noticed low-dose, nonlinear relationships, described as U- or J-shaped. A variety of names have been given to this phenomenon: “autoprotection, heteroprotection, adaptive response, preconditioning, hormesis, xenohormesis, paradoxical…” (Calabrese et al., 2017) Thus, the early observations of Cohen (1989-2008) appear to fit well into a much larger context whereby stress elicits protective effects (Parsons, 2002). In fact, Parsons asserts that “…hormesis for ionizing radiation becomes an evolutionary expectation at exposures substantially exceeding background.” Feinendegen (2015) 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. LC strategy does several things to support reliability: Covariates are controlled via clustering. The single question at issue is examined within each cluster. While the answer to research questions on local effects can be "they depend," LC strategy tends to uncover simple answers by down-playing the ill-effects of between-cluster "noise."

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 at least all "available" *x*-characteristics are indeed ignorable. Otherwise, rigid enforcement of the current EPA threshold for requiring radon mitigation could even increase expected lung cancer mortality in 697 of the 2,881 U.S. counties studied here; namely, counties with average radon exposure > 4 pCi/L that belong to clusters with negative LRC estimates.

Another unique feature of LC strategy is that within-cluster estimation of LRC (or LTD) distributions essentially moves the exposure (or treatment) variable to the left-hand-side of the supervised parametric, model equations commonly used for prediction of local, nonparametric effect-size estimates. This LC feature typically results in models with much better fit to LRCs (or LTDs) than models that attempt to predict *y*-outcomes for individual experimental units. Local fits (within Blocks) can be good even when global (overall) fits are poor or are frustrated by ill-conditioning (near multicollinearity.)

In statistical analysis there can be a tension between explaining and predicting. A paper in Statistical Science (Shmueli 2010) makes a distinction: "To explain or to predict". Local Control is firmly on the side of explaining. In our paper, we attempt to make two points. The first point is that our analysis process proceeds in simple steps so that both the researcher and the reader have a good sense of what is going on. The analysis, using simple steps, explains itself. Our second point is that, after controlling for other key variables via clustering, the overall inverse correlation between lung cancer and radon level persists ...albeit modified by covariates. A single tree can highlight multiple yet simple relationships. When one is seeking clarity of methods and explanations of results, the between-cluster noise suppression provided via LC strategy helps. LC methods aim at simple explanations by applying objective methods to data with subject-matter content. Approaches to multiple-tree recursive partitioning, such as random forests, can focus too exclusively on prediction. Given multiple trees, one can work backwards using statistics across the trees to get at explanations; variable importance can be inferred by how often a variable is used across trees, synergisms and correlations can be examine by looking at the co-occurrence of variables in trees. As our interest here is on simple explanations and we have made the data freely available, interested readers can explore multiple-tree analyses and predictions or whatever else they wish.

The negative LRC estimates observed here at relatively low levels of radon exposure agree with findings in a recent cohort study of Ontario uranium miners (Navaranjan et al., 2015). Cumulative exposure to radon for uranium miners is commonly expressed in units called Working Level Months (WLM). Our Figure 9 focuses on key information from Table 19 of the above cohort study, which maximizes Relative Risk estimates of lung cancer mortality by using

a 5-year lag. In particular, note that Figure 9 shows a hormetic "J-shaped" relationship that is "inverse" only at low levels of occupational exposure, i.e. at levels below 10 WLM.

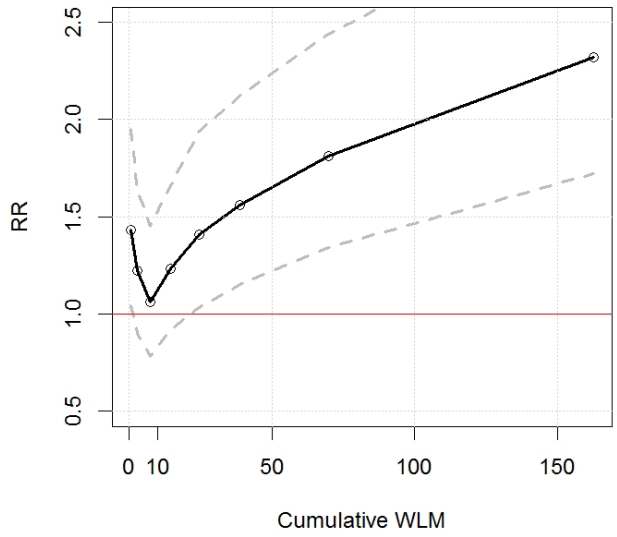


**Figure 9. Lung Cancer Mortality Relative Risk and 95% Confidence Limits from Table 19 of the Ontario Uranium Miner Cohort study.**

One WL of exposure to radon (gas) and its progeny (inhalable particles) is defined as exposure to 100 pCi/L of radon. One WLM of cumulative exposure occurs when a miner works 170 hours per Month at ~100 pCi/L of radon, assuming that mine ventilation is relatively poor. Of the 2,881 U.S. counties studied here, Teller County, CO, (FIPS = 8119) has by far the highest indoor radon: log(radon) = 4.60 in Figure 1, which is radon exposure rate of 99.7 pCi/L or 0.997 WL. (Figure 1 also shows that Teller has relatively low lung cancer mortality; 54.95 deaths per 100K person-years.)

Assuming approximately 7,000 hours spent indoors per year, one WLM/year of exposure to radon would be equivalent to 6.14 pCi/L or 227 Bq/m$^3$ (Health Physics Society, 2012). Our data indicate that the vast majority of U.S. counties show indoor radon levels of less than 36 pCi/L or 1,332 Bq/m$^3$ (6 WLM/year). The median exposure is 2.1 pCi/L or 78 Bq/m$^3$ (0.3 WLM/year), and the 95th percentile is 8.8 pCi/L or 326 Bq/m$^3$ (1.4 WLM/Year). Thus it seems clear that the indoor radon exposures for over 95% of U.S. counties fall well within the range of the "inverse relationship segment" (WLM < 10) of Figure 9.

Krstic (2017) reports a linear regression of data from 26 countries of the Organization for Economic Co-operation and Development (OECD), including North America. His analyses show a weak inverse (negative) correlation for age-standardized lung cancer mortality vs. mean indoor

radon concentration, compared with a significant positive correlation of lung cancer with smoking prevalence. His findings are thus consistent with the LC results presented here. Our LC analyses support the claim that lung cancer mortality decreases as low-level indoor radon exposures increase, with effect-sizes being largely predictable from local confounding characteristics like percentage of residents over 65, percentage of residents who currently smoke and percentage of obese residents.

## COI statements

The authors report no conflicts of interest.

## Grant information

All research work done in developing and writing this manuscript was performed by the authors without support. Work on development of LC methods and software by Obenchain and Young was partially funded by grants to Christophe G. Lambert, PI, University of New Mexico, from PCORI (CER-1507-31607) and NIH (1R21-LM012389). Views expressed in this paper represent those of the authors alone and not of their current or former organizations.

## References

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.

Calabrese EJ, Bachmann KA, Bailer AJ, *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*, 2007; 222: 122-128.

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

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

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

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

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.

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.

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

Health Physics Society (HPS). Answer to Question #10245 Submitted to "Ask the Experts", 2012. Available at: https://hps.org/publicinformation/ate/q10245.html (accessed May 2019).

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

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.

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.

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

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

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

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

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 July 2015).

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.

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.

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

Obenchain RL, Young SS. Advancing statistical thinking in observational health care research. *Journal of Statistical Theory and Practice*, 2013; 7: 456-469. DOI: 10.1080/15598608.2013.772821

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. DOI: 10.1111/risa.12749

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

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

Shmueli G. To Explain or to Predict? *Statistical Science,* 2010; 25, 289–310. (DOI: 10.1214/10)

U.S. Census Bureau - American Fact Finder: 2000 Census of Population and Housing. Accessed July 2015. https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml (pid=DEC_00_SF1_DP1)

U.S. Census Bureau. Small Area Income and Poverty Estimates. U.S. Department of Commerce, Accessed July 2015. https://www.census.gov/programs-surveys/saipe/data/datasets.html

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

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

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