BMJ Open

The validity of the average risk approach in estimating population attributable fractions for continuous exposures

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-045410
Article Type:	Original research
Date Submitted by the Author:	30-Sep-2020
Complete List of Authors:	Ruan, Yibing; Alberta Health Services, Cancer Epidemiology and Prevention Research Walter, Stephen ; McMaster University, Department of Health Research Methods, Evidence, and Impact Gogna, Priyanka ; Queen's University, Department of Public Health Sciences Friedenreich, CM; Alberta Health Services, Cancer Epidemiology and Prevention Research; University of Calgary Cumming School of Medicine, Departments of Oncology and Community Health Sciences Brenner, Darren; Alberta Health Services, Cancer Epidemiology and Prevention Research; University of Calgary Cumming School of Medicine, Departments of Oncology and Community Health Sciences
Keywords:	EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, PUBLIC HEALTH

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

The validity of the average risk approach in estimating population attributable fractions for continuous exposures

Yibing Ruan[1], Stephen D. Walter[2], Priyanka Gogna[3], Christine M. Friedenreich[1, 4], and Darren R. Brenner[1, 4]

1. Department of Cancer Epidemiology and Prevention Research, CancerControl Alberta, Alberta Health Services, Calgary, Alberta, Canada
2. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
3. Department of Public Health Sciences, Queen's University, Kingston, Ontario, Canada
4. Departments of Oncology and Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

*To whom correspondence should be addressed:

Yibing Ruan

yibing.ruan@albertahealthservices.ca

Abstract (298 words)

Main text (2,048 words)

ABSTRACT

Background The population attributable fraction (PAF) is an important metric for estimating disease burden associated with exposure to specific risk factors. In an International Agency for Research on Cancer (IARC) working group report published in 2007, an approach was introduced to estimate the PAF using the weighted average of a continuous exposure and the incremental relative risk (RR) per unit. This "average risk" approach has been subsequently applied in several other studies conducted worldwide. However, no investigation of the validity of this method has been done.

Objective To examine the validity and the potential magnitude of bias of the average risk approach.

Methods We established simulation models based on a variety of risk exposure distributions and a range of RR per unit. We estimated the unbiased PAF from integrating the exposure distribution and RR, and the PAF using the average risk approach. We examined the absolute and relative bias as the direct and relative difference in PAF estimated from the two approaches. We also examined the bias of the average risk approach using real-world data from the Canadian Population Attributable Risk of Cancer (ComPARe) study.

Results The average risk approach involves bias, the magnitude of which is affected by the exposure distribution as well as the value of the unit RR. While this approach is approximately valid when the RR per unit is small or the exposure distribution is symmetrical, its absolute and relative bias can both be large when RR per unit is not small and the exposure distribution is skewed. Under extreme situations, the average risk approach can underestimate PAF by an absolute bias of 17%.

Conclusions We recommend that caution be taken when using the average risk approach to estimate PAF, especially when the shape of the exposure distribution is either unknown or is skewed.

ARTICLE SUMMARY

Strengths and limitations of this study

- This study examined the assumptions and validity of the average risk approach to estimate the PAF, which has not been explored previously.
- We used both simulated and real-world data to demonstrate the factors associated with the bias of the average risk approach.
- As an empirical study, our simulation could only cover a limited number of risk exposure distributions.

INTRODUCTION

An important measure for estimating the burden of disease in a population is the Population Attributable Fraction (PAF) that combines data on the prevalence of an exposure along with the risk of disease associated with this exposure. The validity of PAF estimates is dependent, therefore, on accurate estimates of prevalence as well as population-specific estimates of risk. The International Agency for Research on Cancer has specialized in providing estimates of cancer surveillance and burden of cancer estimates from around the world. In 2007, Boffetta and colleagues,[1] introduced an approach to estimate the population attributable fraction (PAF) when the prevalence data on a continuous exposure in the population under study are only available as a weighted average. This approach, to be referred to here as the "average risk approach", estimated the RR at average exposure of the whole population using the risk of disease per unit increase in exposure, and the average level of exposure of the whole population. At the time that this method was proposed by Boffetta et al., no proof was provided. Hence, the purpose of this paper is to examine the underlying assumptions and validity of this average risk approach when estimating PAF for disease burden in a population. Specifically, we examined several distributions and permuted these distributions to generate a range of random distributions to investigate how the distribution of the exposure and the magnitude of the units of the RR influence the validity of this method.

METHODS

Description of Average Risk Approach

The average risk approach estimated the RR at an average exposure of the whole population using the RR of disease per unit increase in exposure along with the average level of exposure of the whole population as follows:

$$Risk = Exp^{[Ln(Risk\ per\ unit) \times average\ level\ of\ exposure]} = RR_{unit}^{\bar{x}} \qquad (1)$$

where *Risk* is the RR at the population average exposure, RR_{unit} is the RR associated with a per unit increase in exposure, $\bar{x}$ is the weighted average level of exposure. An underlying assumption with this method is that a log-linear relationship exists between the exposure and the risk of cancer. The average risk approach then estimates PAF as:

$$PAF = \frac{Risk - 1}{Risk} \qquad (2)$$

where it was assumed that "each individual has experienced a similar average exposure" (IARC 2007, pg 5). Under this assumption, that all population under study are exposed at the population average level, formula (2) is a simplification of Levin's formula when the prevalence (P) is 100%:

$$PAF = \frac{P(RR - 1)}{1 + P(RR - 1)} \qquad (3)$$

Boffetta et al. stated that "This formula is valid when the risk of cancer per unit of exposure was estimated in a model using log transformation. This is the case for logistic regression or Poisson regression, which are models widely used in case-control and cohort studies respectively" (IARC 2007, pg 5). No proof was shown for this statement, although the authors went on to

acknowledge that "the dose-effect relationship is, in fact, rarely linear (or log-linear) over the whole range of exposures, but this method is considered to be the best approximation available in this respect". Therefore, the validity of the average risk approach has not been fully assessed, particularly concerning its sensitivity to departures from the assumed dose-response relationship, or concerning the impact of the exposure distribution.

When the distribution of a continuous exposure is known, a completely valid method to estimate PAF involves integrating across all levels of exposure:

$$PAF = \frac{\int_{x=0}^{m} RR(x)P(x)dx - 1}{\int_{x=0}^{m} RR(x)P(x)dx} \qquad (4)$$

where RR(x) is the RR at exposure x; P(x) is the population distribution of exposure; and m is the maximum exposure level. Note that if there were to be no bias in the average risk approach, the following equation would have to hold:

$$\int_{x=0}^{m} RR(x)P(x)dx = RR_{unit}^{\bar{x}} \qquad (5)$$

Under log-linear risk assumption, equation (5) becomes:

$$\int_{x=0}^{m} RR_{unit}^{x}P(x)dx = RR_{unit}^{\bar{x}} \qquad (6)$$

The two sides of equation (6) are not guaranteed to equate. The integral on the left contains detailed information on the prevalence distribution, whereas on the right the distribution is condensed into a weighted average. The exposure distribution was not discussed by Boffetta et al., and this distribution affects the validity of their approach. Specifically, it is unknown if the exposure distribution in a population is strongly skewed or bimodal, whether the average risk

approach still provides a good approximation to the actual PAF. In addition, these authors did not explore whether or not the magnitude of RR per unit (RR_{unit}) itself would affect validity.

Investigation of Validity of Average Risk Approach

To investigate whether or not the validity of the average risk approach is affected by the exposure distribution, we simulated several exposure distributions where the exposure is continuous, ranging between standardized values of 0 to 100, with 0 indicating no exposure and 100 indicating the maximal level of exposure in the population (Figure 1). To simplify the calculations, we also assumed that rounding the exposure level to the nearest whole number values provides a sufficient approximation. The prevalence distributions were scaled so that the prevalence of all exposure levels summed to 100%. The details of the distributions are summarized in Table 1. We calculated PAF using both the average risk approach and by integrating across all exposure levels. We calculated the absolute bias ($PAF_{AvgRisk} - PAF_{Integral}$) and the relative bias $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

The choice of these distributions was intended to cover a variety of distributional patterns. To make our study more generalizable, we obtained random exposure distributions by permuting the distributions in Figure 1 10,000 times. The permutation was carried out by randomly re-sampling the exposure levels (0 to 100) with replacement. The re-sampled exposure levels were combined with the original prevalence distribution. This permutation resulted in a new distribution where the prevalence of each exposure level could change, while the sum of prevalence remains 100%. An example of new distributions after permutation is presented in Figure S1. We estimated the mean and standard deviation as well as the range of the absolute and relative bias from the 10,000 permutations.

To examine if the magnitude of risk affects the validity of the average risk approach, we tested a range of values for the RR per standardized unit, from 1.001 to 1.04. Using a standardized unit resolves the scaling issue of the unit. For example, the RR of standardized unit of body mass index and the disease associated with obesity is the same for the RR of 1 kg/m^2 or 5 kg/m^2, as long as it pertained to a single population. In this study, we refer to RR per standardized unit as "RR per unit", unless otherwise stated. We also considered that under our numerical framework, the risk becomes implausible for RR per unit values above 1.04. For example, the RR at maximal exposure level would be 132, if the RR per unit is 1.05 under the log-linear assumption.

Finally, we used real-world data of the distribution of air pollution ($PM_{2.5}$) and residential radon exposures, which were investigated in the Canadian Population Attributable Risk of Cancer (ComPARe) study. The ComPARe study collected national-representative and population-weighted exposure data of $PM_{2.5}$ and residential radon, and used the integral approach to estimate PAF of lung cancer for 2015 for Canada.[2, 3] We compared this PAF to that obtained using the average risk approach, to illustrate the validity of this approach.

Patient or public involvement

No patients involved.

RESULTS

First, we examined the bias of the average risk approach with the distributions we selected as in Table 1. The results at RR per unit of 1.01 and 1.03 were illustrated in Table 2 and the results with a range of RR per unit from 1.001 to 1.04 were shown in Figure 2. At RR per unit of 1.01, the absolute bias was small for all tested distributions (Table 2). The largest absolute bias was seen with the beta (0.5, 0.5) distribution, with higher prevalence at the lower and upper ends of the exposure distribution. When the distribution is symmetrical with a peak (e.g., normal, hypergeometric), the absolute bias was very small (-0.3% and -0.1%). In contrast, the relative bias was substantial with the Poisson and exponential distributions (-28% and -14%). Both the absolute and relative bias increased with larger RR per unit. However, the normal and hypergeometric distributions are more "resistant" than the Poisson and exponential distributions. For example, when the RR per unit increases to 1.03, the absolute and relative bias with the normal and hypergeometric distributions only increased slightly, while the absolute bias with the Poisson and exponential distributions increased to -17% and -18%, and the relative bias increased to -67% and -30%, respectively. For some distributions (uniform, beta (0.5, 0.5), beta (8, 2), and bimodal), the largest absolute and relative bias occurred at an intermediate value of RR per unit (Figure 2). As RR per unit increases, the bias becomes smaller, because the PAF estimates approaches 100%.

When examining the range of the bias of the average risk approach we permuted each distribution (Table 3), we found that when the RR per unit is small (1.002), both PAF differences and percent errors were very small and the average risk approach was valid. When RR per unit was increased to 1.01, the mean absolute and relative bias were larger in most examined distributions. The largest absolute and relative bias in all examined values of RR per unit were

observed in Poisson with and extreme tail, when it was as large as -41.9% and -54.3% when RR per unit is 1.04. Interestingly, we observed that regardless of the distribution and RR per unit, the average risk approach always underestimated PAF.

Finally, when exploring the bias of the average risk approach using real-world data for air pollution ($PM_{2.5}$) and residential radon, we found that neither distribution was normal (Figure S2). The $PM_{2.5}$ distribution had a long left tail, while the distribution of residential radon has a long right tail. We standardized the exposure levels of $PM_{2.5}$ and radon to 0.14 ug/m^3 and 7.4 Bq/m^3 per unit, so that the maximal exposure level is 100 units. The RR per unit of $PM_{2.5}$ associated with lung cancer was 1.0012. The PAFs of $PM_{2.5}$ using the integral and the average risk approach were 6.89% and 6.87%, respectively, indicating very small bias in the average risk approach. The RR per unit of radon associated with lung cancer was 1.011. The PAFs of radon using the integral and average risk approach were 6.87% and 6.37%, respectively. The bias was larger than that seen in $PM_{2.5}$. The absolute bias was -0.5% and the relative bias was -7.3%, indicating slight to moderate bias.

DISCUSSION

Since being introduced by Boffetta and colleagues in 2007, the average risk approach has been used in several PAF estimation projects.[4-8] In addition to the cancer burden study in France,[8] the ComPARe study in Canada,[6] a study of attributable causes in China,[7] and two studies in Brazil[4, 5] have used this method. We illustrated that the average risk approach always underestimates PAF under our simulated scenarios, implying that the direction of bias might be independent of the exposure distributions and the RR per unit. It nonetheless seems to

be sensitive to the log-linear risk assumption, because we observed that the average risk approach overestimates PAF under linear risk assumption (data not shown). When the RR per unit is small, the magnitude of bias is also small and the average risk approach is approximately valid. With larger RR per unit, the validity of the average risk approach would depend on the exposure distribution. We demonstrated that under some circumstances (e.g., Poisson distribution with extreme tail, exponential distribution), the approach could potentially lead to moderate to severe bias.

The limitations of our study need to be considered. To begin, this investigation was an empirical examination of the average risk approach for estimating PAF. We could not mathematically demonstrate the conditions when the average risk approach is valid. Second, we only studied a limited number of exposure distributions that we intentionally chose. Even with permutation, it was not possible to cover all distributions.

In conclusion, we have shown that the average risk approach has some utilities, nonetheless carries the risk of bias. We highly recommend using the integral approach when the exposure distribution data are available. Researchers should set out to gather the exposure distribution in a population before estimating PAF. When such information is unavailable, the average risk approach can be used if the RR per unit is small, or there is evidence that the exposure distribution is not highly skewed.

FOOTNOTES

Contributors: YR participated in study conceptualization, statistical analyses, drafted the initial manuscript and approved the final version of the manuscript. SW participated in study conceptualization, supervision, and critically reviewed and edited the manuscript. PG provided resources (ComPARe datasets), critically reviewed and edited the manuscript, and approved the final version of the manuscript. CF participated in funding acquisition, supervision, critically reviewed and edited the manuscript, and approved the final version of the manuscript. DB participated in funding acquisition, supervision, and approved the final version of the manuscript.

Funding: This study was supported by the Canadian Cancer Society Partner Prevention Research Grant (grant #703106).

Competing interests: None declared.

Patient consent for publication: Not required.

Ethics approval: Not required.

Data sharing statement: Extra data, including the R code for simulation and the exposure datasets from the ComPARe study, are available by emailing to yibing.ruan@albertahealthservices.ca.

Table 1. Description of the exposure distributions used in this study.

Distribution	**Note**
Uniform	Range from 0 to 100
Normal	μ=50, σ=10
Log-normal	μ=5, σ=0.5
Hypergeometric	N=700, K=200, m=200
Beta	α=0.5, β=0.5
Beta	α=2, β=8
Beta	α=8, β=2
Bimodal	Constructed by combining the lognormal distribution (μ=5, σ=0.5) with one-third of beta (8, 2).
Poisson with extreme tail	Constructed by applying the Poisson distribution (k = 0 to 3, λ=1) to exposure level 0 to 3, and one-tenth of the Poisson distribution (k = 70, 75, 80, 85, 90, λ=80) to exposure level 95 to 99
Exponential	Constructed by applying 1/x, where x = 1 to 100.

Note: All distributions were scaled to ensure that the sum of distribution is 100%.

Table 2. Absolute and relative bias in PAF between the average risk approach and the integration approach when RR per unit is 1.01 or 1.03

RRunit	**1.01**				**1.03**			
Distribution	$PAF_{Integral}$	$PAF_{Average\ risk}$	Absolute bias	Relative bias	$PAF_{Integral}$	$PAF_{Average\ risk}$	Absolute bias	Relative bias
Uniform	41.3%	38.9%	-2.4%	-5.9%	83.5%	76.8%	-6.7%	-8.0%
Normal	39.8%	39.5%	-0.3%	-0.8%	78.5%	77.5%	-1.0%	-1.2%
Log-normal	28.7%	27.0%	-1.6%	-5.6%	68.5%	60.8%	-7.7%	-11.3%
Hypergeometric	35.8%	35.7%	-0.1%	-0.2%	73.4%	73.1%	-0.3%	-0.4%
Beta(0.5, 0.5)	42.2%	38.9%	-3.3%	-7.9%	85.2%	76.8%	-8.4%	-9.8%
Beta(2, 8)	18.0%	17.4%	-0.6%	-3.5%	47.2%	43.3%	-3.9%	-8.3%
Beta(8, 2)	55.1%	54.8%	-0.3%	-0.6%	91.1%	90.5%	-0.5%	-0.6%
Bimodal	37.8%	35.3%	-2.6%	-6.7%	80.7%	72.5%	-8.2%	-10.2%
Poisson with extreme tail	4.1%	2.9%	-1.2%	-28.4%	25.8%	8.4%	-17.4%	-67.3%
Exponential	19.4%	16.6%	-2.8%	-14.3%	59.7%	41.7%	-17.9%	-30.0%

Note: the absolute bias is $PAF_{AvgRisk} - PAF_{Integral}$ and the relative bias is $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

Table 3. Mean, standard deviation, minimal, and maximal bias of the average risk approach observed from 10,000 permutations of the selected distributions under selected RR per unit

	Absolute bias				Relative bias			
RRunit = 1.002	Average	SD	Min	Max	Average	SD	Min	Max
Uniform	-0.2%	0.01%	-0.1%	-0.2%	-1.6%	0.17%	-1.0%	-2.4%
Normal	-0.1%	0.02%	-0.1%	-0.2%	-1.6%	0.29%	-0.7%	-2.8%
Log-normal	-0.2%	0.02%	-0.1%	-0.2%	-1.6%	0.21%	-0.9%	-2.7%
Hypergeometric	-0.1%	0.03%	0.0%	-0.3%	-1.5%	0.4%	-0.4%	-3.4%
Beta(0.5, 0.5)	-0.2%	0.02%	-0.1%	-0.2%	-1.6%	0.19%	-0.9%	-2.4%
Beta(2, 8)	-0.1%	0.02%	-0.1%	-0.2%	-1.6%	0.27%	-0.7%	-2.7%
Beta(8, 2)	-0.1%	0.02%	-0.1%	-0.2%	-1.6%	0.27%	-0.7%	-2.8%
Bimodal	-0.2%	0.01%	-0.1%	-0.2%	-1.6%	0.18%	-1.0%	-2.5%
Poisson with extreme tail	-0.1%	0.07%	0.0%	-0.4%	-1.2%	0.87%	0.0%	-7.1%
Exponential	-0.1%	0.03%	-0.1%	-0.3%	-1.5%	0.41%	-0.6%	-3.5%
RRunit = 1.01	Average	SD	Min	Max	Average	SD	Min	Max
Uniform	-2.5%	0.2%	-1.5%	-3.3%	-5.9%	0.7%	-3.5%	-8.8%
Normal	-2.4%	0.4%	-1.1%	-4.1%	-5.9%	1.1%	-2.4%	-10.7%
Log-normal	-2.5%	0.3%	-1.3%	-3.7%	-5.9%	0.8%	-2.8%	-9.6%
Hypergeometric	-2.4%	0.5%	-0.7%	-4.4%	-5.8%	1.6%	-1.6%	-12.9%
Beta(0.5, 0.5)	-2.5%	0.3%	-1.3%	-3.4%	-5.9%	0.8%	-3.2%	-9.2%
Beta(2, 8)	-2.4%	0.4%	-1.1%	-3.9%	-5.9%	1.1%	-2.4%	-10.8%
Beta(8, 2)	-2.4%	0.4%	-1.2%	-3.8%	-5.9%	1.1%	-2.7%	-10.8%
Bimodal	-2.5%	0.3%	-1.4%	-3.5%	-5.9%	0.7%	-3.3%	-8.8%
Poisson with extreme tail	-1.8%	1.2%	0.0%	-6.8%	-4.7%	3.4%	-0.1%	-27.1%
Exponential	-2.4%	0.5%	-1.0%	-4.4%	-5.8%	1.7%	-2.2%	-12.9%
RRunit = 1.04	Average	SD	Min	Max	Average	SD	Min	Max
Uniform	-6.2%	0.9%	-3.4%	-10.8%	-6.8%	1.1%	-3.6%	-12.2%
Normal	-6.2%	1.6%	-1.9%	-15.9%	-6.8%	1.8%	-2.0%	-18.1%
Log-normal	-6.2%	1.2%	-2.6%	-12.4%	-6.8%	1.3%	-2.7%	-13.7%

Hypergeometric	-6.2%	2.2%	-1.1%	-17.1%	-6.8%	2.6%	-1.1%	-20.4%
Beta(0.5, 0.5)	-6.2%	1.1%	-3.1%	-10.9%	-6.8%	1.2%	-3.3%	-12.1%
Beta(2, 8)	-6.3%	1.5%	-1.9%	-13.4%	-6.8%	1.7%	-2.0%	-15.5%
Beta(8, 2)	-6.2%	1.5%	-2.4%	-13.5%	-6.8%	1.7%	-2.5%	-15.0%
Bimodal	-6.2%	1.0%	-3.2%	-10.5%	-6.8%	1.1%	-3.4%	-11.6%
Poisson with extreme tail	-5.6%	4.7%	-0.1%	-35.3%	-6.6%	6.1%	-0.1%	-49.0%
Exponential	-6.3%	2.5%	-1.6%	-17.7%	-7.0%	2.8%	-1.7%	-20.0%

Note: the absolute bias is $PAF_{AvgRisk} - PAF_{Integral}$ and the relative bias is $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

Figure legends:

Figure 1: Probability density curves of selected distributions in this study.

Figure 2: The absolute and relative bias of the average risk approach under the selected distributions and a range of RR per unit. Both absolute and relative bias are presented as a percentage. The absolute bias is the difference in PAF percentage, and the relative bias is the difference in PAF over the PAF using integration and expressed as a percentage.

Figure S1: An example of the probability density curves of selected distributions after permutation

Figure S2: The distribution of $PM_{2.5}$ and residential radon in Canada

REFERENCES

1. IARC, *Attributable Causes of Cancer in France in the Year 2000*, in *IARC Working Group Report Volume 3*. 2007.

2. Gogna, P., et al., *Estimates of the current and future burden of lung cancer attributable to PM2.5 in Canada.* Prev Med, 2019. **122**: p. 91-99.

3. Gogna, P., et al., *Estimates of the current and future burden of lung cancer attributable to residential radon exposure in Canada.* Prev Med, 2019. **122**: p. 100-108.

4. Azevedo, E.S.G., et al., *The Fraction of Cancer Attributable to Ways of Life, Infections, Occupation, and Environmental Agents in Brazil in 2020.* PLoS One, 2016. **11**(2): p. e0148761.

5. Rezende, L.F. and J. Eluf-Neto, *Population attributable fraction: planning of diseases prevention actions in Brazil.* Rev Saude Publica, 2016. **50**.

6. Ruan, Y., et al., *Estimates of the current and future burden of cancer attributable to red and processed meat consumption in Canada.* Prev Med, 2019. **122**: p. 31-39.

7. Wang, J.B., et al., *Attributable causes of cancer in China.* Ann Oncol, 2012. **23**(11): p. 2983-2989.

8. Boffetta, P., et al., *The causes of cancer in France.* Ann Oncol, 2009. **20**(3): p. 550-5.

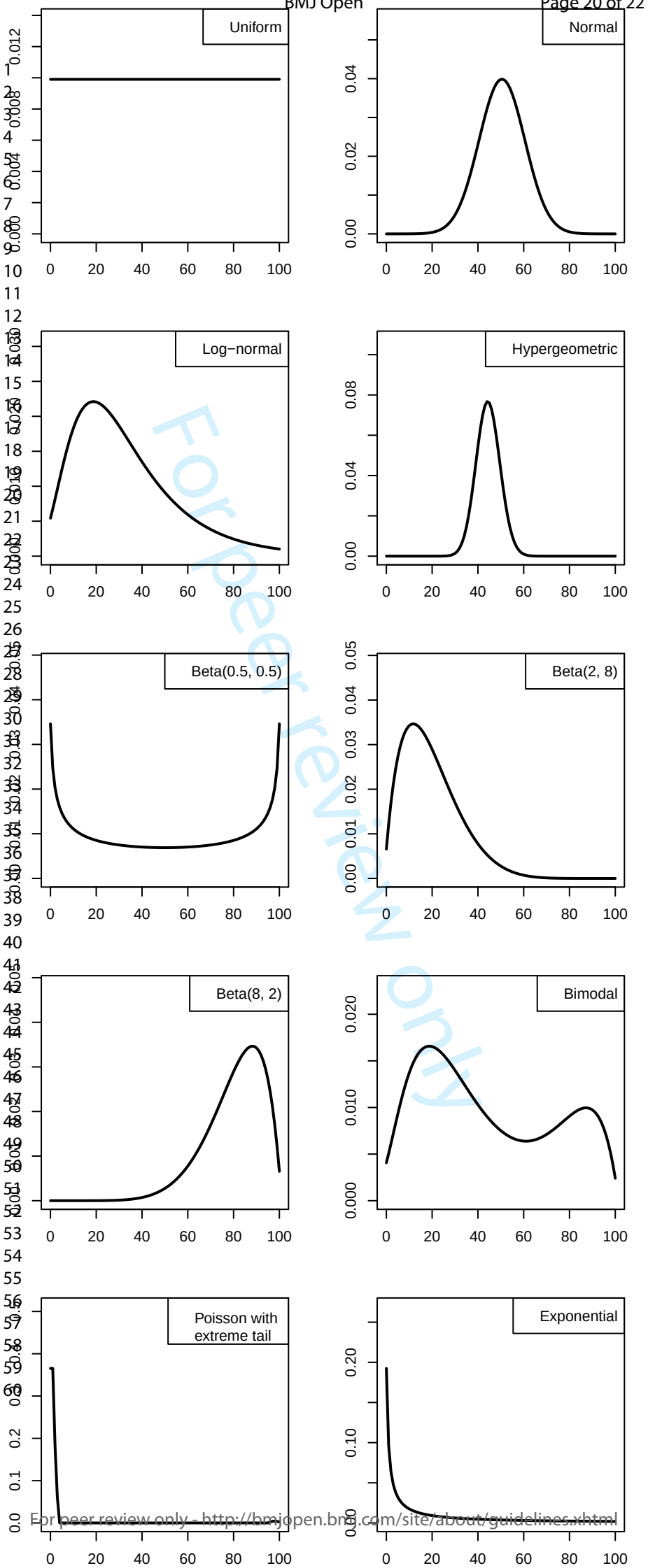
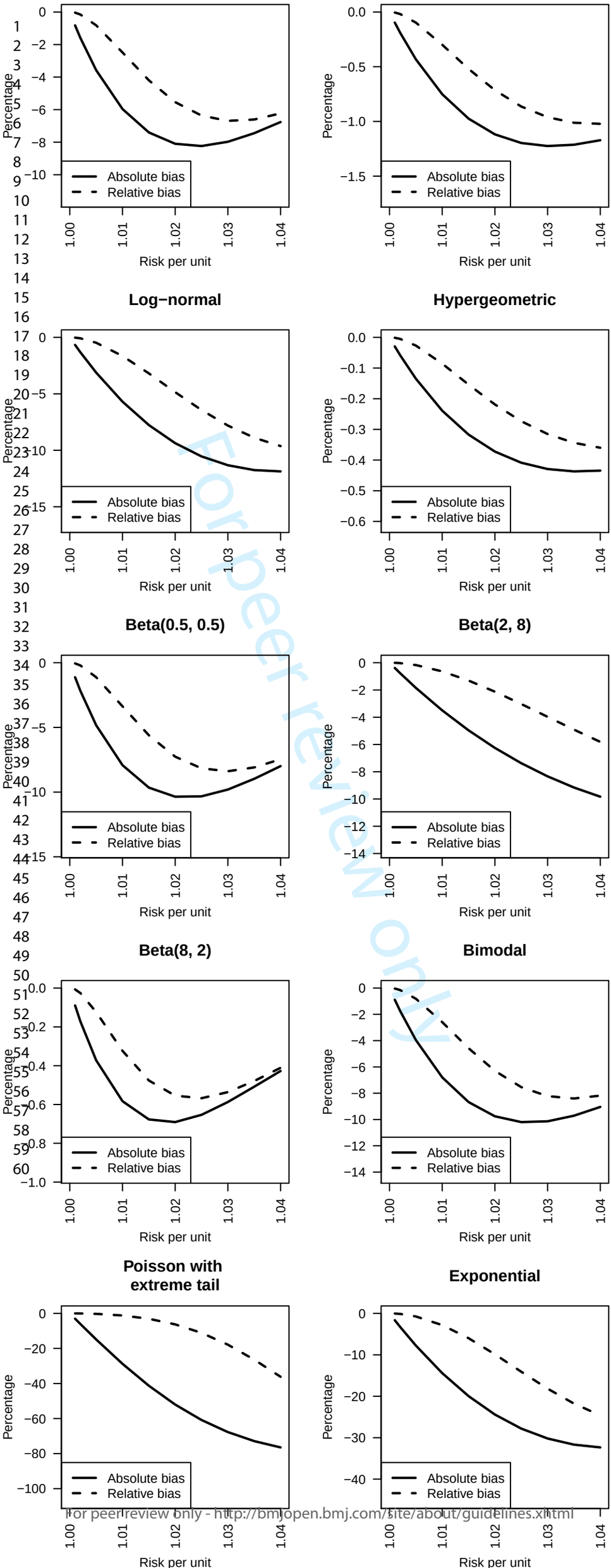
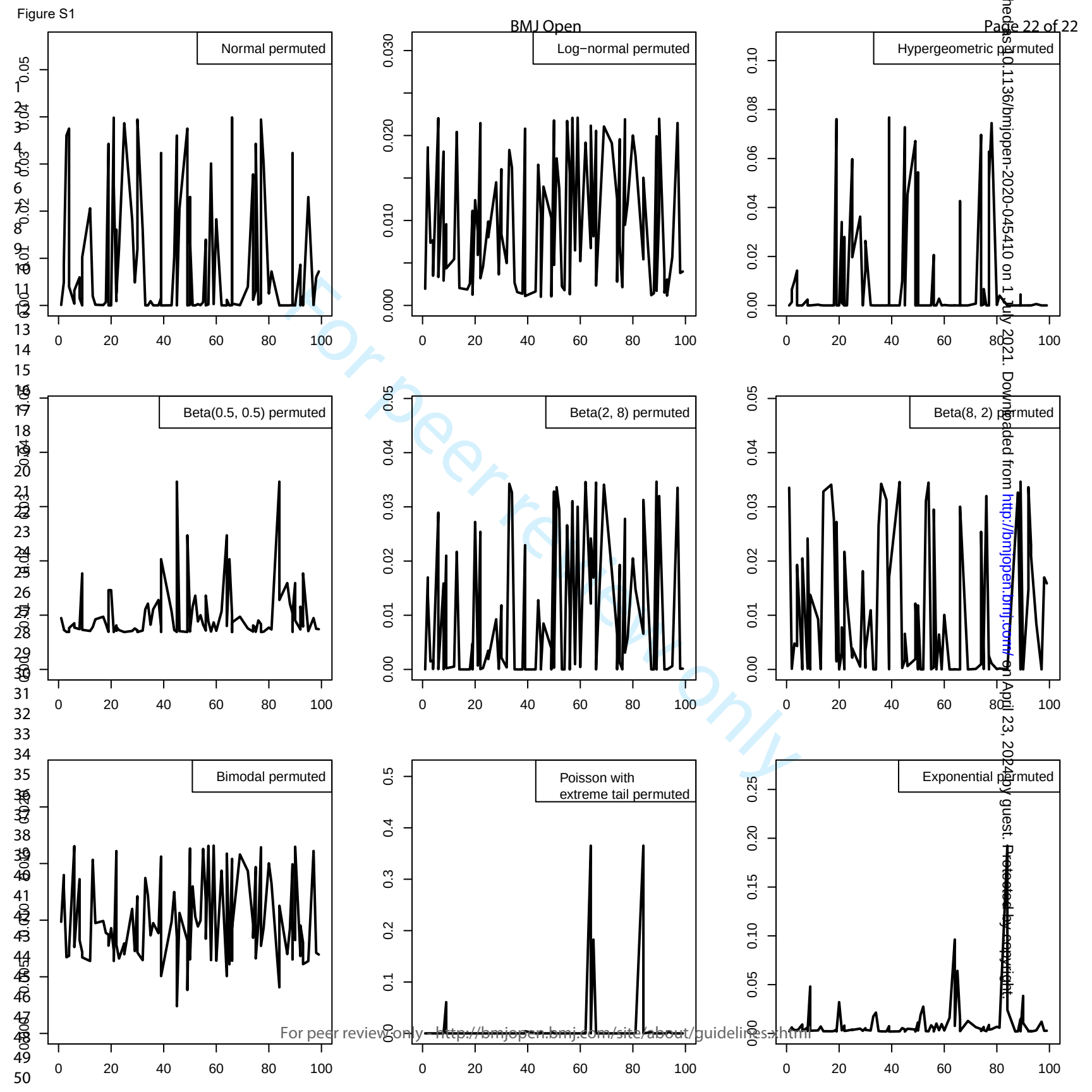
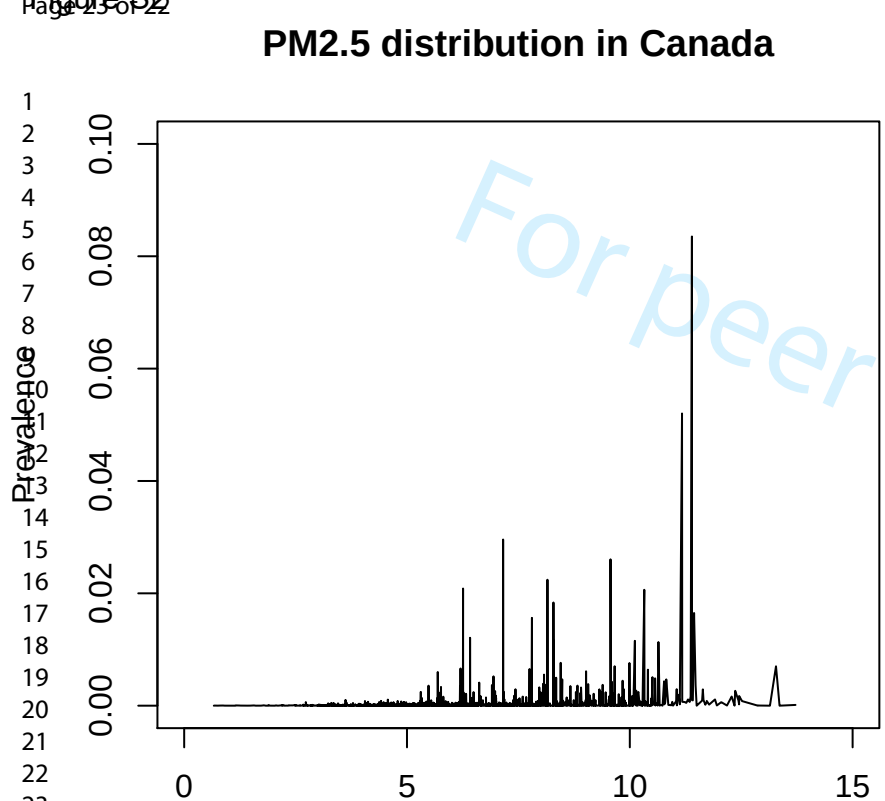
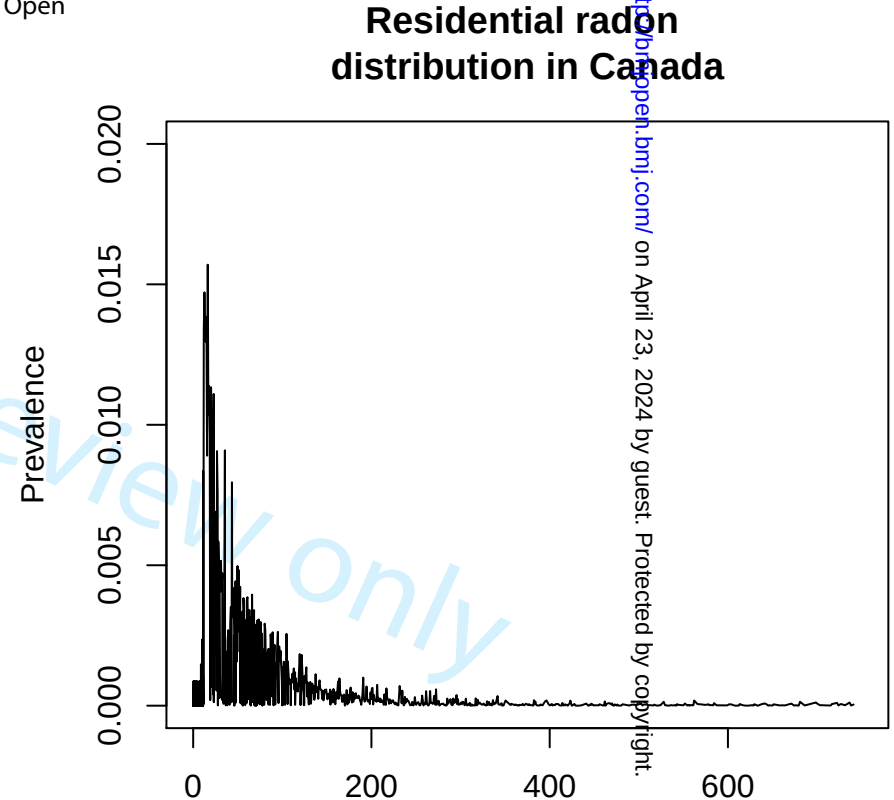


Figure 2









BMJ Open

A simulation study on the validity of the average risk approach in estimating population attributable fractions for continuous exposures

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-045410.R1
Article Type:	Original research
Date Submitted by the Author:	22-Mar-2021
Complete List of Authors:	Ruan, Yibing; Alberta Health Services, Cancer Epidemiology and Prevention Research Walter, Stephen ; McMaster University, Department of Health Research Methods, Evidence, and Impact Gogna, Priyanka ; Queen's University, Department of Public Health Sciences Friedenreich, CM; Alberta Health Services, Cancer Epidemiology and Prevention Research; University of Calgary Cumming School of Medicine, Departments of Oncology and Community Health Sciences Brenner, Darren; Alberta Health Services, Cancer Epidemiology and Prevention Research; University of Calgary Cumming School of Medicine, Departments of Oncology and Community Health Sciences
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Research methods
Keywords:	EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, PUBLIC HEALTH

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Original research: A simulation study on the validity of the average risk approach in estimating population attributable fractions for continuous exposures

Yibing Ruan[1], Stephen D. Walter[2], Priyanka Gogna[3], Christine M. Friedenreich[1,4], and Darren R. Brenner[1,4]

1. Department of Cancer Epidemiology and Prevention Research, Cancer Care Alberta, Alberta Health Services, Calgary, Alberta, Canada
2. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
3. Department of Public Health Sciences, Queen's University, Kingston, Ontario, Canada
4. Departments of Oncology and Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

*To whom correspondence should be addressed:

Yibing Ruan

Department of Cancer Epidemiology and Prevention Research

Cancer Care Alberta, Alberta Health Services

Holy Cross Centre - Room 513

2210-2nd ST. SW.

Calgary, AB, T2S 3C3

yibing.ruan@albertahealthservices.ca

Abstract (300 words)

Main text (2,731 words)

ABSTRACT

Background The population attributable fraction (PAF) is an important metric for estimating disease burden associated with causal risk factors. In an International Agency for Research on Cancer (IARC) working group report, an approach was introduced to estimate the PAF using the average of a continuous exposure and the incremental relative risk (RR) per unit. This "average risk" approach has been subsequently applied in several studies conducted worldwide. However, no investigation of the validity of this method has been done.

Objective To examine the validity and the potential magnitude of bias of the average risk approach.

Methods We established analytically that the direction of the bias is determined by the shape of the RR function. We then used simulation models based on a variety of risk exposure distributions and a range of RR per unit. We estimated the unbiased PAF from integrating the exposure distribution and RR, and the PAF using the average risk approach. We examined the absolute and relative bias as the direct and relative difference in PAF estimated from the two approaches. We also examined the bias of the average risk approach using real-world data from the Canadian Population Attributable Risk of Cancer study.

Results The average risk approach involves bias, which is under- or over-estimation with a convex or concave RR function (a risk profile that increases more/less rapidly at higher levels of exposure). The magnitude of the bias is affected by the exposure distribution as well as the value of RR. This approach is approximately valid when the RR per unit is small or the RR function is

approximately linear. The absolute and relative bias can both be large when RR is not small and the exposure distribution is skewed.

Conclusions We recommend that caution be taken when using the average risk approach to estimate PAF.

ARTICLE SUMMARY

Strengths and limitations of this study

- This study examined the assumptions and validity of the average risk approach to estimate the PAF, which has not been explored previously.
- We used both simulated and real-world data to demonstrate the factors associated with the bias of the average risk approach.
- As an empirical study, our simulation could only analytically establish the direction of bias of this approach and discuss the magnitude of bias using a limited number of risk exposure distributions and RR functions.

INTRODUCTION

Population Attributable Fraction (PAF) is an important measure for estimating the burden of disease in a population that is causally attributable to an exposure. Since its first introduction, PAF has received substantial attention in the field of epidemiology [1]. Many advances have been made in different approaches to calculating PAF of single and multiple risk factors [2-6], in estimating the variance [7, 8] and in the interpretation of PAF [9-11]. There have also been many comprehensive projects, either nationwide or globally, in estimating PAF for the burden of disease associated with its risk factors [12-22]. The International Agency for Research on Cancer has specialized in providing estimates of cancer surveillance and burden of cancer estimates from around the world. In 2007, Boffetta and colleagues [23] introduced an approach to estimating PAF when the prevalence data on a continuous exposure in the population under study are only available as a population average. This approach, to be referred to here as the "average risk approach", estimated the RR at average exposure of the whole population using the risk of disease per unit increase in exposure, and the average level of exposure of the whole population. No proof was provided at the time that this method was proposed. Hence, the purpose of this paper is to examine the underlying assumptions and validity of this average risk approach when estimating PAF for disease burden in a population. Specifically, we examined how the shape of the RR functions and the exposure distributions affect the validity of this approach.

METHODS

Description of Average Risk Approach

The average risk approach estimates the RR at an average exposure of the whole population using the RR of disease per unit increase in exposure along with the average level of exposure of the whole population as follows:

$$Risk = Exp^{[Ln(Risk\ per\ unit) \times average\ level\ of\ exposure]} = RR_{unit}^{\bar{x}} \qquad (1)$$

where *Risk* is the RR at the population average exposure, RR_{unit} is the RR associated with a unit increase in exposure, $\bar{x}$ is the weighted average level of exposure. An underlying assumption with this method is that a log-linear relationship exists between the exposure and the risk of cancer. The average risk approach then estimates PAF as:

$$PAF = \frac{Risk - 1}{Risk} \qquad (2)$$

where it was assumed that "each individual has experienced a similar average exposure" (IARC 2007, pg 5). Under this assumption, that all population under study are exposed at the population average level, formula (2) is a simplification of Levin's formula when the prevalence (*P*) is 100%:

$$PAF = \frac{P(RR - 1)}{1 + P(RR - 1)} \qquad (3)$$

Boffetta et al. stated that "This formula is valid when the risk of cancer per unit of exposure was estimated in a model using log transformation. This is the case for logistic regression or Poisson regression, which are models widely used in case-control and cohort studies respectively" (IARC 2007, pg 5). No proof was shown for this statement, although the authors went on to acknowledge that "the dose-effect relationship is, in fact, rarely linear (or log-linear) over the whole range of exposures, but this method is considered to be the best approximation available in this respect". Therefore, the validity of the average risk approach has not been fully assessed,

particularly concerning its sensitivity to departures from the assumed dose-response relationship, or concerning the impact of the exposure distribution.

When the distribution of a continuous exposure is known and no confounding is assumed, a valid method to estimate PAF involves integrating across all levels of exposure:

$$PAF = \frac{\int_{x=0}^{m} RR(x)P(x)dx - 1}{\int_{x=0}^{m} RR(x)P(x)dx} \quad (4)$$

where RR(x) is the RR at exposure x; P(x) is the population distribution of exposure; and m is the maximum exposure level. Note that if there were to be no bias in the average risk approach, the following equation would have to hold:

$$\int_{x=0}^{m} RR(x)P(x)dx = RR_{unit}^{\bar{x}} \quad (5)$$

Under the log-linear risk assumption, the left-hand side of equation (5) becomes:

$$\int_{x=0}^{m} RR_{unit}^{x}P(x)dx \quad (6)$$

Define $g(x) = RR_{unit}^{x}$ in which x is a random variable with distribution $P(x)$, (6) is E[$g(x)$], and the right-hand side of (5) is g[E(x)], because $g(x)$ is strictly convex (i.e., a line segment connecting any two points on the graph of a function lies above the graph) when RR_{unit} is greater than 1, the Jensen's inequality [24] determines that:

$$RR_{unit}^{\bar{x}} \leq \int_{x=0}^{m} RR_{unit}^{x}P(x)dx \quad (7)$$

According to (7), the average risk approach will not overestimate PAF. The magnitude of the bias is determined by the extent of the convexity of $g(x)$ over the effective range of x. When RR_{unit} is small (i.e., close to 1.00), $g(x)$ is approximately linear and there is little bias. However, whether or not the choice of the exposure distribution $P(x)$ affects the validity of this approach is unexplored. Specifically, it is unknown, if the exposure distribution in a population is strongly skewed or bimodal, whether or not the average risk approach still provides a good approximation to the actual PAF. Therefore, we studied the validity of the average risk approach under the loglinear RR function and a variety of exposure distributions.

In broad terms, when the loglinear function of RR is not assumed, the average risk approach can still be generalized as equation (2), in which "*Risk*" is the RR at the population average exposure level. It can be reasoned that the curvature of the RR function determines the direction and the magnitude of the bias. When RR is a linear function of the exposure (i.e., $RR(x) = 1 + k \cdot x, x \in [0,m]$), there is no bias, because the integral PAF ($\int_{x=0}^{m}(1+kx)P(x)dx$) and the average risk PAF ($1 + k\int_{x=0}^{m}xP(x)dx$) are equivalent. When the RR function has a convex form, which indicates a risk profile that increases more rapidly at higher levels of exposure, this approach underestimates PAF. In contrast, it overestimates PAF with a concave RR function, which indicates a risk profile that increases less rapidly at higher levels of exposure. To illustrate the latter point, we included two examples of simulated concave RR functions and calculated the bias of the average risk approach.

Investigation of Validity of Average Risk Approach

To investigate whether or not the validity of the average risk approach is affected by the exposure distribution, we simulated several exposure distributions where the exposure is

continuous, ranging between standardized values of 0 to 100, with 0 indicating no exposure and 100 indicating the maximal level of exposure in the population (Figure 1). The prevalence distributions were scaled so that the prevalence of all exposure levels summed to 100%. The details of the distributions are summarized in Table 1. We calculated PAF using both the average risk approach and by integrating across all exposure levels. We calculated the absolute bias ($PAF_{AvgRisk} - PAF_{Integral}$) and the relative bias $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

To examine if the magnitude of risk affects the validity of the average risk approach, we tested a range of values for the RR per standardized unit, from 1.001 to 1.04. Using a standardized unit resolves the scaling issue of the unit. For example, the RR of standardized unit of body mass index and the disease associated with obesity is the same for the RR of 1 kg/m^2 or 5 kg/m^2, as long as it pertained to a single population. In this study, we refer to RR per standardized unit as "RR per unit", unless otherwise stated. We also considered that the risk becomes implausible for RR per unit values above 1.04. For example, the RR at maximal exposure level would be 132, if the RR per unit is 1.05 under the log-linear assumption.

In addition, we illustrated the bias of the average risk approach when the RR function is non-linear or loglinear. In particular, we used two simulated examples of quadratic and cubic spline RR functions, which are both concave (Figure S1). The quadratic RR function has a form of $RR(x) = 3 - 2(\frac{x}{k} - 1)^2$, in which $k \in [\frac{m}{2}, m]$. This quadratic form has RR=1 when x=0, and RR has a maximum of 5 when $x = k$. In the illustrated example, we used k = 75, i.e., 75% of the maximal exposure. The cubic spline RR function is based on simulated data, with the function

being approximately quadratic in the lower exposure range, and approximately linear at higher exposures.

Finally, we used real-world data of the distribution of air pollution ($PM_{2.5}$) and residential radon exposures, which were investigated in the Canadian Population Attributable Risk of Cancer (ComPARe) study. The ComPARe study collected national-representative and population-weighted exposure data of $PM_{2.5}$ and residential radon and used the integral approach to estimate PAF of lung cancer for 2015 for Canada.[25, 26] We compared this PAF to that obtained using the average risk approach, to illustrate the validity of this approach.

Patient or public involvement

No patients involved.

RESULTS

First, we examined the bias of the average risk approach under the loglinear RR function with the exposure distributions we selected in Table 1. The results at RR per unit of 1.001, 1.01 and 1.03 were illustrated in Table 2 and the results with a range of RR per unit from 1.001 to 1.04 were shown in Figure 2. At RR of 1.001, the absolute and relative biases were very small and the average risk approach can be regarded unbiased. At RR of 1.01, the absolute bias remained small for all tested distributions although the relative bias started to increase substantially in the power distribution and in the Poisson distribution with an extreme tail (Table 2). At RR of 1.03, large absolute and relative biases were observed in several distributions. However, the normal and hypergeometric distributions were more robust than the Poisson with extreme tail and power distributions with the increase in RR (Table 2, Figure 2). For some

distributions (uniform, beta (0.5, 0.5), beta (8, 2), and bimodal), the largest absolute and relative bias occurred at an intermediate value of RR (Figure 2). As RR increases, the bias becomes smaller, because the PAF estimates approaches 100%. Regardless of the exposure distribution and the magnitude of RR, the direction of the bias is underestimation in the case of loglinear RR.

We then illustrated the direction of the bias when the RR function is concave. Table 3 showed the resulting bias of the two RR functions in Figure S1 when the exposure distributions were as reported in Table 2. With concave RR functions, the direction of the bias in the average risk approach is overestimation. Similar to the loglinear RR function, we observed little bias in normal, hypergeometric, and beta(8, 2) distributions, whereas substantial bias was observed in power, Poisson with extreme tail, and beta(0.5, 0.5) distributions.

Finally, we explored the bias of the average risk approach using real-world data for air pollution ($PM_{2.5}$) and residential radon. Epidemiologic studies support a loglinear RR function between exposure to residential radon and lung cancer [27, 28]. A loglinear dose response between PM2.5 and lung cancer risk was less consistent. The loglinear relationship was supported by several studies [29-32], while two studies reported some deviation from it [33, 34]. The 2019 Global Burden of Disease Study of 87 risk factors suggested that PM2.5 has a loglinear relation with lung cancer in low exposure range (0-50 ug/m3) and a linear relation in high exposure range (>50 ug/m3) {Collaborators, 2020 #20}. We assumed a loglinear relation for PM2.5 because the level is typically below 20 ug/m3 in Canada. We found that both exposures had skewed distributions (Figure S2). The $PM_{2.5}$ distribution had a long left tail, while the distribution of residential radon has a long right tail. We standardized the exposure levels of $PM_{2.5}$ and radon to 0.14 ug/m^3 and 7.4 Bq/m^3 per unit, so that the maximal exposure level is 100 units. The RR per unit of $PM_{2.5}$ associated with lung cancer was 1.0012. The PAFs of

$PM_{2.5}$ using the integral and the average risk approach were 6.89% and 6.87%, respectively, indicating very small bias in the average risk approach. The RR per unit of radon associated with lung cancer was 1.011. The PAFs of radon using the integral and average risk approach were 6.87% and 6.37%, respectively. The bias was larger than that seen in $PM_{2.5}$. The absolute bias was -0.5% and the relative bias was -7.3%, indicating slight to moderate bias. The observations were consistent with the simulations, in that small RRs yield little bias ($PM_{2.5}$), and moderate to large RRs could produce bias with some skewed exposure distributions (radon).

DISCUSSION

Since being introduced by Boffetta and colleagues in 2007, the average risk approach has been used in several PAF estimation projects.[12-15, 35] In addition to the cancer burden study in France,[15] the ComPARe study in Canada,[35] a study of attributable causes in China,[12] and two studies in Brazil[13, 14] have used this method. We illustrated that the direction of bias of the average risk approach is determined by whether the RR function is convex or concave, while the magnitude of bias is affected by the degree of convexity or concavity, as well as the exposure distribution. When the RR per unit is small under a loglinear RR function, the magnitude of bias is also small and the average risk approach is approximately valid. With larger RR and increased convexity, the validity of the average risk approach would also depend on the exposure distribution. We demonstrated that under some circumstances (e.g., Poisson distribution with extreme tail, power distribution), the approach could potentially lead to moderate to severe bias.

The average risk approach has an implicit assumption that the minimal risk exposure value is 0. When the minimal risk exposure value is not 0, this approach generates invalid estimates. To offer a simplified example, overweight and obesity defined as body mass index (BMI) $\geq$ 25.0 kg/m^2 is associated with postmenopausal breast cancer. The minimal risk exposure value of BMI is 25.0 kg/m^2. Assuming a log-linear relationship between BMI above 25.0 kg/m^2 and the risk of breast cancer and that a postmenopausal female population has a normal distribution of body mass index (BMI) at a mean and standard deviation of 25.0 and 5.0 kg/m^2. The average risk approach yields a PAF of 0 in this population, because the population average risk exposure is 25.0 kg/m^2, which has a RR of 1.0. Although it is possible to recode the exposure so that the minimal exposure is zero, a new average of the recoded exposure must be estimated, which requires the information of the exposure distribution. On the other hand, the prerequisite of applying the average risk approach is that such information is only available as a population average. In practice, the minimal risk exposure level of many natural or physiological exposures is not 0. Therefore, this implicit assumption is a substantial limitation of this approach. For the same reason, the average risk approach cannot be applied in the framework of generalized impact fraction, in which the impact of partial reduction of exposure is considered.

Our study has some limitations. First, this study is an empirical examination of the validity of the average risk approach. We have mathematically demonstrated the direction of the bias in this approach. However, we only qualitatively discussed the magnitude of the bias associated with the RR function and the exposure distribution. We illustrated the magnitude of the bias through several RR functions and exposure distributions. However, this pragmatic approach could not cover all RR functions and distributions. Second, we compared the average risk approach to the integral approach under the assumption of no confounding. The integral

approach is an extension of Levin's formula, which is biased in the presence of confounding [1, 11]. Ideally, the validity of the average risk approach should be tested against the integral form of Miettinen's formula, which is based on the prevalence of exposure among the cases and is valid in the presence of confounding [6]. However, because the average risk approach was developed under the framework of Levin's formula, we considered that a comparison of two approaches under the same framework would be more appropriate. Nevertheless, it should be noted that the validity of the average risk approach is also prone to the presence of confounding, just like Levin's formula.

In conclusion, we have shown that the average risk approach has some utility, but nonetheless carries the risk of bias. We highly recommend using alternative approaches when the RR per unit is not small in the range of the exposure, the RR functions depart from linear, or the exposure distribution data are available. The average risk approach can be used if the RR per unit is small, or there is evidence that the exposure distribution is not highly skewed. Nevertheless, researchers using this approach should discuss the direction of the bias based on the RR functions.

FOOTNOTES

Contributors: YR participated in study conceptualization, statistical analyses, drafted the initial manuscript and approved the final version of the manuscript. SW participated in study conceptualization, supervision, and critically reviewed and edited the manuscript. PG provided resources (ComPARe datasets), critically reviewed and edited the manuscript, and approved the final version of the manuscript. CF participated in funding acquisition, supervision, critically reviewed and edited the manuscript, and approved the final version of the manuscript. DB participated in funding acquisition, supervision, and approved the final version of the manuscript.

Funding: This study was supported by the Canadian Cancer Society Partner Prevention Research Grant (grant #703106).

Competing interests: None declared.

Patient consent for publication: Not required.

Ethics approval: Not required.

Data sharing statement: Extra data, including the R code for simulation and the exposure datasets from the ComPARe study, are available by emailing to yibing.ruan@albertahealthservices.ca.

Table 1. Description of the exposure distributions used in this study.

Distribution	**Note**
Uniform	Range from 0 to 100
Normal	$\mu=50$, $\sigma=10$
Log-normal	$\mu=5$, $\sigma=0.5$
Hypergeometric	N=700, K=200, m=200
Beta	$\alpha=0.5$, $\beta=0.5$
Beta	$\alpha=2$, $\beta=8$
Beta	$\alpha=8$, $\beta=2$
Bimodal	Constructed by combining the lognormal distribution ($\mu=5$, $\sigma=0.5$) with one-third of beta (8, 2).
Poisson with extreme tail	Constructed by applying the Poisson distribution (k = 0 to 3, $\lambda=1$) to exposure level 0 to 3, and one-tenth of the Poisson distribution (k = 70, 75, 80, 85, 90, $\lambda=80$) to exposure level 95 to 99
Power	Constructed by rescaling the function of 1/x, where $x \in [0.1, 2.5]$.

Note: All distributions were scaled to ensure that the sum of distribution is 100%.

Table 2. Absolute and relative bias in PAF between the average risk approach and the integration approach in selected exposure distributions when RR per unit is 1.001, 1.01 or 1.03 for the loglinear function

RRunit	1.001				1.01				1.03			
Distribution	PAF_{Integr}	PAF_{Avera}	Absolute bias	Relative bias	PAF_{Integr}	PAF_{Avera}	Absolute bias	Relative bias	PAF_{Integr}	PAF_{Avera}	Absolute bias	Relative bias
Uniform	4.9%	4.8%	0%	-0.9%	41.4%	38.9%	-2.6%	-6.2%	83.8%	76.8%	-7%	-8.3%
Normal	4.8%	4.8%	0%	-0.1%	38.9%	38.6%	-0.3%	-0.8%	77.6%	76.6%	-1%	-1.3%
Log-normal	3.1%	3.0%	0%	-0.7%	28.3%	26.5%	-1.7%	-6.1%	68.3%	60.0%	-8.4%	-12.3%
Hypergeometric	4.3%	4.3%	0%	0%	35.3%	35.3%	0%	0%	72.6%	72.6%	0%	0%
Beta(0.5, 0.5)	4.9%	4.8%	-0.1%	-1.1%	42.3%	38.9%	-3.4%	-7.9%	85.3%	76.9%	-8.5%	-9.9%
Beta(2, 8)	1.8%	1.8%	0%	-0.4%	17.2%	16.5%	-0.7%	-3.8%	45.7%	41.5%	-4.2%	-9.2%
Beta(8, 2)	7.8%	7.8%	0%	-0.1%	55.7%	55.3%	-0.3%	-0.6%	91.4%	90.9%	-0.5%	-0.6%
Bimodal	4.3%	4.3%	0%	-0.9%	37.9%	35.1%	-2.7%	-7.2%	81.1%	72.3%	-8.7%	-10.8%
Poisson with extreme tail	0.6%	0.6%	0%	-3.4%	8.3%	5.9%	-2.4%	-29.3%	42.7%	16.5%	-26.2%	-61.4%
Power	2.6%	2.6%	0%	-1.4%	25.9%	22.8%	-3%	-11.6%	69.0%	53.7%	-15.2%	-22.1%

Note: the absolute bias is $PAF_{AvgRisk} - PAF_{Integral}$ and the relative bias is $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

Table 3. Absolute and relative bias in PAF between the average risk approach and the integration approach in two illustrated examples of concave RR functions.

RR funtion	Cubic spline				Quadratic			
Distribution	$PAF_{Integral}$	$PAF_{Average\ risk}$	Absolute bias	Relative bias	$PAF_{Integral}$	$PAF_{Average\ risk}$	Absolute bias	Relative bias
Uniform	49.0%	52.8%	3.8%	7.8%	59.6%	64.0%	4.4%	7.4%
Normal	52.6%	52.8%	0.1%	0.3%	63.6%	64.1%	0.5%	0.7%
Log-normal	46.4%	51.3%	4.9%	10.6%	54.3%	57.3%	3%	5.6%
Hypergeometric	52.6%	52.7%	0.1%	0.1%	62.4%	62.5%	0.1%	0.2%
Beta(0.5, 0.5)	46.9%	52.8%	5.8%	12.4%	57.7%	64.0%	6.3%	11%
Beta(2, 8)	40.9%	43.9%	3.1%	7.6%	45.9%	47.4%	1.5%	3.3%
Beta(8, 2)	53.1%	53.1%	0%	0%	65.9%	66.5%	0.6%	0.9%
Bimodal	48.3%	52.7%	4.4%	9.2%	57.9%	62.4%	4.5%	7.8%
Poisson with extreme tail	6.1%	11.1%	5%	81%	8.5%	13.6%	5.1%	60.6%
Power	38.7%	49.1%	10.4%	26.9%	47.1%	53.4%	6.4%	13.5%

Note: the absolute bias is $PAF_{AvgRisk} - PAF_{Integral}$ and the relative bias is $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

Figure legends:

Figure 1: Probability density curves of selected distributions in this study.

Figure 2: The absolute and relative bias of the average risk approach under the selected distributions and a range of RR per unit. Both absolute and relative bias are presented as a percentage. The absolute bias is the difference in PAF percentage, and the relative bias is the difference in PAF over the PAF using integration and expressed as a percentage.

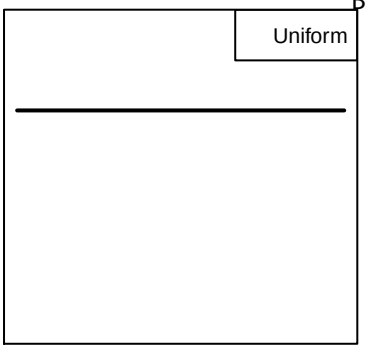
Figure S1: Graph of the two concave RR functions used in this study to illustrate the direction and the magnitude of bias of the average risk approach

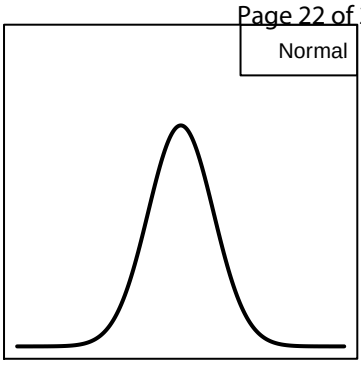
Figure S2: The smoothed density plot of the distributions of $PM_{2.5}$ and residential radon in Canada

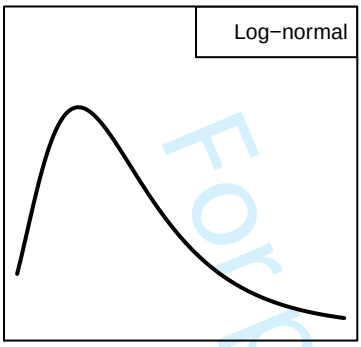
REFERENCES

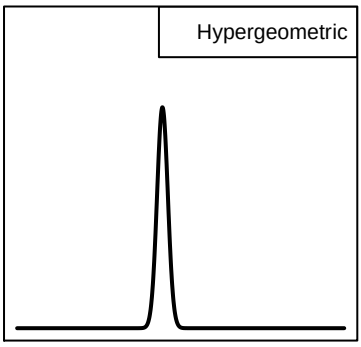
1. Levin, M.L., *The occurrence of lung cancer in man.* Acta Unio Int Contra Cancrum, 1953. **9**(3): p. 531-41.
2. Bruzzi, P., et al., *Estimating the population attributable risk for multiple risk factors using case-control data.* Am J Epidemiol, 1985. **122**(5): p. 904-14.
3. Eide, G.E. and I. Heuch, *Average attributable fractions: a coherent theory for apportioning excess risk to individual risk factors and subpopulations.* Biom J, 2006. **48**(5): p. 820-37.
4. Walter, S.D., *The estimation and interpretation of attributable risk in health research.* Biometrics, 1976. **32**(4): p. 829-49.
5. Whittemore, A.S., *Statistical methods for estimating attributable risk from retrospective data.* Stat Med, 1982. **1**(3): p. 229-43.
6. Miettinen, O.S., *Proportion of disease caused or prevented by a given exposure, trait or intervention.* Am J Epidemiol, 1974. **99**(5): p. 325-32.
7. Benichou, J. and M.H. Gail, *Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models.* Biometrics, 1990. **46**(4): p. 991-1003.
8. Greenland, S., *Variance estimators for attributable fraction estimates consistent in both large strata and sparse data.* Stat Med, 1987. **6**(6): p. 701-8.
9. Mansournia, M.A. and D.G. Altman, *Population attributable fraction.* BMJ, 2018. **360**: p. k757.
10. Di Maso, M., et al., *Attributable fraction for multiple risk factors: Methods, interpretations, and examples.* Stat Methods Med Res, 2020. **29**(3): p. 854-865.
11. Rockhill, B., B. Newman, and C. Weinberg, *Use and misuse of population attributable fractions.* Am J Public Health, 1998. **88**(1): p. 15-9.
12. Wang, J.B., et al., *Attributable causes of cancer in China.* Ann Oncol, 2012. **23**(11): p. 2983-2989.
13. Azevedo, E.S.G., et al., *The Fraction of Cancer Attributable to Ways of Life, Infections, Occupation, and Environmental Agents in Brazil in 2020.* PLoS One, 2016. **11**(2): p. e0148761.
14. Rezende, L.F. and J. Eluf-Neto, *Population attributable fraction: planning of diseases prevention actions in Brazil.* Rev Saude Publica, 2016. **50**.
15. Boffetta, P., et al., *The causes of cancer in France.* Ann Oncol, 2009. **20**(3): p. 550-5.
16. Collaborators, G.B.D.R.F., *Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet, 2020. **396**(10258): p. 1223-1249.
17. Parkin, D.M., L. Boyd, and L.C. Walker, *16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010.* Br J Cancer, 2011. **105 Suppl 2**: p. S77-81.
18. Whiteman, D.C., et al., *Cancers in Australia in 2010 attributable to modifiable factors: summary and conclusions.* Aust N Z J Public Health, 2015. **39**(5): p. 477-84.
19. Diseases, G.B.D. and C. Injuries, *Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet, 2020. **396**(10258): p. 1204-1222.
20. Poirier, A.E., et al., *The current and future burden of cancer attributable to modifiable risk factors in Canada: Summary of results.* Prev Med, 2019. **122**: p. 140-147.
21. Arnold, M., et al., *Global burden of cancer attributable to high body-mass index in 2012: a population-based study.* Lancet Oncol, 2015. **16**(1): p. 36-46.
22. Islami, F., et al., *Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States.* CA Cancer J Clin, 2018. **68**(1): p. 31-54.

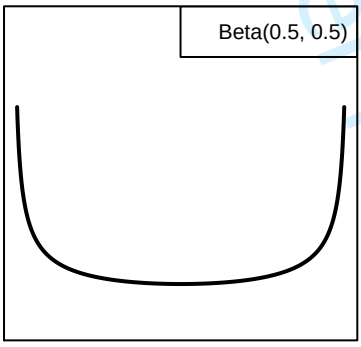
23. IARC, *Attributable Causes of Cancer in France in the Year 2000*, in *IARC Working Group Report Volume 3*. 2007.
24. Durrett, R., *Probability: Theory and Examples*. 5th ed. ed. 2019: Cambridge University Press.
25. Gogna, P., et al., *Estimates of the current and future burden of lung cancer attributable to PM2.5 in Canada.* Prev Med, 2019. **122**: p. 91-99.
26. Gogna, P., et al., *Estimates of the current and future burden of lung cancer attributable to residential radon exposure in Canada.* Prev Med, 2019. **122**: p. 100-108.
27. Krewski, D., et al., *A combined analysis of North American case-control studies of residential radon and lung cancer.* J Toxicol Environ Health A, 2006. **69**(7): p. 533-97.
28. Puskin, J.S., *Perspective on the use of LNT for radiation protection and risk assessment by the U.S. Environmental Protection Agency.* Dose Response, 2009. **7**(4): p. 284-91.
29. Hystad, P., et al., *Long-term residential exposure to air pollution and lung cancer risk.* Epidemiology, 2013. **24**(5): p. 762-72.
30. Lepeule, J., et al., *Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009.* Environ Health Perspect, 2012. **120**(7): p. 965-70.
31. Puett, R.C., et al., *Particulate matter air pollution exposure, distance to road, and incident lung cancer in the nurses' health study cohort.* Environ Health Perspect, 2014. **122**(9): p. 926-32.
32. Turner, M.C., et al., *Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers.* Am J Respir Crit Care Med, 2011. **184**(12): p. 1374-81.
33. Crouse, D.L., et al., *Ambient PM2.5, O(3), and NO(2) Exposures and Associations with Mortality over 16 Years of Follow-Up in the Canadian Census Health and Environment Cohort (CanCHEC).* Environ Health Perspect, 2015. **123**(11): p. 1180-6.
34. Pinault, L., A. van Donkelaar, and R.V. Martin, *Exposure to fine particulate matter air pollution in Canada.* Health Rep, 2017. **28**(3): p. 9-16.
35. Ruan, Y., et al., *Estimates of the current and future burden of cancer attributable to red and processed meat consumption in Canada.* Prev Med, 2019. **122**: p. 31-39.

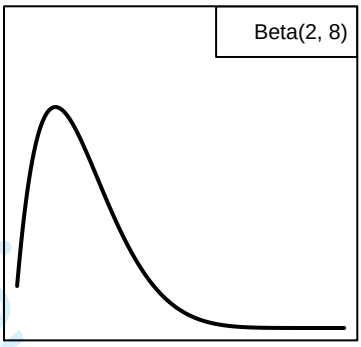


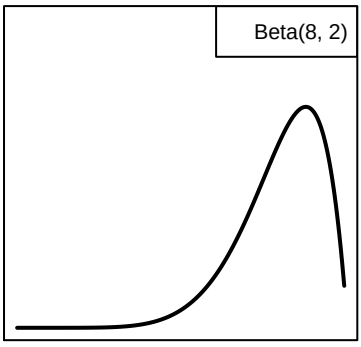


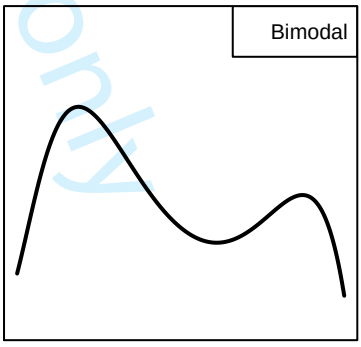


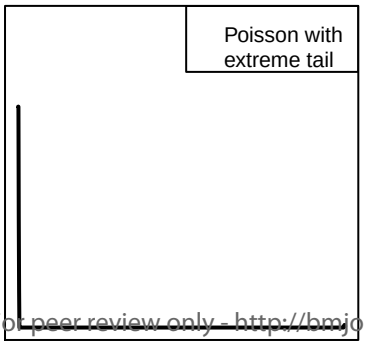


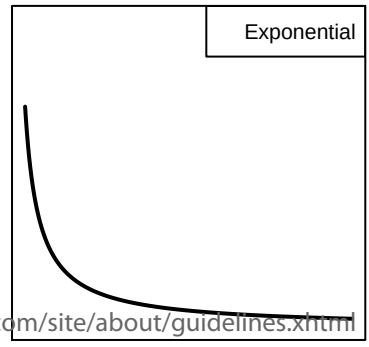


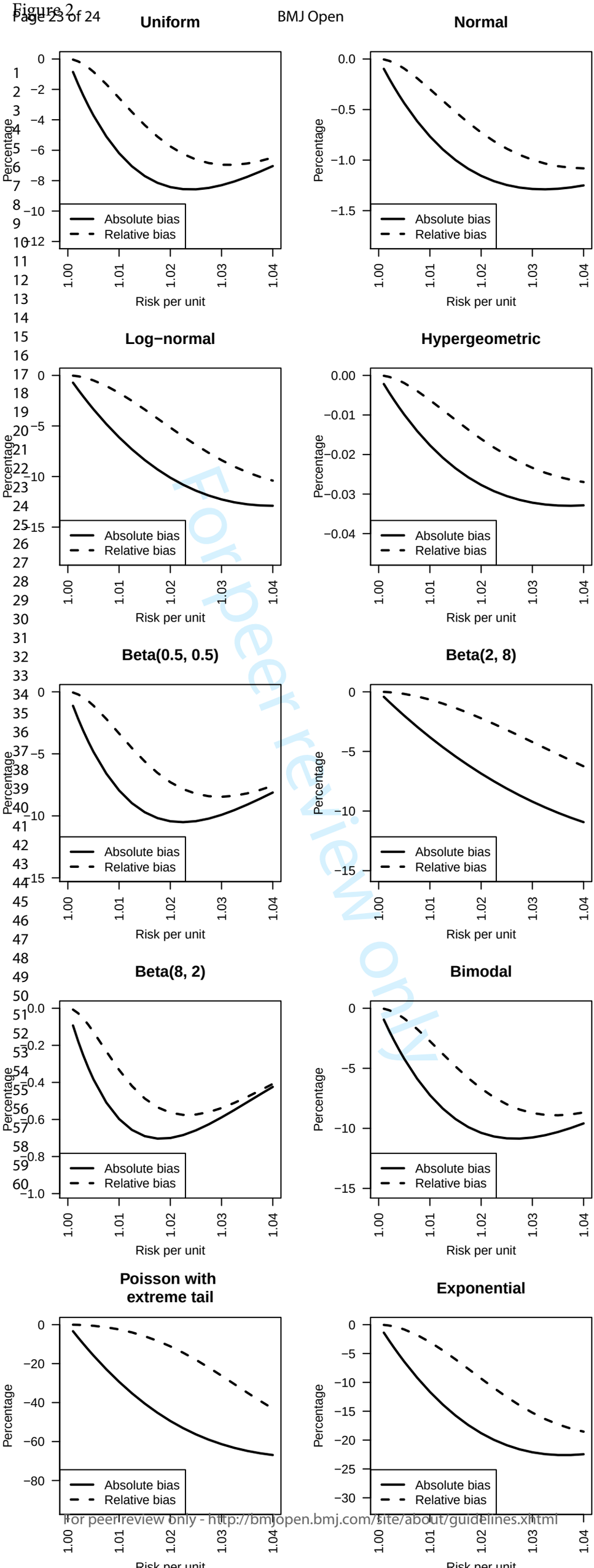


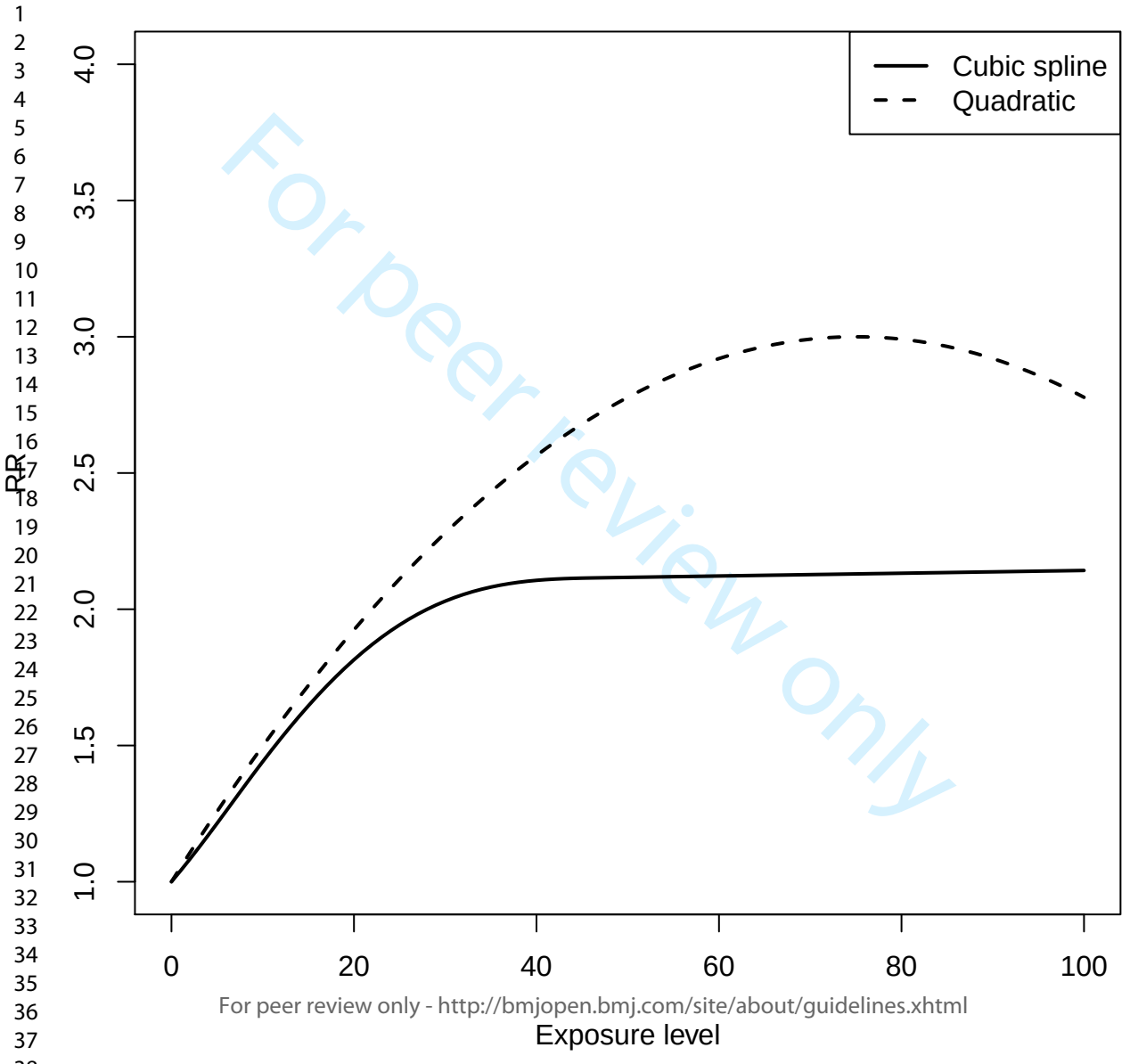


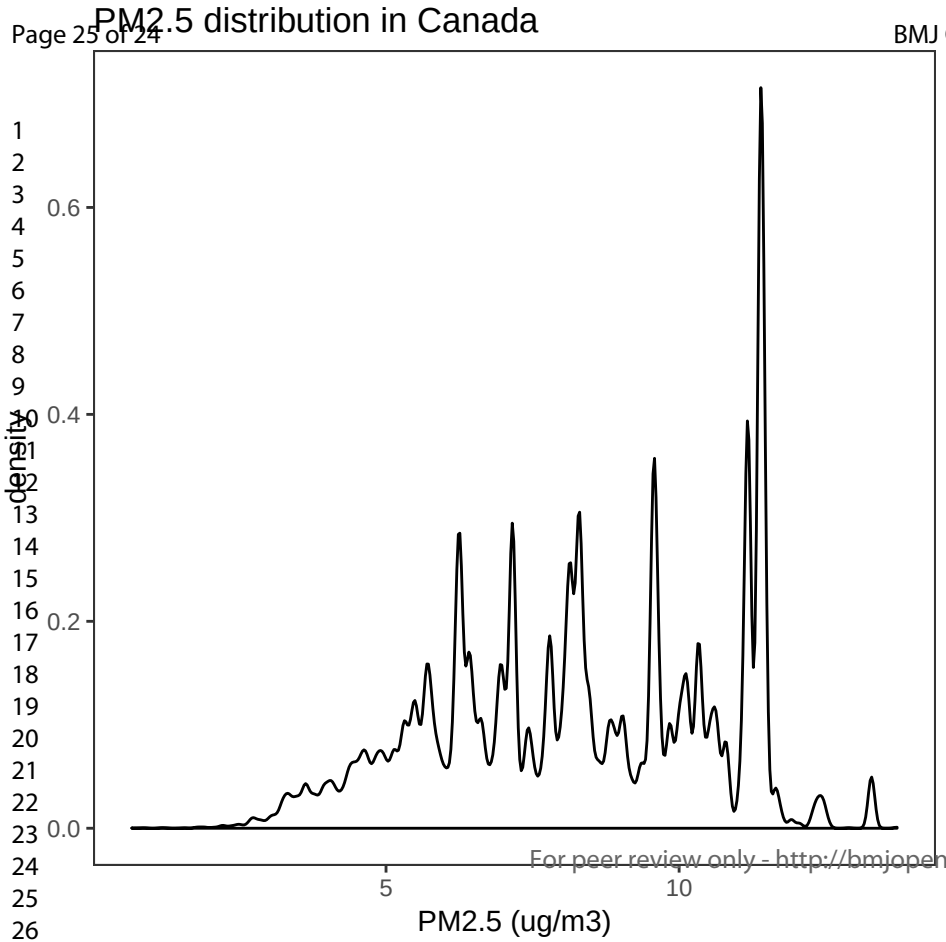


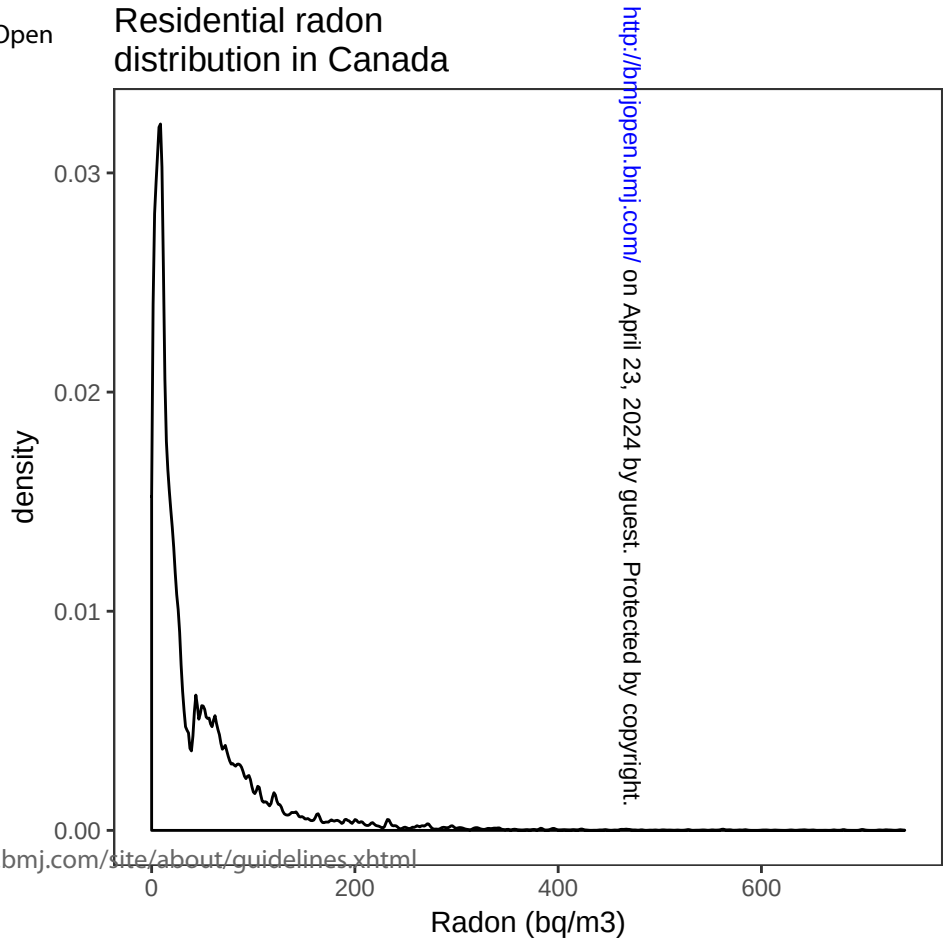












BMJ Open

A simulation study on the validity of the average risk approach in estimating population attributable fractions for continuous exposures

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-045410.R2
Article Type:	Original research
Date Submitted by the Author:	24-Apr-2021
Complete List of Authors:	Ruan, Yibing; Alberta Health Services, Cancer Epidemiology and Prevention Research Walter, Stephen ; McMaster University, Department of Health Research Methods, Evidence, and Impact Gogna, Priyanka ; Queen's University, Department of Public Health Sciences Friedenreich, CM; Alberta Health Services, Cancer Epidemiology and Prevention Research; University of Calgary Cumming School of Medicine, Departments of Oncology and Community Health Sciences Brenner, Darren; Alberta Health Services, Cancer Epidemiology and Prevention Research; University of Calgary Cumming School of Medicine, Departments of Oncology and Community Health Sciences
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Research methods
Keywords:	EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, PUBLIC HEALTH

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Original research: A simulation study on the validity of the average risk approach in estimating population attributable fractions for continuous exposures

Yibing Ruan[1], Stephen D. Walter[2], Priyanka Gogna[3], Christine M. Friedenreich[1,4], and Darren R. Brenner[1,4]

1. Department of Cancer Epidemiology and Prevention Research, Cancer Care Alberta, Alberta Health Services, Calgary, Alberta, Canada
2. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
3. Department of Public Health Sciences, Queen's University, Kingston, Ontario, Canada
4. Departments of Oncology and Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

*To whom correspondence should be addressed:

Yibing Ruan

Department of Cancer Epidemiology and Prevention Research

Cancer Care Alberta, Alberta Health Services

Holy Cross Centre - Room 513

2210-2nd ST. SW.

Calgary, AB, T2S 3C3

yibing.ruan@albertahealthservices.ca

Abstract (300 words)

Main text (2,888 words)

ABSTRACT

Background The population attributable fraction (PAF) is an important metric for estimating disease burden associated with causal risk factors. In an International Agency for Research on Cancer (IARC) working group report, an approach was introduced to estimate the PAF using the average of a continuous exposure and the incremental relative risk (RR) per unit. This "average risk" approach has been subsequently applied in several studies conducted worldwide. However, no investigation of the validity of this method has been done.

Objective To examine the validity and the potential magnitude of bias of the average risk approach.

Methods We established analytically that the direction of the bias is determined by the shape of the RR function. We then used simulation models based on a variety of risk exposure distributions and a range of RR per unit. We estimated the unbiased PAF from integrating the exposure distribution and RR, and the PAF using the average risk approach. We examined the absolute and relative bias as the direct and relative difference in PAF estimated from the two approaches. We also examined the bias of the average risk approach using real-world data from the Canadian Population Attributable Risk of Cancer study.

Results The average risk approach involves bias, which is under- or over-estimation with a convex or concave RR function (a risk profile that increases more/less rapidly at higher levels of exposure). The magnitude of the bias is affected by the exposure distribution as well as the value of RR. This approach is approximately valid when the RR per unit is small or the RR function is

approximately linear. The absolute and relative bias can both be large when RR is not small and the exposure distribution is skewed.

Conclusions We recommend that caution be taken when using the average risk approach to estimate PAF.

ARTICLE SUMMARY

Strengths and limitations of this study

- This study examined the assumptions and validity of the average risk approach to estimate the PAF, which has not been explored previously.
- We used both simulated and real-world data to demonstrate the factors associated with the bias of the average risk approach.
- As an empirical study, our simulation could only analytically establish the direction of bias of this approach and discuss the magnitude of bias using a limited number of risk exposure distributions and RR functions.

INTRODUCTION

Population Attributable Fraction (PAF) is an important measure for estimating the burden of disease in a population that is causally attributable to an exposure. Since its first introduction, PAF has received substantial attention in the field of epidemiology [1]. Many advances have been made in different approaches to calculating PAF of single and multiple risk factors [2-6], in estimating the variance [7, 8] and in the interpretation of PAF [9-11]. There have also been many comprehensive projects, either nationwide or globally, in estimating PAF for the burden of disease associated with its risk factors [12-22]. The International Agency for Research on Cancer has specialized in providing estimates of cancer surveillance and burden of cancer estimates from around the world. In 2007, Boffetta and colleagues [23] introduced an approach to estimating PAF when the prevalence data on a continuous exposure in the population under study are only available as a population average. This approach, to be referred to here as the "average risk approach", estimated the RR at average exposure of the whole population using the risk of disease per unit increase in exposure, and the average level of exposure of the whole population. No proof was provided at the time that this method was proposed. Hence, the purpose of this paper is to examine the underlying assumptions and validity of this average risk approach when estimating PAF for disease burden in a population. Specifically, we examined how the shape of the RR functions and the exposure distributions affect the validity of this approach.

METHODS

Description of Average Risk Approach

The average risk approach estimates the RR at an average exposure of the whole population using the RR of disease per unit increase in exposure along with the average level of exposure of the whole population as follows:

$$Risk = Exp^{[Ln(Risk\ per\ unit) \times average\ level\ of\ exposure]} = RR_{unit}^{\bar{x}} \qquad (1)$$

where *Risk* is the RR at the population average exposure, RR_{unit} is the RR associated with a unit increase in exposure, $\bar{x}$ is the weighted average level of exposure. An underlying assumption with this method is that a log-linear relationship exists between the exposure and the risk of cancer. The average risk approach then estimates PAF as:

$$PAF = \frac{Risk - 1}{Risk} \qquad (2)$$

where it was assumed that "each individual has experienced a similar average exposure" (IARC 2007, pg 5). Under this assumption, that all population under study are exposed at the population average level, formula (2) is a simplification of Levin's formula when the prevalence (*P*) is 100%:

$$PAF = \frac{P(RR - 1)}{1 + P(RR - 1)} \qquad (3)$$

Boffetta et al. stated that "This formula is valid when the risk of cancer per unit of exposure was estimated in a model using log transformation. This is the case for logistic regression or Poisson regression, which are models widely used in case-control and cohort studies respectively" (IARC 2007, pg 5). No proof was shown for this statement, although the authors went on to acknowledge that "the dose-effect relationship is, in fact, rarely linear (or log-linear) over the whole range of exposures, but this method is considered to be the best approximation available in this respect". Therefore, the validity of the average risk approach has not been fully assessed,

particularly concerning its sensitivity to departures from the assumed dose-response relationship, or concerning the impact of the exposure distribution.

When the distribution of a continuous exposure is known and no confounding is assumed, a valid method to estimate PAF involves integrating across all levels of exposure:

$$PAF = \frac{\int_{x=0}^{m} RR(x)P(x)dx - 1}{\int_{x=0}^{m} RR(x)P(x)dx} \quad (4)$$

where RR(x) is the RR at exposure x; P(x) is the population distribution of exposure; and m is the maximum exposure level. Note that if there were to be no bias in the average risk approach, the following equation would have to hold:

$$\int_{x=0}^{m} RR(x)P(x)dx = RR_{unit}^{\bar{x}} \quad (5)$$

Under the log-linear risk assumption, the left-hand side of equation (5) becomes:

$$\int_{x=0}^{m} RR_{unit}^{x}P(x)dx \quad (6)$$

Define $g(x) = RR_{unit}^{x}$ in which x is a random variable with distribution $P(x)$, (6) is E[$g(x)$], and the right-hand side of (5) is g[E(x)], because $g(x)$ is strictly convex (i.e., a line segment connecting any two points on the graph of a function lies above the graph) when RR_{unit} is greater than 1, the Jensen's inequality [24] determines that:

$$RR_{unit}^{\bar{x}} \leq \int_{x=0}^{m} RR_{unit}^{x}P(x)dx \quad (7)$$

According to (7), the average risk approach will not overestimate PAF. The magnitude of the bias is determined by the extent of the convexity of $g(x)$ over the effective range of x. When RR_{unit} is small (i.e., close to 1.00), $g(x)$ is approximately linear and there is little bias. However, whether or not the choice of the exposure distribution $P(x)$ affects the validity of this approach is unexplored. Specifically, it is unknown, if the exposure distribution in a population is strongly skewed or bimodal, whether or not the average risk approach still provides a good approximation to the actual PAF. Therefore, we studied the validity of the average risk approach under the loglinear RR function and a variety of exposure distributions.

In broad terms, when the loglinear function of RR is not assumed, the average risk approach can still be generalized as equation (2), in which "*Risk*" is the RR at the population average exposure level. It can be reasoned that the curvature of the RR function determines the direction and the magnitude of the bias. When RR is a linear function of the exposure (i.e., $RR(x) = 1 + k \cdot x, x \in [0,m]$), there is no bias, because the integral PAF ($\int_{x=0}^{m}(1+kx)P(x)dx$) and the average risk PAF ($1 + k\int_{x=0}^{m}xP(x)dx$) are equivalent. When the RR function has a convex form, which indicates a risk profile that increases more rapidly at higher levels of exposure, this approach underestimates PAF. In contrast, it overestimates PAF with a concave RR function, which indicates a risk profile that increases less rapidly at higher levels of exposure. To illustrate the latter point, we included two examples of simulated concave RR functions and calculated the bias of the average risk approach.

Investigation of Validity of Average Risk Approach

To investigate whether or not the validity of the average risk approach is affected by the exposure distribution, we simulated several exposure distributions where the exposure is

continuous, ranging between standardized values of 0 to 100, with 0 indicating no exposure and 100 indicating the maximal level of exposure in the population (Figure 1). The prevalence distributions were scaled so that the prevalence of all exposure levels summed to 100%. The details of the distributions are summarized in Table 1. We calculated PAF using both the average risk approach and by integrating across all exposure levels. We calculated the absolute bias ($PAF_{AvgRisk} - PAF_{Integral}$) and the relative bias $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$. Note that because PAF is often expressed as a percentage, their absolute and relative biases are both in percentage units. However, the meaning of the former is the percentage points, and the meaning of the latter is an actual percentage. For example, an absolute bias of -5% from the difference of $PAF_{AvgRisk}$ of 15% and $PAF_{Integral}$ of 20% indicates a relative bias of -25%.

To examine if the magnitude of risk affects the validity of the average risk approach, we tested a range of values for the RR per standardized unit, from 1.001 to 1.04. Using a standardized unit resolves the scaling issue of the unit. For example, the RR of standardized unit of body mass index and the disease associated with obesity is the same for the RR of 1 kg/m^2 or 5 kg/m^2, as long as it pertained to a single population. In this study, we refer to RR per standardized unit as "RR per unit", unless otherwise stated. We also considered that the risk becomes implausible for RR per unit values above 1.04. For example, the RR at maximal exposure level would be 132, if the RR per unit is 1.05 under the log-linear assumption.

In addition, we illustrated the bias of the average risk approach when the RR function is non-linear or loglinear. In particular, we used two simulated examples of quadratic and cubic spline RR functions, which are both concave (Figure S1). The quadratic RR function has a form of $RR(x) = 3 - 2(\frac{x}{k} - 1)^2$, in which $k \in [\frac{m}{2}, m]$. This quadratic form has RR=1 when x=0, and

RR has a maximum of 5 when $x = k$. In the illustrated example, we used k = 75, i.e., 75% of the maximal exposure. The cubic spline RR function is based on simulated data, with the function being approximately quadratic in the lower exposure range, and approximately linear at higher exposures.

Finally, we used real-world data of the distribution of air pollution ($PM_{2.5}$) and residential radon exposures, which were investigated in the Canadian Population Attributable Risk of Cancer (ComPARe) study. The ComPARe study collected national-representative and population-weighted exposure data of $PM_{2.5}$ and residential radon and used the integral approach to estimate PAF of lung cancer for 2015 for Canada.[25, 26] We compared this PAF to that obtained using the average risk approach, to illustrate the validity of this approach. We also estimated the approximate 95%CI of the PAFs and the bias, assuming a fixed prevalence distribution for simplicity and a lognormal distribution of the RR. We resampled 10,000 RRs from this distribution and calculated PAF and bias. We used the 2.5% and 97.5% quantiles as the approximate 95%CI.

Patient or public involvement

No patients involved.

RESULTS

First, we examined the bias of the average risk approach under the loglinear RR function with the exposure distributions we selected in Table 1. The results at RR per unit of 1.001, 1.01 and 1.03 were illustrated in Table 2 and the results with a range of RR per unit from 1.001 to 1.04 were shown in Figure 2. At RR of 1.001, the absolute and relative biases were very small

and the average risk approach can be regarded unbiased. At RR of 1.01, the absolute bias remained small for all tested distributions although the relative bias started to increase substantially in the power distribution and in the Poisson distribution with an extreme tail (Table 2). At RR of 1.03, large absolute and relative biases were observed in several distributions. However, the normal and hypergeometric distributions were more robust than the Poisson with extreme tail and power distributions with the increase in RR (Table 2, Figure 2). For some distributions (uniform, beta (0.5, 0.5), beta (8, 2), and bimodal), the largest absolute and relative bias occurred at an intermediate value of RR (Figure 2). As RR increases, the bias becomes smaller, because the PAF estimates approaches 100%. Regardless of the exposure distribution and the magnitude of RR, the direction of the bias is underestimation in the case of loglinear RR.

We then illustrated the direction of the bias when the RR function is concave. Table 3 showed the resulting bias of the two RR functions in Figure S1 when the exposure distributions were as reported in Table 2. With concave RR functions, the direction of the bias in the average risk approach is overestimation. Similar to the loglinear RR function, we observed little bias in normal, hypergeometric, and beta(8, 2) distributions, whereas substantial bias was observed in power, Poisson with extreme tail, and beta(0.5, 0.5) distributions.

Finally, we explored the bias of the average risk approach using real-world data for air pollution ($PM_{2.5}$) and residential radon. Epidemiologic studies support a loglinear RR function between exposure to residential radon and lung cancer [27, 28]. A loglinear dose response between PM2.5 and lung cancer risk was less consistent. The loglinear relationship was supported by several studies [29-32], while two studies reported some deviation from it [33, 34]. The 2019 Global Burden of Disease Study of 87 risk factors suggested that PM2.5 has a loglinear relation with lung cancer in low exposure range (0-50 ug/m3) and a linear relation in

high exposure range (>50 ug/m3) [16]We assumed a loglinear relation for PM2.5 because the level is typically below 20 ug/m3 in Canada. We found that both exposures had skewed distributions (Figure S2). The $PM_{2.5}$ distribution had a long left tail, while the distribution of residential radon has a long right tail. We standardized the exposure levels of $PM_{2.5}$ and radon to 0.14 ug/m^3 and 7.4 Bq/m^3 per unit, so that the maximal exposure level is 100 units. The RR per unit of $PM_{2.5}$ associated with lung cancer was 1.0012 (95%CI: 1.0008 to 1.0016). The PAFs of $PM_{2.5}$ using the integral and the average risk approach were 6.89% (95%CI: 4.71% to 8.98%) and 6.87% (95%CI: 4.70% to 8.95%), respectively, indicating very small bias in the average risk approach (-0.02%, 95%CI: -0.03% to -0.01%). The RR per unit of radon associated with lung cancer was 1.011 (95%CI: 1.005 to 1.016). The PAFs of radon using the integral and average risk approach were 6.87% (95%CI: 3.33% to 10.52%) and 6.37% (95%CI: 3.21% to 9.37%), respectively. The bias was larger than that seen in $PM_{2.5}$. The absolute bias was -0.5% (95%CI: -1.2% to -0.1%) and the relative bias was -7.3% (95%CI: -11.0% to -3.5%), indicating slight to moderate bias. The observations were consistent with the simulations, in that small RRs yield little bias ($PM_{2.5}$), and moderate to large RRs could produce bias with some skewed exposure distributions (radon).

DISCUSSION

Since being introduced by Boffetta and colleagues in 2007, the average risk approach has been used in several PAF estimation projects.[12-15, 35] In addition to the cancer burden study in France,[15] the ComPARe study in Canada,[35] a study of attributable causes in China,[12] and two studies in Brazil[13, 14] have used this method. We illustrated that the direction of bias of the average risk approach is determined by whether the RR function is convex or concave,

while the magnitude of bias is affected by the degree of convexity or concavity, as well as the exposure distribution. When the RR per unit is small under a loglinear RR function, the magnitude of bias is also small and the average risk approach is approximately valid. With larger RR and increased convexity, the validity of the average risk approach would also depend on the exposure distribution. We demonstrated that under some circumstances (e.g., Poisson distribution with extreme tail, power distribution), the approach could potentially lead to moderate to severe bias.

The average risk approach has an implicit assumption that the minimal risk exposure value is 0. When the minimal risk exposure value is not 0, this approach generates invalid estimates. To offer a simplified example, overweight and obesity defined as body mass index (BMI) $\geq$ 25.0 kg/m^2 is associated with postmenopausal breast cancer. The minimal risk exposure value of BMI is 25.0 kg/m^2. Assuming a log-linear relationship between BMI above 25.0 kg/m^2 and the risk of breast cancer and that a postmenopausal female population has a normal distribution of body mass index (BMI) at a mean and standard deviation of 25.0 and 5.0 kg/m^2. The average risk approach yields a PAF of 0 in this population, because the population average risk exposure is 25.0 kg/m^2, which has a RR of 1.0. Although it is possible to recode the exposure so that the minimal exposure is zero, a new average of the recoded exposure must be estimated, which requires the information of the exposure distribution. On the other hand, the prerequisite of applying the average risk approach is that such information is only available as a population average. In practice, many natural or physiological exposures have a non-zero minimal risk exposure value and the estimation of PAF for such exposures requires additional considerations [36]. Therefore, this implicit assumption is a substantial limitation of this

approach. For the same reason, the average risk approach cannot be applied in the framework of generalized impact fraction, in which the impact of partial reduction of exposure is considered.

Our study has some limitations. First, this study is an empirical examination of the validity of the average risk approach. We have mathematically demonstrated the direction of the bias in this approach. However, we only qualitatively discussed the magnitude of the bias associated with the RR function and the exposure distribution. We illustrated the magnitude of the bias through several RR functions and exposure distributions. However, this pragmatic approach could not cover all RR functions and distributions. Second, we compared the average risk approach to the integral approach under the assumption of no confounding. The integral approach is an extension of Levin's formula, which is biased in the presence of confounding [1, 11]. Ideally, the validity of the average risk approach should be tested against the integral form of Miettinen's formula, which is based on the prevalence of exposure among the cases and is valid in the presence of confounding [6]. However, because the average risk approach was developed under the framework of Levin's formula, we considered that a comparison of two approaches under the same framework would be more appropriate. Nevertheless, it should be noted that the validity of the average risk approach is also prone to the presence of confounding, just like Levin's formula.

In conclusion, we have shown that the average risk approach has some utility, but nonetheless carries the risk of bias. We highly recommend using alternative approaches when the RR per unit is not small in the range of the exposure, the RR functions depart from linear, or the exposure distribution data are available. The average risk approach can be used if the RR per unit is small, or there is evidence that the exposure distribution is not highly skewed. Nevertheless,

researchers using this approach should discuss the direction of the bias based on the RR functions.

FOOTNOTES

Contributors: YR participated in study conceptualization, statistical analyses, drafted the initial manuscript and approved the final version of the manuscript. SW participated in study conceptualization, supervision, and critically reviewed and edited the manuscript. PG provided resources (ComPARe datasets), critically reviewed and edited the manuscript, and approved the final version of the manuscript. CF participated in funding acquisition, supervision, critically reviewed and edited the manuscript, and approved the final version of the manuscript. DB participated in funding acquisition, supervision, and approved the final version of the manuscript.

Funding: This study was supported by the Canadian Cancer Society Partner Prevention Research Grant (grant #703106).

Competing interests: None declared.

Patient consent for publication: Not required.

Ethics approval: Not required.

Data sharing statement: Extra data, including the R code for simulation and the exposure datasets from the ComPARe study, are available by emailing to yibing.ruan@albertahealthservices.ca.

Table 1. Description of the exposure distributions used in this study.

Distribution	**Note**
Uniform	Range from 0 to 100
Normal	μ=50, σ=10
Log-normal	μ=5, σ=0.5
Hypergeometric	N=700, K=200, m=200
Beta	α=0.5, β=0.5
Beta	α=2, β=8
Beta	α=8, β=2
Bimodal	Constructed by combining the lognormal distribution (μ=5, σ=0.5) with one-third of beta (8, 2).
Poisson with extreme tail	Constructed by applying the Poisson distribution (k = 0 to 3, λ=1) to exposure level 0 to 3, and one-tenth of the Poisson distribution (k = 70, 75, 80, 85, 90, λ=80) to exposure level 95 to 99
Power	Constructed by rescaling the function of 1/x, where x $\in$ [0.1, 2.5].

Note: All distributions were scaled to ensure that the sum of distribution is 100%.

Table 2. Absolute and relative bias in PAF between the average risk approach and the integration approach in selected exposure distributions when RR per unit is 1.001, 1.01 or 1.03 for the loglinear function

RRunit	1.001				1.01				1.03			
Distribution	PAF_{Integr}	PAF_{Avera}	Absolute bias	Relative bias	PAF_{Integr}	PAF_{Avera}	Absolute bias	Relative bias	PAF_{Integr}	PAF_{Avera}	Absolute bias	Relative bias
Uniform	4.9%	4.8%	0%	-0.9%	41.4%	38.9%	-2.6%	-6.2%	83.8%	76.8%	-7%	-8.3%
Normal	4.8%	4.8%	0%	-0.1%	38.9%	38.6%	-0.3%	-0.8%	77.6%	76.6%	-1%	-1.3%
Log-normal	3.1%	3.0%	0%	-0.7%	28.3%	26.5%	-1.7%	-6.1%	68.3%	60.0%	-8.4%	-12.3%
Hypergeometric	4.3%	4.3%	0%	0%	35.3%	35.3%	0%	0%	72.6%	72.6%	0%	0%
Beta(0.5, 0.5)	4.9%	4.8%	-0.1%	-1.1%	42.3%	38.9%	-3.4%	-7.9%	85.3%	76.9%	-8.5%	-9.9%
Beta(2, 8)	1.8%	1.8%	0%	-0.4%	17.2%	16.5%	-0.7%	-3.8%	45.7%	41.5%	-4.2%	-9.2%
Beta(8, 2)	7.8%	7.8%	0%	-0.1%	55.7%	55.3%	-0.3%	-0.6%	91.4%	90.9%	-0.5%	-0.6%
Bimodal	4.3%	4.3%	0%	-0.9%	37.9%	35.1%	-2.7%	-7.2%	81.1%	72.3%	-8.7%	-10.8%
Poisson with extreme tail	0.6%	0.6%	0%	-3.4%	8.3%	5.9%	-2.4%	-29.3%	42.7%	16.5%	-26.2%	-61.4%
Power	2.6%	2.6%	0%	-1.4%	25.9%	22.8%	-3%	-11.6%	69.0%	53.7%	-15.2%	-22.1%

Note: the absolute bias is $PAF_{AvgRisk} - PAF_{Integral}$ and the relative bias is $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

Table 3. Absolute and relative bias in PAF between the average risk approach and the integration approach in two illustrated examples of concave RR functions.

RR funtion	Cubic spline				Quadratic			
Distribution	$PAF_{Integral}$	$PAF_{Average\ risk}$	Absolute bias	Relative bias	$PAF_{Integral}$	$PAF_{Average\ risk}$	Absolute bias	Relative bias
Uniform	49.0%	52.8%	3.8%	7.8%	59.6%	64.0%	4.4%	7.4%
Normal	52.6%	52.8%	0.1%	0.3%	63.6%	64.1%	0.5%	0.7%
Log-normal	46.4%	51.3%	4.9%	10.6%	54.3%	57.3%	3%	5.6%
Hypergeometric	52.6%	52.7%	0.1%	0.1%	62.4%	62.5%	0.1%	0.2%
Beta(0.5, 0.5)	46.9%	52.8%	5.8%	12.4%	57.7%	64.0%	6.3%	11%
Beta(2, 8)	40.9%	43.9%	3.1%	7.6%	45.9%	47.4%	1.5%	3.3%
Beta(8, 2)	53.1%	53.1%	0%	0%	65.9%	66.5%	0.6%	0.9%
Bimodal	48.3%	52.7%	4.4%	9.2%	57.9%	62.4%	4.5%	7.8%
Poisson with extreme tail	6.1%	11.1%	5%	81%	8.5%	13.6%	5.1%	60.6%
Power	38.7%	49.1%	10.4%	26.9%	47.1%	53.4%	6.4%	13.5%

Note: the absolute bias is $PAF_{AvgRisk} - PAF_{Integral}$ and the relative bias is $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

Figure legends:

Figure 1: Probability density curves of selected distributions in this study.

Figure 2: The absolute and relative bias of the average risk approach under the selected distributions and a range of RR per unit. Both absolute and relative bias are presented as a percentage. The absolute bias is the difference in PAF percentage, and the relative bias is the difference in PAF over the PAF using integration and expressed as a percentage.

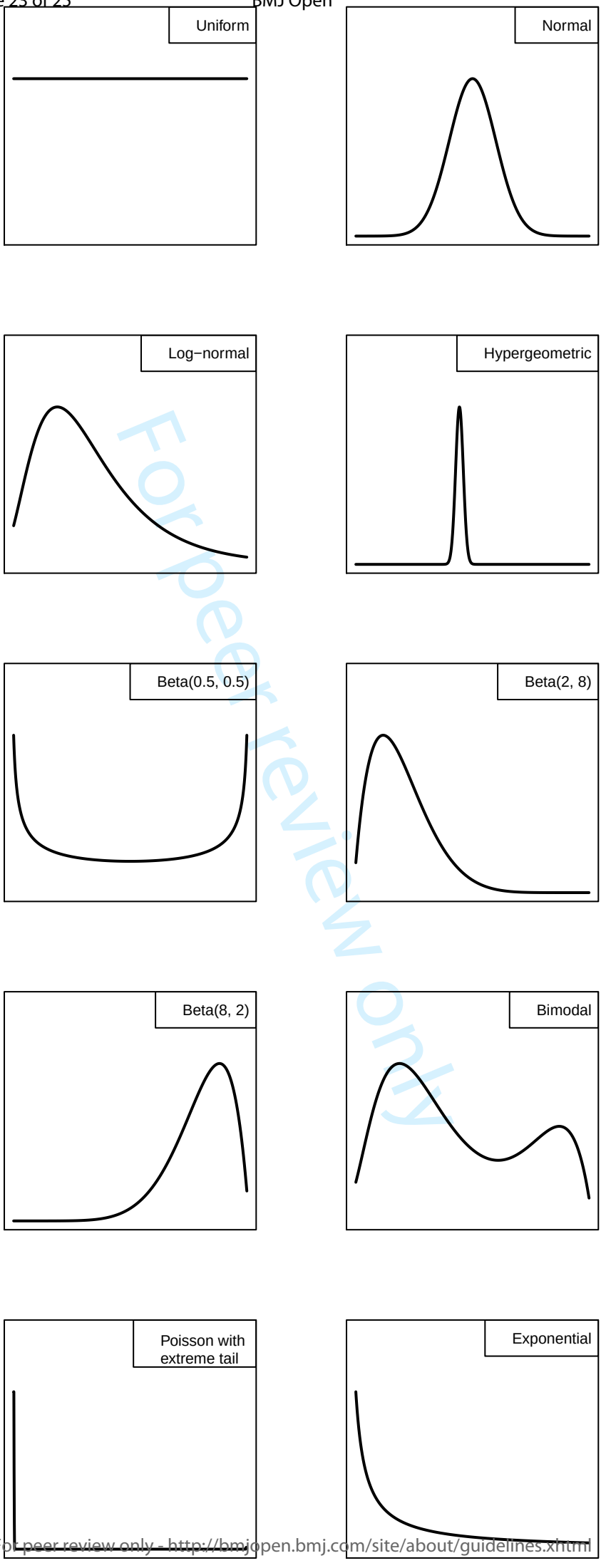
Figure S1: Graph of the two concave RR functions used in this study to illustrate the direction and the magnitude of bias of the average risk approach

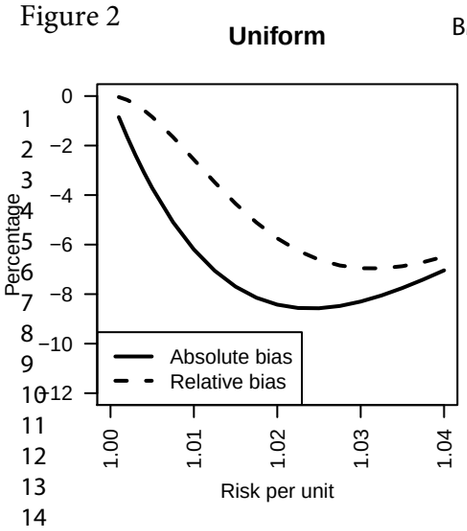
Figure S2: The smoothed density plot of the distributions of $PM_{2.5}$ and residential radon in Canada

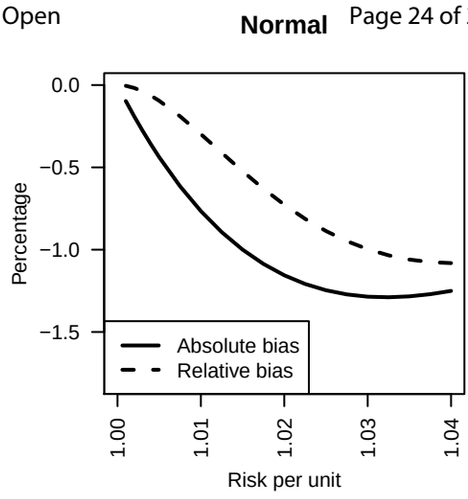
REFERENCES

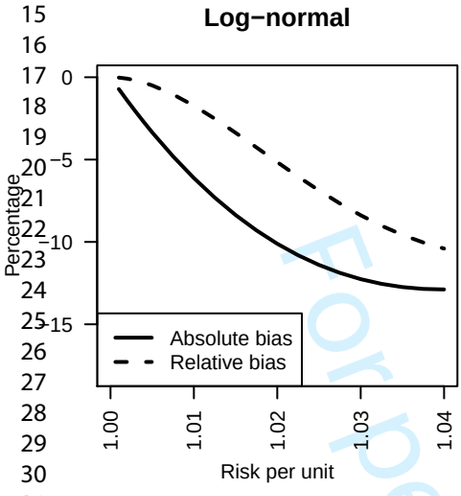
1. Levin, M.L., *The occurrence of lung cancer in man.* Acta Unio Int Contra Cancrum, 1953. **9**(3): p. 531-41.
2. Bruzzi, P., et al., *Estimating the population attributable risk for multiple risk factors using case-control data.* Am J Epidemiol, 1985. **122**(5): p. 904-14.
3. Eide, G.E. and I. Heuch, *Average attributable fractions: a coherent theory for apportioning excess risk to individual risk factors and subpopulations.* Biom J, 2006. **48**(5): p. 820-37.
4. Walter, S.D., *The estimation and interpretation of attributable risk in health research.* Biometrics, 1976. **32**(4): p. 829-49.
5. Whittemore, A.S., *Statistical methods for estimating attributable risk from retrospective data.* Stat Med, 1982. **1**(3): p. 229-43.
6. Miettinen, O.S., *Proportion of disease caused or prevented by a given exposure, trait or intervention.* Am J Epidemiol, 1974. **99**(5): p. 325-32.
7. Benichou, J. and M.H. Gail, *Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models.* Biometrics, 1990. **46**(4): p. 991-1003.
8. Greenland, S., *Variance estimators for attributable fraction estimates consistent in both large strata and sparse data.* Stat Med, 1987. **6**(6): p. 701-8.
9. Mansournia, M.A. and D.G. Altman, *Population attributable fraction.* BMJ, 2018. **360**: p. k757.
10. Di Maso, M., et al., *Attributable fraction for multiple risk factors: Methods, interpretations, and examples.* Stat Methods Med Res, 2020. **29**(3): p. 854-865.
11. Rockhill, B., B. Newman, and C. Weinberg, *Use and misuse of population attributable fractions.* Am J Public Health, 1998. **88**(1): p. 15-9.
12. Wang, J.B., et al., *Attributable causes of cancer in China.* Ann Oncol, 2012. **23**(11): p. 2983-2989.
13. Azevedo, E.S.G., et al., *The Fraction of Cancer Attributable to Ways of Life, Infections, Occupation, and Environmental Agents in Brazil in 2020.* PLoS One, 2016. **11**(2): p. e0148761.
14. Rezende, L.F. and J. Eluf-Neto, *Population attributable fraction: planning of diseases prevention actions in Brazil.* Rev Saude Publica, 2016. **50**.
15. Boffetta, P., et al., *The causes of cancer in France.* Ann Oncol, 2009. **20**(3): p. 550-5.
16. Collaborators, G.B.D.R.F., *Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet, 2020. **396**(10258): p. 1223-1249.
17. Parkin, D.M., L. Boyd, and L.C. Walker, *16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010.* Br J Cancer, 2011. **105 Suppl 2**: p. S77-81.
18. Whiteman, D.C., et al., *Cancers in Australia in 2010 attributable to modifiable factors: summary and conclusions.* Aust N Z J Public Health, 2015. **39**(5): p. 477-84.
19. Diseases, G.B.D. and C. Injuries, *Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet, 2020. **396**(10258): p. 1204-1222.
20. Poirier, A.E., et al., *The current and future burden of cancer attributable to modifiable risk factors in Canada: Summary of results.* Prev Med, 2019. **122**: p. 140-147.
21. Arnold, M., et al., *Global burden of cancer attributable to high body-mass index in 2012: a population-based study.* Lancet Oncol, 2015. **16**(1): p. 36-46.
22. Islami, F., et al., *Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States.* CA Cancer J Clin, 2018. **68**(1): p. 31-54.

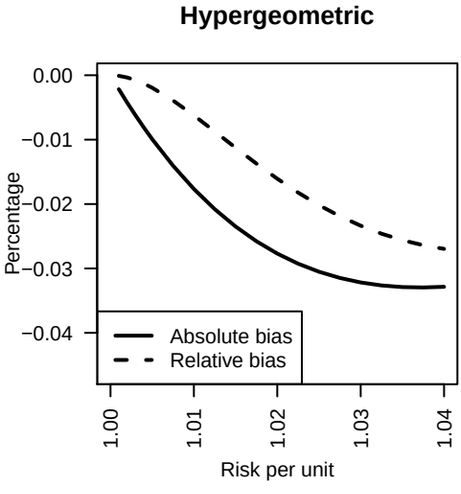
23. IARC, *Attributable Causes of Cancer in France in the Year 2000*, in *IARC Working Group Report Volume 3*. 2007.
24. Durrett, R., *Probability: Theory and Examples*. 5th ed. ed. 2019: Cambridge University Press.
25. Gogna, P., et al., *Estimates of the current and future burden of lung cancer attributable to PM2.5 in Canada.* Prev Med, 2019. **122**: p. 91-99.
26. Gogna, P., et al., *Estimates of the current and future burden of lung cancer attributable to residential radon exposure in Canada.* Prev Med, 2019. **122**: p. 100-108.
27. Krewski, D., et al., *A combined analysis of North American case-control studies of residential radon and lung cancer.* J Toxicol Environ Health A, 2006. **69**(7): p. 533-97.
28. Puskin, J.S., *Perspective on the use of LNT for radiation protection and risk assessment by the U.S. Environmental Protection Agency.* Dose Response, 2009. **7**(4): p. 284-91.
29. Hystad, P., et al., *Long-term residential exposure to air pollution and lung cancer risk.* Epidemiology, 2013. **24**(5): p. 762-72.
30. Lepeule, J., et al., *Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009.* Environ Health Perspect, 2012. **120**(7): p. 965-70.
31. Puett, R.C., et al., *Particulate matter air pollution exposure, distance to road, and incident lung cancer in the nurses' health study cohort.* Environ Health Perspect, 2014. **122**(9): p. 926-32.
32. Turner, M.C., et al., *Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers.* Am J Respir Crit Care Med, 2011. **184**(12): p. 1374-81.
33. Crouse, D.L., et al., *Ambient PM2.5, O(3), and NO(2) Exposures and Associations with Mortality over 16 Years of Follow-Up in the Canadian Census Health and Environment Cohort (CanCHEC).* Environ Health Perspect, 2015. **123**(11): p. 1180-6.
34. Pinault, L., A. van Donkelaar, and R.V. Martin, *Exposure to fine particulate matter air pollution in Canada.* Health Rep, 2017. **28**(3): p. 9-16.
35. Ruan, Y., et al., *Estimates of the current and future burden of cancer attributable to red and processed meat consumption in Canada.* Prev Med, 2019. **122**: p. 31-39.
36. Ferguson, J., et al., *Population attributable fractions for continuously distributed exposures.* Epidemiologic Methods, 2020. **9**(1).

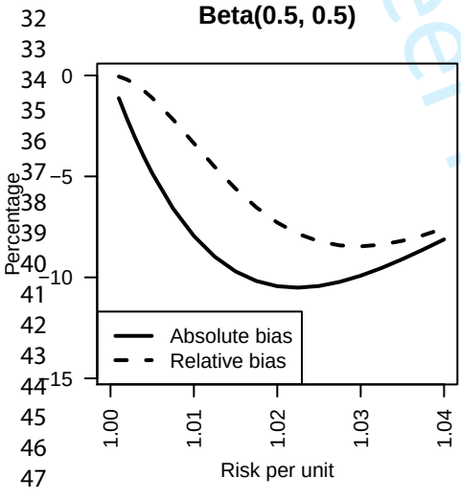


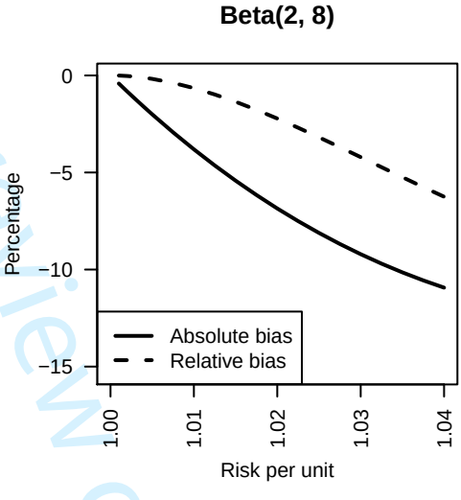


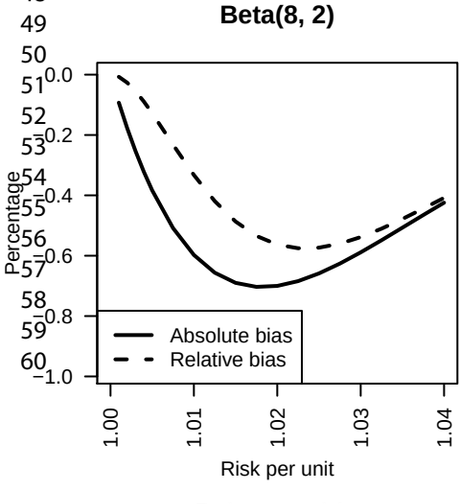


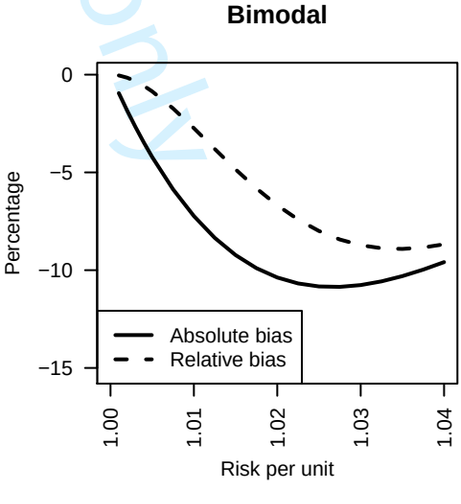


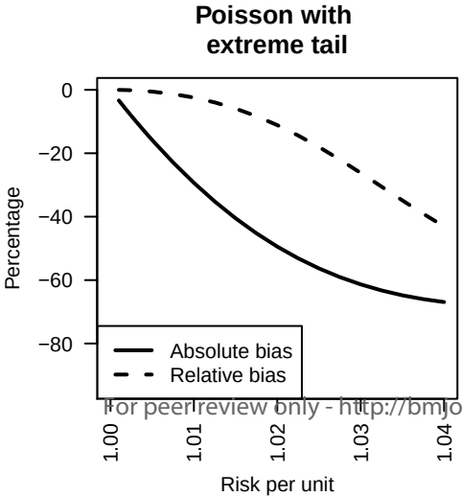












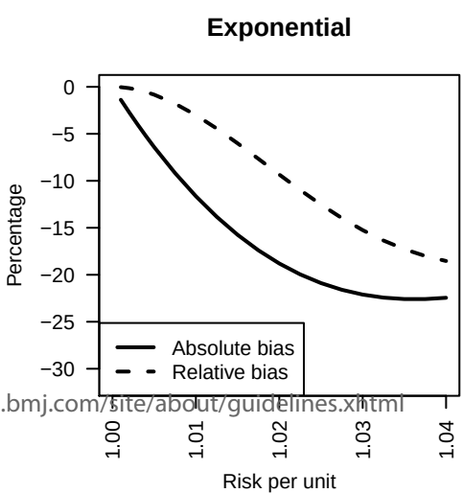
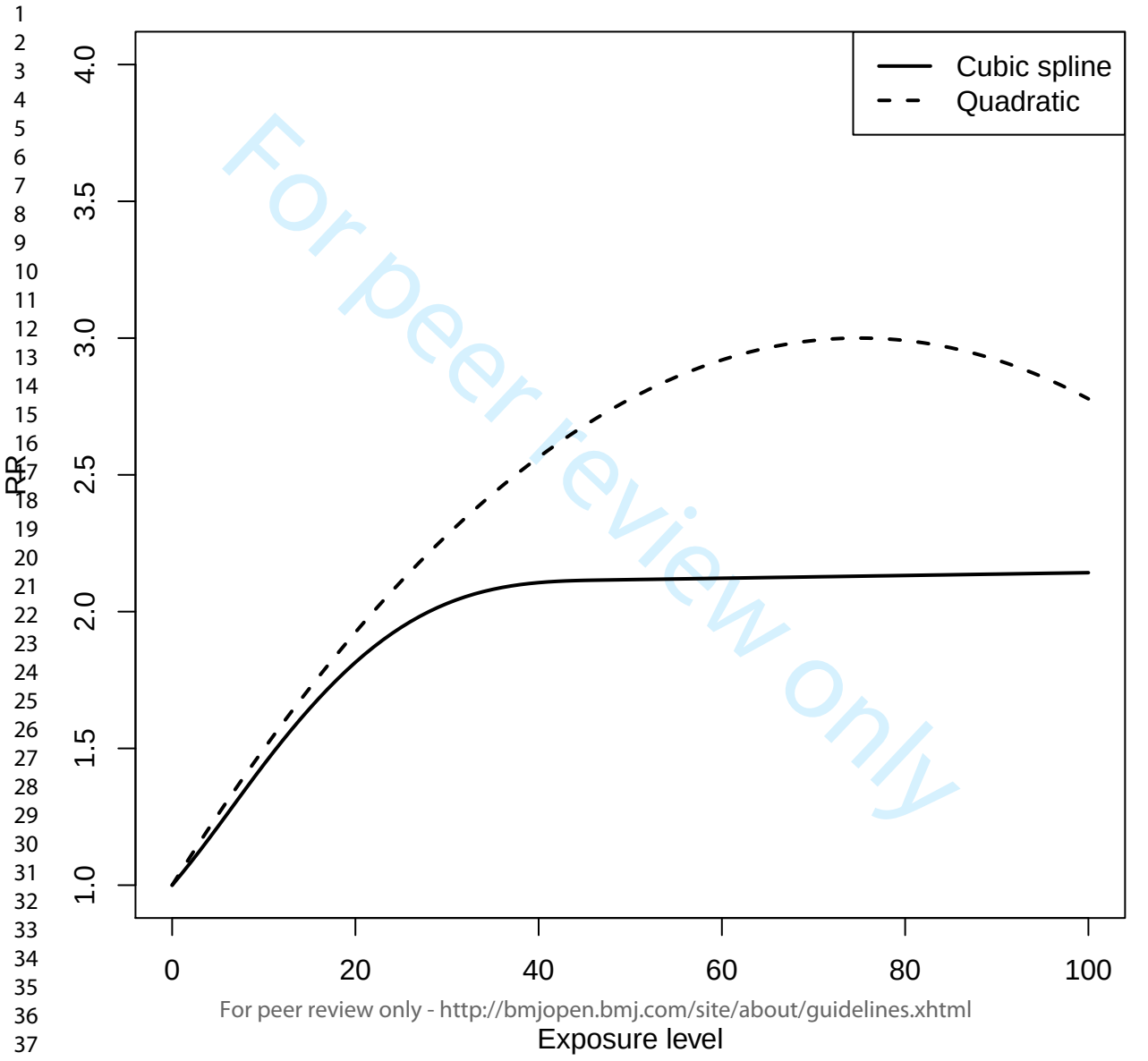
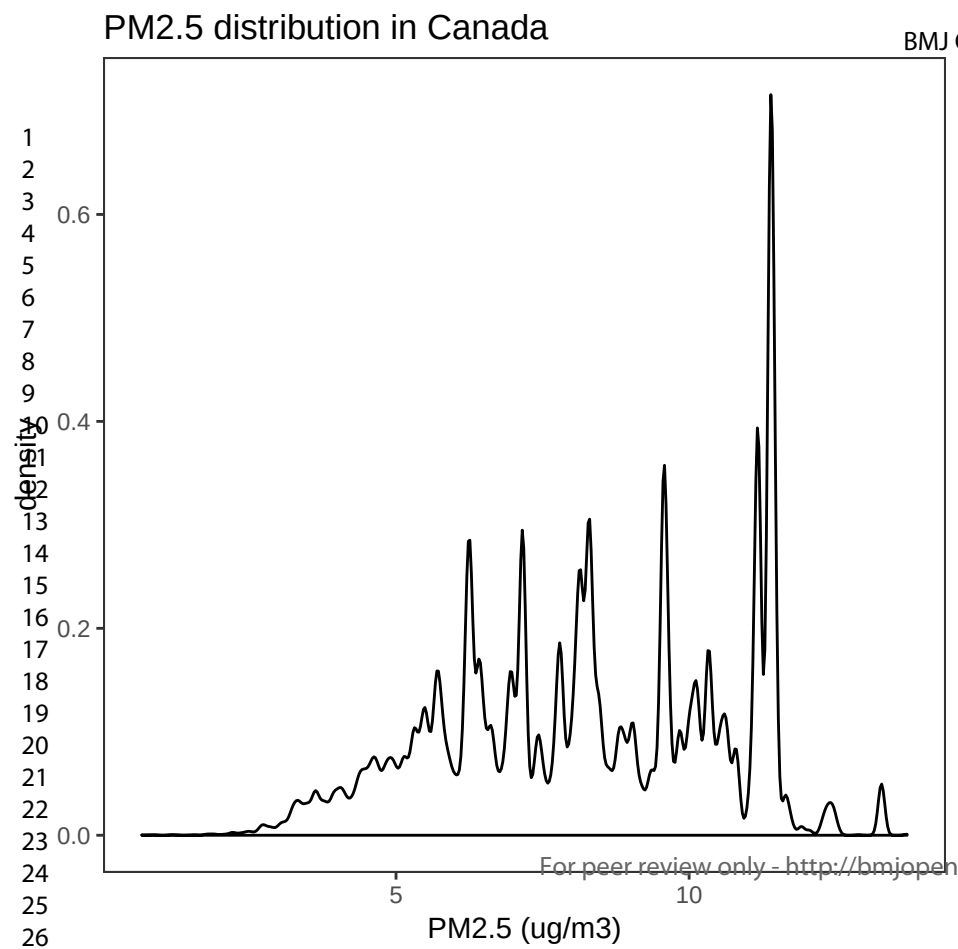
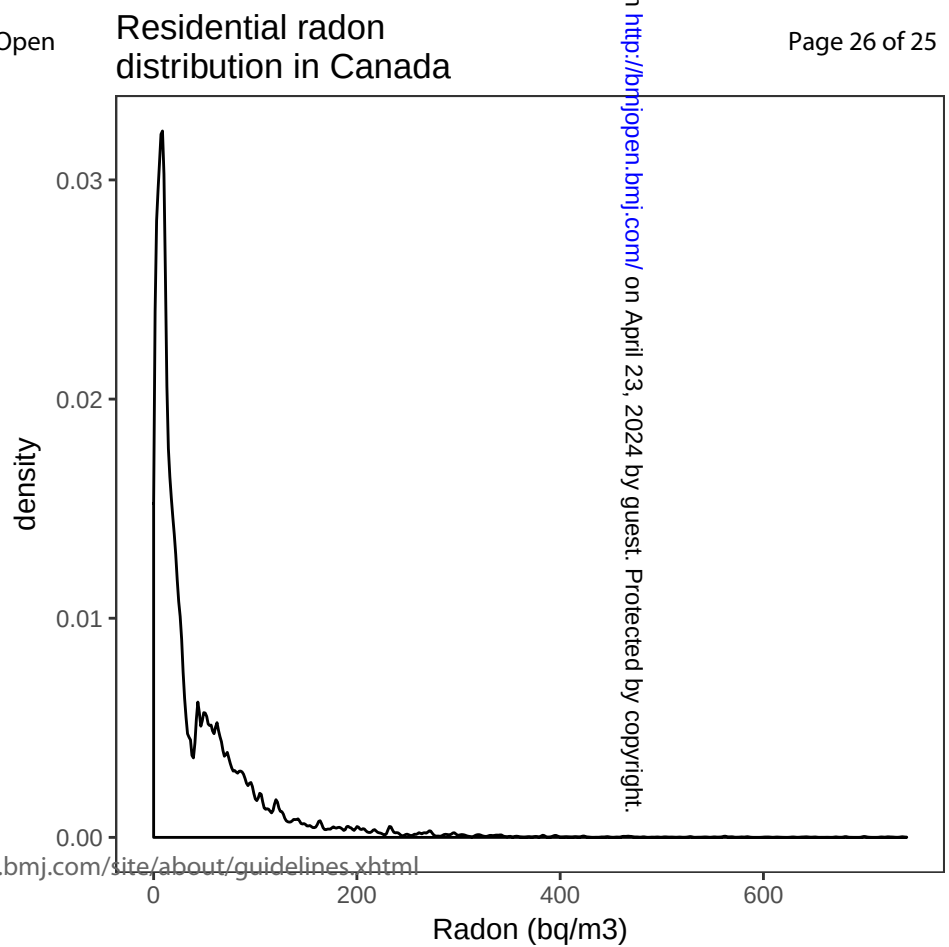


Figure S1







BMJ Open

A simulation study on the validity of the average risk approach in estimating population attributable fractions for continuous exposures

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2020-045410.R3
Article Type:	Original research
Date Submitted by the Author:	01-Jun-2021
Complete List of Authors:	Ruan, Yibing; Alberta Health Services, Cancer Epidemiology and Prevention Research Walter, Stephen ; McMaster University, Department of Health Research Methods, Evidence, and Impact Gogna, Priyanka ; Queen's University, Department of Public Health Sciences Friedenreich, CM; Alberta Health Services, Cancer Epidemiology and Prevention Research; University of Calgary Cumming School of Medicine, Departments of Oncology and Community Health Sciences Brenner, Darren; Alberta Health Services, Cancer Epidemiology and Prevention Research; University of Calgary Cumming School of Medicine, Departments of Oncology and Community Health Sciences
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Epidemiology, Research methods
Keywords:	EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, PUBLIC HEALTH

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Original research: A simulation study on the validity of the average risk approach in estimating population attributable fractions for continuous exposures

Yibing Ruan[1], Stephen D. Walter[2], Priyanka Gogna[3], Christine M. Friedenreich[1,4], and Darren R. Brenner[1,4]

1. Department of Cancer Epidemiology and Prevention Research, Cancer Care Alberta, Alberta Health Services, Calgary, Alberta, Canada
2. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
3. Department of Public Health Sciences, Queen's University, Kingston, Ontario, Canada
4. Departments of Oncology and Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

*To whom correspondence should be addressed:

Yibing Ruan

Department of Cancer Epidemiology and Prevention Research

Cancer Care Alberta, Alberta Health Services

Holy Cross Centre - Room 513

2210-2nd ST. SW.

Calgary, AB, T2S 3C3

yibing.ruan@albertahealthservices.ca

Abstract (300 words)

Main text (2,857 words)

ABSTRACT

Background The population attributable fraction (PAF) is an important metric for estimating disease burden associated with causal risk factors. In an International Agency for Research on Cancer (IARC) working group report, an approach was introduced to estimate the PAF using the average of a continuous exposure and the incremental relative risk (RR) per unit. This "average risk" approach has been subsequently applied in several studies conducted worldwide. However, no investigation of the validity of this method has been done.

Objective To examine the validity and the potential magnitude of bias of the average risk approach.

Methods We established analytically that the direction of the bias is determined by the shape of the RR function. We then used simulation models based on a variety of risk exposure distributions and a range of RR per unit. We estimated the unbiased PAF from integrating the exposure distribution and RR, and the PAF using the average risk approach. We examined the absolute and relative bias as the direct and relative difference in PAF estimated from the two approaches. We also examined the bias of the average risk approach using real-world data from the Canadian Population Attributable Risk of Cancer study.

Results The average risk approach involves bias, which is under- or over-estimation with a convex or concave RR function (a risk profile that increases more/less rapidly at higher levels of exposure). The magnitude of the bias is affected by the exposure distribution as well as the value of RR. This approach is approximately valid when the RR per unit is small or the RR function is

approximately linear. The absolute and relative bias can both be large when RR is not small and the exposure distribution is skewed.

Conclusions We recommend that caution be taken when using the average risk approach to estimate PAF.

ARTICLE SUMMARY

Strengths and limitations of this study

- This study examined the assumptions and validity of the average risk approach to estimate the PAF, which has not been explored previously.
- We used both simulated and real-world data to demonstrate the factors associated with the bias of the average risk approach.
- As an empirical study, our simulation could only analytically establish the direction of bias of this approach and discuss the magnitude of bias using a limited number of risk exposure distributions and RR functions.

INTRODUCTION

Population Attributable Fraction (PAF) is an important measure for estimating the burden of disease in a population that is causally attributable to an exposure. Since its first introduction, PAF has received substantial attention in the field of epidemiology [1]. Many advances have been made in different approaches to calculating PAF of single and multiple risk factors [2-6], in estimating the variance [7, 8] and in the interpretation of PAF [9-11]. There have also been many comprehensive projects, either nationwide or globally, in estimating PAF for the burden of disease associated with its risk factors [12-22]. The International Agency for Research on Cancer has specialized in providing estimates of cancer surveillance and burden of cancer estimates from around the world. In 2007, Boffetta and colleagues [23] introduced an approach to estimating PAF when the prevalence data on a continuous exposure in the population under study are only available as a population average. This approach, to be referred to here as the "average risk approach", estimated the RR at average exposure of the whole population using the risk of disease per unit increase in exposure, and the average level of exposure of the whole population. No proof was provided at the time that this method was proposed. Hence, the purpose of this paper is to examine the underlying assumptions and validity of this average risk approach when estimating PAF for disease burden in a population. Specifically, we examined how the shape of the RR functions and the exposure distributions affect the validity of this approach.

METHODS

Description of Average Risk Approach

The average risk approach estimates the RR at an average exposure of the whole population using the RR of disease per unit increase in exposure along with the average level of exposure of the whole population as follows:

$$Risk = Exp^{[Ln(Risk\ per\ unit) \times average\ level\ of\ exposure]} = RR_{unit}^{\bar{x}} \qquad (1)$$

where *Risk* is the RR at the population average exposure, RR_{unit} is the RR associated with a unit increase in exposure, $\bar{x}$ is the weighted average level of exposure. An underlying assumption with this method is that a log-linear relationship exists between the exposure and the risk of cancer. The average risk approach then estimates PAF as:

$$PAF = \frac{Risk - 1}{Risk} \qquad (2)$$

where it was assumed that "each individual has experienced a similar average exposure" (IARC 2007, pg 5). Under this assumption, that all population under study are exposed at the population average level, formula (2) is a simplification of Levin's formula when the prevalence (*P*) is 100%:

$$PAF = \frac{P(RR - 1)}{1 + P(RR - 1)} \qquad (3)$$

Boffetta et al. stated that "This formula is valid when the risk of cancer per unit of exposure was estimated in a model using log transformation. This is the case for logistic regression or Poisson regression, which are models widely used in case-control and cohort studies respectively" (IARC 2007, pg 5). No proof was shown for this statement, although the authors went on to acknowledge that "the dose-effect relationship is, in fact, rarely linear (or log-linear) over the whole range of exposures, but this method is considered to be the best approximation available in this respect". Therefore, the validity of the average risk approach has not been fully assessed,

particularly concerning its sensitivity to departures from the assumed dose-response relationship, or concerning the impact of the exposure distribution.

When the distribution of a continuous exposure is known and no confounding is assumed, a valid method to estimate PAF involves integrating across all levels of exposure:

$$PAF = \frac{\int_{x=0}^{m} RR(x)P(x)dx - 1}{\int_{x=0}^{m} RR(x)P(x)dx} \quad (4)$$

where RR(x) is the RR at exposure x; P(x) is the population distribution of exposure; and m is the maximum exposure level. Note that if there were to be no bias in the average risk approach, the following equation would have to hold:

$$\int_{x=0}^{m} RR(x)P(x)dx = RR_{unit}^{\bar{x}} \quad (5)$$

Under the log-linear risk assumption, the left-hand side of equation (5) becomes:

$$\int_{x=0}^{m} RR_{unit}^{x}P(x)dx \quad (6)$$

Define $g(x) = RR_{unit}^{x}$ in which x is a random variable with distribution $P(x)$, (6) is E[$g(x)$], and the right-hand side of (5) is g[E(x)], because $g(x)$ is strictly convex (i.e., a line segment connecting any two points on the graph of a function lies above the graph) when RR_{unit} is greater than 1, the Jensen's inequality [24] determines that:

$$RR_{unit}^{\bar{x}} \leq \int_{x=0}^{m} RR_{unit}^{x}P(x)dx \quad (7)$$

According to (7), the average risk approach will not overestimate PAF. The magnitude of the bias is determined by the extent of the convexity of $g(x)$ over the effective range of x. When RR_{unit} is small (i.e., close to 1.00), $g(x)$ is approximately linear and there is little bias. However, whether or not the choice of the exposure distribution $P(x)$ affects the validity of this approach is unexplored. Specifically, it is unknown, if the exposure distribution in a population is strongly skewed or bimodal, whether or not the average risk approach still provides a good approximation to the actual PAF. Therefore, we studied the validity of the average risk approach under the loglinear RR function and a variety of exposure distributions.

In broad terms, when the loglinear function of RR is not assumed, the average risk approach can still be generalized as equation (2), in which "*Risk*" is the RR at the population average exposure level. It can be reasoned that the curvature of the RR function determines the direction and the magnitude of the bias. When RR is a linear function of the exposure (i.e., $RR(x) = 1 + k \cdot x, x \in [0,m]$), there is no bias, because the integral PAF ($\int_{x=0}^{m}(1+kx)P(x)dx$) and the average risk PAF ($1 + k\int_{x=0}^{m}xP(x)dx$) are equivalent. When the RR function has a convex form, which indicates a risk profile that increases more rapidly at higher levels of exposure, this approach underestimates PAF. In contrast, it overestimates PAF with a concave RR function, which indicates a risk profile that increases less rapidly at higher levels of exposure. To illustrate the latter point, we included two examples of simulated concave RR functions and calculated the bias of the average risk approach.

Investigation of Validity of Average Risk Approach

To investigate whether or not the validity of the average risk approach is affected by the exposure distribution, we simulated several exposure distributions where the exposure is

continuous, ranging between standardized values of 0 to 100, with 0 indicating no exposure and 100 indicating the maximal level of exposure in the population (Figure 1). The prevalence distributions were scaled so that the prevalence of all exposure levels summed to 100%. The details of the distributions are summarized in Table 1. We calculated PAF using both the average risk approach and by integrating across all exposure levels. We calculated the absolute bias ($PAF_{AvgRisk} - PAF_{Integral}$) and the relative bias $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$. Note that because PAF is often expressed as a percentage, their absolute and relative biases are both in percentage units. However, the meaning of the former is the percentage points, and the meaning of the latter is an actual percentage. For example, an absolute bias of -5% from the difference of $PAF_{AvgRisk}$ of 15% and $PAF_{Integral}$ of 20% indicates a relative bias of -25%.

To examine if the magnitude of risk affects the validity of the average risk approach, we tested a range of values for the RR per standardized unit, from 1.001 to 1.04. Using a standardized unit resolves the scaling issue of the unit. For example, the RR of standardized unit of body mass index and the disease associated with obesity is the same for the RR of 1 kg/m^2 or 5 kg/m^2, as long as it pertained to a single population. In this study, we refer to RR per standardized unit as "RR per unit", unless otherwise stated. We also considered that the risk becomes implausible for RR per unit values above 1.04. For example, the RR at maximal exposure level would be 132, if the RR per unit is 1.05 under the log-linear assumption.

In addition, we illustrated the bias of the average risk approach when the RR function is non-linear or loglinear. In particular, we used two simulated examples of quadratic and cubic spline RR functions, which are both concave (Figure S1). The quadratic RR function has a form of $RR(x) = 3 - 2(\frac{x}{k} - 1)^2$, in which $k \in [\frac{m}{2}, m]$. This quadratic form has RR=1 when x=0, and

RR has a maximum of 5 when $x = k$. In the illustrated example, we used k = 75, i.e., 75% of the maximal exposure. The cubic spline RR function is based on simulated data, with the function being approximately quadratic in the lower exposure range, and approximately linear at higher exposures.

Finally, we used real-world data of the distribution of air pollution ($PM_{2.5}$) and residential radon exposures, which were investigated in the Canadian Population Attributable Risk of Cancer (ComPARe) study. The ComPARe study collected national-representative and population-weighted exposure data of $PM_{2.5}$ and residential radon and used the integral approach to estimate PAF of lung cancer for 2015 for Canada.[25, 26] We compared this PAF to that obtained using the average risk approach, to illustrate the validity of this approach. We also estimated the approximate 95%CI of the PAFs and the bias, assuming a fixed prevalence distribution for simplicity and a lognormal distribution of the RR. We resampled 10,000 RRs from this distribution and calculated PAF and bias. We used the 2.5% and 97.5% quantiles as the approximate 95%CI.

Patient or public involvement

No patients involved.

RESULTS

First, we examined the bias of the average risk approach under the loglinear RR function with the exposure distributions we selected in Table 1. The results at RR per unit of 1.001, 1.01 and 1.03 were illustrated in Table 2 and the results with a range of RR per unit from 1.001 to 1.04 were shown in Figure 2. At RR of 1.001, the absolute and relative biases were very small

and the average risk approach can be regarded unbiased. At RR of 1.01, the absolute bias remained small for all tested distributions although the relative bias started to increase substantially in the power distribution and in the Poisson distribution with an extreme tail (Table 2). At RR of 1.03, large absolute and relative biases were observed in several distributions. However, the normal and hypergeometric distributions were more robust than the Poisson with extreme tail and power distributions with the increase in RR (Table 2, Figure 2). For some distributions (uniform, beta (0.5, 0.5), beta (8, 2), and bimodal), the largest absolute and relative bias occurred at an intermediate value of RR (Figure 2). As RR increases, the bias becomes smaller, because the PAF estimates approaches 100%. Regardless of the exposure distribution and the magnitude of RR, the direction of the bias is underestimation in the case of loglinear RR.

We then illustrated the direction of the bias when the RR function is concave. Table 3 showed the resulting bias of the two RR functions in Figure S1 when the exposure distributions were as reported in Table 2. With concave RR functions, the direction of the bias in the average risk approach is overestimation. Similar to the loglinear RR function, we observed little bias in normal, hypergeometric, and beta(8, 2) distributions, whereas substantial bias was observed in power, Poisson with extreme tail, and beta(0.5, 0.5) distributions.

Finally, we explored the bias of the average risk approach using real-world data for air pollution ($PM_{2.5}$) and residential radon. Epidemiologic studies support a loglinear RR function between exposure to residential radon and lung cancer [27, 28]. A loglinear dose response between PM2.5 and lung cancer risk was less consistent. The loglinear relationship was supported by several studies [29-32], while two studies reported some deviation from it [33, 34]. The 2019 Global Burden of Disease Study of 87 risk factors suggested that PM2.5 has a loglinear relation with lung cancer in low exposure range (0-50 ug/m3) and a linear relation in

high exposure range (>50 ug/m3) [16]We assumed a loglinear relation for PM2.5 because the level is typically below 20 ug/m3 in Canada. We found that both exposures had skewed distributions (Figure S2). The $PM_{2.5}$ distribution had a long left tail, while the distribution of residential radon has a long right tail. We standardized the exposure levels of $PM_{2.5}$ and radon to 0.14 ug/m^3 and 7.4 Bq/m^3 per unit, so that the maximal exposure level is 100 units. The RR per unit of $PM_{2.5}$ associated with lung cancer was 1.0012 (95%CI: 1.0008 to 1.0016). The PAFs of $PM_{2.5}$ using the integral and the average risk approach were 6.89% (95%CI: 4.71% to 8.98%) and 6.87% (95%CI: 4.70% to 8.95%), respectively, indicating very small bias in the average risk approach (-0.02%, 95%CI: -0.03% to -0.01%). The RR per unit of radon associated with lung cancer was 1.011 (95%CI: 1.005 to 1.016). The PAFs of radon using the integral and average risk approach were 6.87% (95%CI: 3.33% to 10.52%) and 6.37% (95%CI: 3.21% to 9.37%), respectively. The bias was larger than that seen in $PM_{2.5}$. The absolute bias was -0.5% (95%CI: -1.2% to -0.1%) and the relative bias was -7.3% (95%CI: -11.0% to -3.5%), indicating slight to moderate bias. The observations were consistent with the simulations, in that small RRs yield little bias ($PM_{2.5}$), and moderate to large RRs could produce bias with some skewed exposure distributions (radon).

DISCUSSION

Since being introduced by Boffetta and colleagues in 2007, the average risk approach has been used in several PAF estimation projects.[12-15, 35] In addition to the cancer burden study in France,[15] the ComPARe study in Canada,[35] a study of attributable causes in China,[12] and two studies in Brazil[13, 14] have used this method. We illustrated that the direction of bias of the average risk approach is determined by whether the RR function is convex or concave,

while the magnitude of bias is affected by the degree of convexity or concavity, as well as the exposure distribution. When the RR per unit is small under a loglinear RR function, the magnitude of bias is also small and the average risk approach is approximately valid. With larger RR and increased convexity, the validity of the average risk approach would also depend on the exposure distribution. We demonstrated that under some circumstances (e.g., Poisson distribution with extreme tail, power distribution), the approach could potentially lead to moderate to severe bias.

The average risk approach has an implicit assumption that the minimal risk exposure value is 0. When the minimal risk exposure value is not 0, this approach generates invalid estimates. To offer a simplified example, overweight and obesity defined as body mass index (BMI) $\geq$ 25.0 kg/m^2 is associated with postmenopausal breast cancer. The minimal risk exposure value of BMI is 25.0 kg/m^2. Assuming a log-linear relationship between BMI above 25.0 kg/m^2 and the risk of breast cancer and that a postmenopausal female population has a normal distribution of body mass index (BMI) at a mean and standard deviation of 25.0 and 5.0 kg/m^2. The average risk approach yields a PAF of 0 in this population, because the population average risk exposure is 25.0 kg/m^2, which has a RR of 1.0. Although it is possible to recode the exposure so that the minimal exposure is zero, a new average of the recoded exposure must be estimated, which requires the information of the exposure distribution. On the other hand, the prerequisite of applying the average risk approach is that such information is only available as a population average. In practice, many natural or physiological exposures have a non-zero minimal risk exposure value and the estimation of PAF for such exposures requires additional considerations [36]. Therefore, this implicit assumption is a substantial limitation of this

approach. For the same reason, the average risk approach cannot be applied in the framework of generalized impact fraction, in which the impact of partial reduction of exposure is considered.

Our study has some limitations. First, this study is an empirical examination of the validity of the average risk approach. We have mathematically demonstrated the direction of the bias in this approach. However, we only qualitatively discussed the magnitude of the bias associated with the RR function and the exposure distribution. We illustrated the magnitude of the bias through several RR functions and exposure distributions. However, this pragmatic approach could not cover all RR functions and distributions. Second, we compared the average risk approach to the integral approach under the assumption of no confounding. The integral approach is an extension of Levin's formula, which is biased in the presence of confounding [1, 11]. Ideally, the validity of the average risk approach should be tested against the integral form of Miettinen's formula, which is based on the prevalence of exposure among the cases and is valid in the presence of confounding [6]. However, because the average risk approach was developed under the framework of Levin's formula, we considered that a comparison of two approaches under the same framework would be more appropriate. Nevertheless, it should be noted that the validity of the average risk approach is also prone to the presence of confounding, just like Levin's formula.

In conclusion, we have shown that the average risk approach has some utility, but nonetheless carries the risk of bias. This approach should not be used when the minimal exposure level is not zero. We recommend using approaches with smaller risk of bias, such as the integral approach, to estimate PAF when the information regarding the RR function and the exposure distribution data are available.

FOOTNOTES

Contributors: YR participated in study conceptualization, statistical analyses, drafted the initial manuscript and approved the final version of the manuscript. SW participated in study conceptualization, supervision, and critically reviewed and edited the manuscript. PG provided resources (ComPARe datasets), critically reviewed and edited the manuscript, and approved the final version of the manuscript. CF participated in funding acquisition, supervision, critically reviewed and edited the manuscript, and approved the final version of the manuscript. DB participated in funding acquisition, supervision, and approved the final version of the manuscript.

Funding: This study was supported by the Canadian Cancer Society Partner Prevention Research Grant (grant #703106).

Competing interests: None declared.

Patient consent for publication: Not required.

Ethics approval: Not applicable. This is a simulation and secondary data analysis study based on both simulated data and population-based aggregated data. Human participants are not involved.

Data sharing statement: Extra data, including the R code for simulation and the exposure datasets from the ComPARe study, are available by emailing to yibing.ruan@albertahealthservices.ca.

Table 1. Description of the exposure distributions used in this study.

Distribution	**Note**
Uniform	Range from 0 to 100
Normal	$\mu=50$, $\sigma=10$
Log-normal	$\mu=5$, $\sigma=0.5$
Hypergeometric	N=700, K=200, m=200
Beta	$\alpha=0.5$, $\beta=0.5$
Beta	$\alpha=2$, $\beta=8$
Beta	$\alpha=8$, $\beta=2$
Bimodal	Constructed by combining the lognormal distribution ($\mu=5$, $\sigma=0.5$) with one-third of beta (8, 2).
Poisson with extreme tail	Constructed by applying the Poisson distribution (k = 0 to 3, $\lambda=1$) to exposure level 0 to 3, and one-tenth of the Poisson distribution (k = 70, 75, 80, 85, 90, $\lambda=80$) to exposure level 95 to 99
Power	Constructed by rescaling the function of 1/x, where $x \in [0.1, 2.5]$.

Note: All distributions were scaled to ensure that the sum of distribution is 100%.

Table 2. Absolute and relative bias in PAF between the average risk approach and the integration approach in selected exposure distributions when RR per unit is 1.001, 1.01 or 1.03 for the loglinear function

RRunit	1.001				1.01				1.03			
Distribution	PAF_{Integr}	PAF_{Avera}	Absolute bias	Relative bias	PAF_{Integr}	PAF_{Avera}	Absolute bias	Relative bias	PAF_{Integr}	PAF_{Avera}	Absolute bias	Relative bias
Uniform	4.9%	4.8%	0%	-0.9%	41.4%	38.9%	-2.6%	-6.2%	83.8%	76.8%	-7%	-8.3%
Normal	4.8%	4.8%	0%	-0.1%	38.9%	38.6%	-0.3%	-0.8%	77.6%	76.6%	-1%	-1.3%
Log-normal	3.1%	3.0%	0%	-0.7%	28.3%	26.5%	-1.7%	-6.1%	68.3%	60.0%	-8.4%	-12.3%
Hypergeometric	4.3%	4.3%	0%	0%	35.3%	35.3%	0%	0%	72.6%	72.6%	0%	0%
Beta(0.5, 0.5)	4.9%	4.8%	-0.1%	-1.1%	42.3%	38.9%	-3.4%	-7.9%	85.3%	76.9%	-8.5%	-9.9%
Beta(2, 8)	1.8%	1.8%	0%	-0.4%	17.2%	16.5%	-0.7%	-3.8%	45.7%	41.5%	-4.2%	-9.2%
Beta(8, 2)	7.8%	7.8%	0%	-0.1%	55.7%	55.3%	-0.3%	-0.6%	91.4%	90.9%	-0.5%	-0.6%
Bimodal	4.3%	4.3%	0%	-0.9%	37.9%	35.1%	-2.7%	-7.2%	81.1%	72.3%	-8.7%	-10.8%
Poisson with extreme tail	0.6%	0.6%	0%	-3.4%	8.3%	5.9%	-2.4%	-29.3%	42.7%	16.5%	-26.2%	-61.4%
Power	2.6%	2.6%	0%	-1.4%	25.9%	22.8%	-3%	-11.6%	69.0%	53.7%	-15.2%	-22.1%

Note: the absolute bias is $PAF_{AvgRisk} - PAF_{Integral}$ and the relative bias is $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

Table 3. Absolute and relative bias in PAF between the average risk approach and the integration approach in two illustrated examples of concave RR functions.

RR funtion	Cubic spline				Quadratic			
Distribution	$PAF_{Integral}$	$PAF_{Average\ risk}$	Absolute bias	Relative bias	$PAF_{Integral}$	$PAF_{Average\ risk}$	Absolute bias	Relative bias
Uniform	49.0%	52.8%	3.8%	7.8%	59.6%	64.0%	4.4%	7.4%
Normal	52.6%	52.8%	0.1%	0.3%	63.6%	64.1%	0.5%	0.7%
Log-normal	46.4%	51.3%	4.9%	10.6%	54.3%	57.3%	3%	5.6%
Hypergeometric	52.6%	52.7%	0.1%	0.1%	62.4%	62.5%	0.1%	0.2%
Beta(0.5, 0.5)	46.9%	52.8%	5.8%	12.4%	57.7%	64.0%	6.3%	11%
Beta(2, 8)	40.9%	43.9%	3.1%	7.6%	45.9%	47.4%	1.5%	3.3%
Beta(8, 2)	53.1%	53.1%	0%	0%	65.9%	66.5%	0.6%	0.9%
Bimodal	48.3%	52.7%	4.4%	9.2%	57.9%	62.4%	4.5%	7.8%
Poisson with extreme tail	6.1%	11.1%	5%	81%	8.5%	13.6%	5.1%	60.6%
Power	38.7%	49.1%	10.4%	26.9%	47.1%	53.4%	6.4%	13.5%

Note: the absolute bias is $PAF_{AvgRisk} - PAF_{Integral}$ and the relative bias is $(PAF_{AvgRisk} - PAF_{Integral})/PAF_{Integral} \times 100\%$.

Figure legends:

Figure 1: Probability density curves of selected distributions in this study.

Figure 2: The absolute and relative bias of the average risk approach under the selected distributions and a range of RR per unit. Both absolute and relative bias are presented as a percentage. The absolute bias is the difference in PAF percentage, and the relative bias is the difference in PAF over the PAF using integration and expressed as a percentage.

Figure S1: Graph of the two concave RR functions used in this study to illustrate the direction and the magnitude of bias of the average risk approach

Figure S2: The smoothed density plot of the distributions of $PM_{2.5}$ and residential radon in Canada

REFERENCES

1. Levin, M.L., *The occurrence of lung cancer in man.* Acta Unio Int Contra Cancrum, 1953. **9**(3): p. 531-41.
2. Bruzzi, P., et al., *Estimating the population attributable risk for multiple risk factors using case-control data.* Am J Epidemiol, 1985. **122**(5): p. 904-14.
3. Eide, G.E. and I. Heuch, *Average attributable fractions: a coherent theory for apportioning excess risk to individual risk factors and subpopulations.* Biom J, 2006. **48**(5): p. 820-37.
4. Walter, S.D., *The estimation and interpretation of attributable risk in health research.* Biometrics, 1976. **32**(4): p. 829-49.
5. Whittemore, A.S., *Statistical methods for estimating attributable risk from retrospective data.* Stat Med, 1982. **1**(3): p. 229-43.
6. Miettinen, O.S., *Proportion of disease caused or prevented by a given exposure, trait or intervention.* Am J Epidemiol, 1974. **99**(5): p. 325-32.
7. Benichou, J. and M.H. Gail, *Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models.* Biometrics, 1990. **46**(4): p. 991-1003.
8. Greenland, S., *Variance estimators for attributable fraction estimates consistent in both large strata and sparse data.* Stat Med, 1987. **6**(6): p. 701-8.
9. Mansournia, M.A. and D.G. Altman, *Population attributable fraction.* BMJ, 2018. **360**: p. k757.
10. Di Maso, M., et al., *Attributable fraction for multiple risk factors: Methods, interpretations, and examples.* Stat Methods Med Res, 2020. **29**(3): p. 854-865.
11. Rockhill, B., B. Newman, and C. Weinberg, *Use and misuse of population attributable fractions.* Am J Public Health, 1998. **88**(1): p. 15-9.
12. Wang, J.B., et al., *Attributable causes of cancer in China.* Ann Oncol, 2012. **23**(11): p. 2983-2989.
13. Azevedo, E.S.G., et al., *The Fraction of Cancer Attributable to Ways of Life, Infections, Occupation, and Environmental Agents in Brazil in 2020.* PLoS One, 2016. **11**(2): p. e0148761.
14. Rezende, L.F. and J. Eluf-Neto, *Population attributable fraction: planning of diseases prevention actions in Brazil.* Rev Saude Publica, 2016. **50**.
15. Boffetta, P., et al., *The causes of cancer in France.* Ann Oncol, 2009. **20**(3): p. 550-5.
16. Collaborators, G.B.D.R.F., *Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet, 2020. **396**(10258): p. 1223-1249.
17. Parkin, D.M., L. Boyd, and L.C. Walker, *16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010.* Br J Cancer, 2011. **105 Suppl 2**: p. S77-81.
18. Whiteman, D.C., et al., *Cancers in Australia in 2010 attributable to modifiable factors: summary and conclusions.* Aust N Z J Public Health, 2015. **39**(5): p. 477-84.
19. Diseases, G.B.D. and C. Injuries, *Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.* Lancet, 2020. **396**(10258): p. 1204-1222.
20. Poirier, A.E., et al., *The current and future burden of cancer attributable to modifiable risk factors in Canada: Summary of results.* Prev Med, 2019. **122**: p. 140-147.
21. Arnold, M., et al., *Global burden of cancer attributable to high body-mass index in 2012: a population-based study.* Lancet Oncol, 2015. **16**(1): p. 36-46.
22. Islami, F., et al., *Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States.* CA Cancer J Clin, 2018. **68**(1): p. 31-54.

23. IARC, *Attributable Causes of Cancer in France in the Year 2000*, in *IARC Working Group Report Volume 3*. 2007.
24. Durrett, R., *Probability: Theory and Examples*. 5th ed. ed. 2019: Cambridge University Press.
25. Gogna, P., et al., *Estimates of the current and future burden of lung cancer attributable to PM2.5 in Canada.* Prev Med, 2019. **122**: p. 91-99.
26. Gogna, P., et al., *Estimates of the current and future burden of lung cancer attributable to residential radon exposure in Canada.* Prev Med, 2019. **122**: p. 100-108.
27. Krewski, D., et al., *A combined analysis of North American case-control studies of residential radon and lung cancer.* J Toxicol Environ Health A, 2006. **69**(7): p. 533-97.
28. Puskin, J.S., *Perspective on the use of LNT for radiation protection and risk assessment by the U.S. Environmental Protection Agency.* Dose Response, 2009. **7**(4): p. 284-91.
29. Hystad, P., et al., *Long-term residential exposure to air pollution and lung cancer risk.* Epidemiology, 2013. **24**(5): p. 762-72.
30. Lepeule, J., et al., *Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009.* Environ Health Perspect, 2012. **120**(7): p. 965-70.
31. Puett, R.C., et al., *Particulate matter air pollution exposure, distance to road, and incident lung cancer in the nurses' health study cohort.* Environ Health Perspect, 2014. **122**(9): p. 926-32.
32. Turner, M.C., et al., *Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers.* Am J Respir Crit Care Med, 2011. **184**(12): p. 1374-81.
33. Crouse, D.L., et al., *Ambient PM2.5, O(3), and NO(2) Exposures and Associations with Mortality over 16 Years of Follow-Up in the Canadian Census Health and Environment Cohort (CanCHEC).* Environ Health Perspect, 2015. **123**(11): p. 1180-6.
34. Pinault, L., A. van Donkelaar, and R.V. Martin, *Exposure to fine particulate matter air pollution in Canada.* Health Rep, 2017. **28**(3): p. 9-16.
35. Ruan, Y., et al., *Estimates of the current and future burden of cancer attributable to red and processed meat consumption in Canada.* Prev Med, 2019. **122**: p. 31-39.
36. Ferguson, J., et al., *Population attributable fractions for continuously distributed exposures.* Epidemiologic Methods, 2020. **9**(1).