A Polygon-Based Locally-Weighted-Average Method for Smoothing Disease Rates of Small Units.
Xun Shi; Eric Duell; Eugene Demidenko; Tracy Onega; Benjamin Wilson; David Hoftiezer

Shiro Tanaka; Yutaka Matsuyama; Masataka Shiraki; Yasuo Ohashi

Estimating the Longitudinal Prevalence of Diarrhea and Other Episodic Diseases: Continuous Versus Intermittent Surveillance.
Wolf-Peter Schmidt; Stephen P. Luby; Bernd Genser; Mauricio L. Barreto; Thomas Clasen

Potential Confounding by Exposure History and Prior Outcomes: An Example From Perinatal Epidemiology.
Penelope P. Howards; Enrique F. Schisterman; Patrick J. Heagerty

The Importance and Role of Intracluster Correlations in Planning Cluster Trials.
John S. Preisser; Beth A. Reboussin; Eun-Young Song; Mark Wolfson

Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs.
Tyler J. VanderWeele; James M. Robins

Can DAGs Clarify Effect Modification?
Clarice R. Weinberg

Are Girls More Susceptible to the Effects of Prenatal Exposure to Tobacco Smoke on Asthma?
Jouni J. K. Jaakkola; Mika Gissler

Prenatal Ultrasound Scanning and the Risk of Schizophrenia and Other Psychoses.
Karin Stålberg; Bengt Haglund; Ove Axelsson; Sven Cnattingius; Christina M. Hultman; Helle Kieler

Belated Concerns and Latent Effects: The Example of Schizophrenia.
Michaeline Bresnahan; Ezra Susser
### Original Article

585  **Ambient Air Pollution and Cardiac Arrhythmias in Patients With Implantable Defibrillators.**
Kristina B. Metzger; Mitchel Klein; W Dana Flanders; Jennifer L. Peel; James A. Mulholland; Jonathan J. Langberg; Paige E. Tolbert

593  **Short-Term Effects of Particulate Air Pollution on Male Smokers and Never-Smokers.**
Chit-Ming Wong; Chun-Quan Ou; Nga-Wing Lee; King-Pan Chan; Thuan-Quoc Thach; Yuen-Kwan Chau; Sai-Yin Ho; Anthony Johnson Hedley; Tai-Hing Lam

599  **Body Mass Index in Adolescence and Number of Children in Adulthood.**
Markus Jokela; Mika Kivimäki; Marko Elovainio; Jorma Viikari; Olli T. Raitakari; Liisa Keltikangas-Järvinen

607  **Body Mass Index and Lung Cancer Risk in Women.**
Geoffrey C. Kabat; Anthony B. Miller; Thomas E. Rohan

613  **Pet Ownership and Blood Pressure in Old Age.**
Joel David Wright; Donna Kritz-Silverstein; Deborah J. Morton; Deborah L. Wingard; Elizabeth Barrett-Connor

### Brief Report

619  **Does Low Maternal Blood Pressure During Pregnancy Increase the Risk of Perinatal Death?**
Aimin Chen; Olga Basso

### Original Article

623  **How Much Would Closing Schools Reduce Transmission During an Influenza Pandemic?**
Kathryn Glass; Belinda Barnes

629  **Cigarette Smoking and Risk of Breast Carcinoma In Situ.**
Amy Trentham-Dietz; Hazel B. Nichols; Kathleen M. Egan; Linda Titus-Ernstoff; John M. Hampton; Polly A. Newcomb

639  **Cigarette Smoking and Lung Cancer: Modeling Effect Modification of Total Exposure and Intensity.**
Jay H. Lubin; Neil Caporaso; H Erich Wichmann; Angelika Schaffrath-Rosario; Michael C. R. Alavanja

### Commentary

649  **Models of Smoking and Lung Cancer Risk: A Means to an End.**
Jonathan M. Samet; Michael J. Thun; Amy Berrington de Gonzalez

### Remembrance

652  **Stephan Karl Maria Weiland, 1958-2007.**
Ulrich Keil
A Polygon-Based Locally-Weighted-Average Method for Smoothing Disease Rates of Small Units

Xun Shi,* Eric Duell,† Eugene Demidenko,† Tracy Onega,† Benjamin Wilson,* and David Hoftiezer*

Background: Disease rates for geographic areas with small populations may be unstable. Therefore, accurate nonparametric methods for smoothing or stabilizing rates are needed.

Methods: We propose an innovative locally-weighted-average method as an easy tool for disease surveillance. Our approach has several important advantages over existing locally-weighted-average methods. One advantage is that the buffer zone is created based on a polygon rather than centroid. Second, the buffer distance is determined by a user-specified population threshold. Third, a weighting factor that accounts for variability in the rate is used in the smoothing process. We further propose a variance-driven procedure to reduce arbitrariness in selecting the population threshold, and a binary search technique to quickly and precisely find the buffer distance according to the specified population threshold. Lastly, we develop a software tool using ArcObjects® (ESRI, Redland, CA) to implement this method.

Results: Our method was applied to town-level lung cancer incidence rates for New Hampshire. A comparison with a traditional point-based method indicated that our method produced less under- and over-smoothing.

Conclusion: Our method and the software tool are suitable for researchers and public health workers who want to apply geographic information systems to map smoothed disease rates for exploratory purposes.

(Epidemiology 2007;18: 523–528)

For privacy reasons, disease incidence and mortality data are usually available only at the aggregate level (eg, at the level of zip code zones, census tracts, towns, etc.). Such data are often used to estimate and map disease rates for these geographic areas to show spatial patterns. A major problem, however, is the instability of rates for regions with small populations; a small change in the number of cases may cause the rates to change drastically. One solution is to incorporate information from other units to stabilize the rate. This type of operation tends to reduce differences between nearby units and is considered a smoothing process. Model-based methods, such as hierarchical and Bayesian methods, have been developed for this purpose.1–7 Unfortunately, the use of these methods requires substantial statistical skills, which impedes their widespread use. In addition, few if any of these methods have been integrated into commonly used commercial geographic information systems (GIS) software packages.

Locally-weighted-average methods smooth the rate of an individual unit by averaging all the rates of the units in a region around the unit being smoothed; we refer to this special region as the “neighborhood” of the unit being smoothed. In this process, each rate is weighted by its population.8 The effects of locally-weighted-average and model-based methods on individual spatial units are shown in Figure 1. Both types of methods behave similarly in that they apply more smoothing to those units with high variance (small populations). However, in locally-weighted averaging, some units below a certain threshold will not be included in smoothing. The threshold can be set according to spatial extent, population size, or rate variation associated with the spatial units. Determining this threshold in traditional locally-weighted averaging is highly subjective and therefore less statistically rigorous. However, the model-based methods are also not immune to subjectivity. Those models include assumptions that may result from specifying the underlying distribution (eg, Poisson distribution), setting prior probabilities, using models relying on random effects, and choosing the spatial correlation matrix. Strictly speaking, locally-weighted averaging is distribution-free, and hence may be less subject to the assumptions of model-based methods. Using simulated data in which the true spatial variation was known, Kafadar8 found locally-weighted averaging to be more accurate (in terms of preserving the true variation and smoothing over artificially elevated rates) than the other tested methods including a local polynomial regression method (the loess method9) and an empirical Bayes method.10 In practice, locally-weighted averaging is easier to conceptualize and implement than model-based methods. Additionally, this approach can be readily incorporated into commercial GIS software. Therefore, locally-weighted averaging offers an alternative for researchers and public health workers to use in routine applications, especially in settings where a preliminary exploration of possible spatial patterns is desired.

In this paper, we present a polygon-based locally-weighted-average method, which is an improvement to the traditional point-based locally-weighted-average method. We implement our method using a commercial GIS package,
Data and Rate Calculation

Tumor registry records of 10,500 incident lung cancer cases in New Hampshire from 1989 through 2002 were obtained from the New Hampshire State Cancer Registry. U.S. Census population counts from 1990 and 2000 for New Hampshire were used to estimate the population for each New Hampshire town. Age-adjusted lung cancer incidence rates were calculated for each New Hampshire town using a conventional direct standardization method.\(^8\) The calculated rate is a yearly average over a 14-year period (1989–2002). The equation for the calculation is as follows:

\[
r_i = \frac{1}{M} \sum_{m=1}^{M} \frac{N_m}{n_{im}} r_{im} \sum_{m=1}^{M} N_m
\]

In this equation, \(r_i\) is the age-adjusted rate of town \(i\); \(m\) represents an age category, and \(M\) is the total number of the categories. Thus, \(y_{im}\) is the case count in age category \(m\) in town \(i\); \(N_m\) is the total number of people in \(m^{th}\) age category in New Hampshire; \(n_{im}\) is the number of people in the \(m^{th}\) age category in town \(i\); and \(k\) is the number of years for which the data are available. We refer to the resulting rates as the “original rates.”

Age adjustment of lung cancer rates was performed over the following age categories: 30–39, 40–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75+. People under 30 years of age were excluded because of the small number of lung cancer cases in that age group.\(^11\) Because we were interested in estimating rates over a 14-year period, we used both the 1990 and 2000 U.S. Census data in the rate calculation. Specifically, for cases reported in the period of 1989–1995 we used the 1990 Census data to derive the standard population, and for cases in the period of 1996–2002 we used the 2000 data.

Smoothing Method

The general process of locally-weighted averaging can be represented as follows:\(^8\):

\[
\tilde{r}_i = \frac{\sum_{j \in N(i)} w_j r_j}{\sum_{j \in N(i)} w_j}
\]

Within the context of our study, \(\tilde{r}_i\) is the smoothed rate of town \(i\) (which we refer to as the “index town”); \(r_j\) is the original rate of town \(j\); \(w_j\) is the weight of \(r_j\) in the smoothing; and \(\sum_{j \in N(i)}\) means summation over towns in the neighborhood of town \(i\) (which we refer to as the “neighboring towns”). To conduct smoothing for a given index town, the user needs to specify (a) how to determine a neighboring town, and (b) how to calculate the weight for a neighboring town. A simple locally-weighted-average method described by Waller and Gotway\(^10\) uses a constant distance to create a buffer around the centroid of the index town, as shown in Figure 2A. Those towns whose centroids fall into the buffer are considered to be “neighboring towns.” We will refer to this method, as the point-based method, because it assumes that all disease cases and all people in a town are concentrated on the centroid. Using this method, whether a town is a neighboring town of the index town is determined solely by the geographical distance between the centroids of the 2 towns, disregarding how much of the area of the former town is actually inside or outside the buffer of the index town. This strategy generates 2 unwanted consequences. First, the buffer zone always has a circular shape and ignores the actual shape of the index town. Thus, the circular buffer favors those towns whose centroids happen to be closer to the centroid of the index town and may lead to a significant directional bias, especially when the shapes of the towns are highly irregular. Second, if a town is deemed a neighboring town, its entire population will be counted in the weight calculation, regardless of the proportion of the population that actually lives close to the index town. The point-based strategy has been adopted by some researchers for disease mapping and cluster detection.\(^12,13\)

In contrast, our method incorporates town geometry, as displayed in Figure 2B. Rather than assuming that all disease cases and people are concentrated on a single point, we assume that cases and the population are evenly distributed across the town. Based on this assumption, our method

---

A. Model-based method

B. Locally-weighted-average method

FIGURE 1. Comparison of the effects of different smoothing methods on individual spatial units.

ArcGIS\(^\text{®}\) (ESRI, Redland, CA), and demonstrate its features by applying it to town-level lung cancer incidence data from the US state of New Hampshire.

METHODS

A. Point-based LWA

B. Polygon-based LWA

FIGURE 2. Neighborhood defined by buffering polygon.
creates a buffer around the polygon, rather than around the centroid. One advantage of this polygon-based method is that it includes the surrounding towns in all directions in a geographically uniform fashion. The relevance of a town to the index town is not “all or none,” but rather is determined by the number of individuals in that town who are within the buffer of the index town. This number is estimated based on the proportion of the town’s area falling into the buffer of the index town, assuming an even distribution. For example, referring to Figure 2B, the number of individuals in town B who are within the buffer of town A, \( n_{AB} \), is calculated as follows:

\[
 n_{AB} = n_B \ast (a_{AB}/a_B)
\]

where \( n_B \) is the total population of B, \( a_{AB} \) is the area of the part of B that is within the buffer of A (the darker area), and \( a_B \) is the total area of B.

Our method also takes into account the variability of a rate when weighting each rate:

\[
w_j = n_j/c_j
\]

where \( n_j \) is the number of people in town \( j \) that fall into the buffer of town \( i \), and \( c_j \) is the variability of the original rate \( r_j \). The more individuals in \( j \) close to \( i \), the greater the weight for \( r_j \), while the larger the variability of \( r_j \), the less the weight.

In this study, the variability of the rate for a town was represented by the temporal fluctuation of annual case counts for that town. Specifically, since we have annual case-count data for each town during a 14-year period, we were able to use 14 count values to quantify the variability for each town. The specific statistic we used is coefficient of variation, calculated by dividing the standard deviation by the mean. We used the temporal variability to represent general variability, because the annual case-counts for multiple years can be considered as the results of repetitive sampling of the same population, and therefore can be used to calculate empirical variability. When data for calculating such empirical variability are not available, variability derived from probability distribution models can be used. In this study, we calculated the empirical variability as well as the variability derived from the binomial model, so as to assess the comparability of the 2.

Instead of directly specifying a buffer distance, our method uses GIS to derive the buffer distance according to a user-specified population threshold. With this approach, only towns whose populations are smaller or equal to the threshold will be smoothed. An advantage of the population-determined neighborhood is that the user can control the minimum population base for calculating (smoothing) the rates to control the upper limit of variances of the rates. In other words, the user can easily specify the desired stability for the towns with small populations without concern about over-smoothing the towns with bigger populations. The population-determined neighborhood has been used by Talbot et al\(^{12}\) to create smoothed maps of low birth weight in New York State and by Turnbull et al\(^{13}\) to detect clusters of leukemia cases in upstate New York.

The key to our population-based method is the choice of the population threshold. This choice is inherently present in all nonparametric statistics, and in locally-weighted-average methods particularly.\(^{14}\) Here we propose a method to facilitate this choice based on the idea of stabilizing threshold. This process is based on the observation that the variance of the entire area (ie, variance across all towns) decreases as the threshold increases, and that there may be a threshold above which the variance becomes relatively stable. We suggest using this “stabilizing” threshold for smoothing. The rationale of this suggestion is as follows. Disease incidence rates possess 2 variation components: the variation due to natural causes, such as environmental, demographic, and other unadjusted factors, and the variation representing the instability due to a small population. The goal is to preserve the former and reduce the latter. Although the variation due to natural causes in irreducible, the latter source of variation can be substantially eliminated by increasing the population size via expanding the buffer zone. In the proposed process, we assume that the relatively stable variance beyond the “stabilizing” threshold represents the natural variation. Technically, the “stabilizing” threshold can be identified by locating where the rate of variance decrease reaches a minimum. Figure 3 illustrates the results from this process in our case study.

To identify the threshold for the town-level lung cancer incidence rates in New Hampshire, we tested a series of population sizes with small increments between 2 consecutive test sizes. We first used each of the population sizes to perform the smoothing, and calculated the variance of the resulting rates (Fig. 3A). Subsequently, we used a simple way to locate the minimum rate of variance decrease: we calculated the difference between 2 consecutive variances and found the minimum difference (Fig. 3B). The threshold identified in this way for the New Hampshire data is 3,000 individuals. While we propose this variance-driving process as an approach to reducing arbitrariness in traditional LWA, we are aware of the assumptions it requires: first, the true variation is much more stable in smoothing than the variation due to small population size; and second, the relatively stable variance, if identifiable, correctly represents the true variation. When the factors causing the true variation are highly localized (with small influence ranges) and coincident to the areas with small populations, the effects of those factors may be smoothed out along with the variation due to population size, which makes those 2 assumptions questionable. Thus we suggest that this method be used with caution.

To derive a buffer distance according to the specified population size, we employed an iterative binary search algorithm. The algorithm starts with a small distance. If the distance is indeed too small, it jumps to a fairly large distance. If the second distance is too large, the algorithm tests the “midpoint” between the last tested small and large distance. If the second distance is indeed too small, it jumps to a fairly large distance. If the second distance is too large, the algorithm tests the “midpoint” between the last tested small and large distances. Depending on the middle distance being too small or too large, the next test distance will either be between the small and the middle distances or between the middle and the large distances. This algorithm can approach the desired distance quickly with high precision.

We implemented our method using the Visual Basic for Applications language and ArcObjects® of ArcGIS®. The output from the program is saved to a new field inserted into the attribute table of the original polygon Shapefile®. This field contains the smoothed rate for each polygon; these values can be
immediately mapped using ArcMap®. The graphical user interface of the program is shown in Figure 4. An ArcMap® Template (.mxt) containing the script can be downloaded as freeware from the authors’ website (www.dartmouth.edu/~xunshi).

For comparison, we also implemented the point-based method described by Waller and Gotway and applied it to the same New Hampshire lung cancer incidence data. After several trials, 12 km was selected as the buffer distance for the point-based method, because it gives a result that based on visual inspection, is comparable to the result from the 3000-people threshold. We also applied the 3000-people threshold to the point-based method, ie, both methods smoothed the rates only for towns with populations >3000. For towns on or near the border of New Hampshire both methods defined the neighborhoods only toward the interior of the state due to the absence of data for the areas surrounding the state.

**RESULTS**

Figure 5A shows a map of the town populations in New Hampshire and Figure 5B shows the original age-adjusted rate for each town. The smoothing results from the 2 methods are shown in Figures 5C and 5D. Figure 5B shows that some towns have zero rates. Most of these towns have very small populations and are concentrated in the northern part of the state. In this study, we specified that if a town had a zero population, it would be excluded from the smoothing operation. The rates for these zero-population towns remain at zero in the maps from both smoothing methods. Note that towns with nonzero populations but zero original rates may gain positive rate values from their neighbors through smoothing.

Here we use towns labeled as “A” and “B” in Figures 5C and 5D to illustrate the effects of the 2 smoothing methods.

**FIGURE 3.** Identifying the optimal threshold in population size through variance calculation. A, Variance is a decreasing function of the population size. B, The optimal threshold is where the difference between 2 consecutive variances is minimal (indicated by the white circle).
Town A has a high rate and a small population (population = 298). During the 14-year period, it had 6 lung cancer cases; thus, one more or one fewer case would cause a significant change to town A’s rate. The point-based method was not able to smooth (ie, stabilize) its rate, because all the towns whose centroids are within 12 km of town A’s centroid have even smaller populations and thus cannot lend much power. Further, town A’s unstable rate propagates to its tiny-population neighbors through the smoothing process and forms a visually striking but actually unstable “hot region” in northern New Hampshire (Fig. 5C). The polygon-based method, on the other hand, reduced the influence of town A considerably. After being smoothed using a 3,000-people base, the rate of town A is no longer among the highest in the state (Fig. 5D).

Town B illustrates the under-smoothing problem of the point-based method. Town B’s original rate is zero. Because the town has a small population (n = 257), the zero rate is unstable. However, due to its relatively large area, the 12-km circle around town B’s centroid cannot reach the centroids of any neighboring towns and thus its rate remains zero after the smoothing operation of the point-based method (Fig. 5C). In contrast, the polygon-based method was able to incorporate disease information from the surrounding towns and assign a rate to town B.

One could increase the buffer distance in the point-based method to include more neighboring towns for towns A and B. The tradeoff is that increasing the buffer will worsen the oversmoothing problems for other towns, eg, Town C. Town C has a population of 2321, which means its rate does not need much smoothing. The 12-km buffer distance of the point-based method, however, gives more emphasis to Town C’s low-rate neighbors, which considerably lowers its rate. In contrast, our polygon-based method does not noticeably alter that rate.

Finally, the towns’ 2 coefficients of variation, calculated using the actual annual counts and derived from the binomial model, correlate very well ($R^2 = 0.94$, number of towns = 259). This correlation indicates that when data are insufficient for calculating the empirical coefficient of variation, an approximation can be derived from the total number of cases and the population using the binomial model.

DISCUSSION

Our polygon-based, locally-weighted-average method has several advantages. First, the impact of towns surrounding the index town is determined spatially by the proportion of its area falling into the buffer of the index town. Second, the shape of the buffer defined for the index town specifically reflects the original shape of the town. Third, the population-determined buffer distance allows the user to control the population base for deriving the rate, and therefore is better able to assess the stabilities of the smoothed rates. Fourth information about the variability of a rate can be incorporated into the calculation of the weighting factor. The rationale for incorporating the variability information is obvious: if the rate of a neighbor is used in smoothing, the stability of that rate should also be considered.
To our knowledge, no previous study using locally-weighted-average methods have explicitly incorporated rate variability information, perhaps due to the lack of such information. The good match between the coefficient of variation values derived from the binomial model and those calculated using the empirical data suggests that the variability of a rate can be reliably estimated from a theoretical model.

The difference between the constant buffer distances used in the traditional locally-weighted-average method and those determined by population size deserves further discussion. Population usually has an uneven distribution over space, and thus applying a constant distance to the entire area is likely to result in different population bases for rates at different locations, which makes the rates less comparable. Furthermore, under a constant distance, a unit with a small population but large area may be under-smoothed; and a unit with a large population but small area may be over-smoothed. The reason is that the constant-distance method does not directly address population size, which is the underlying cause of instability.

We favor the population-determined distance method because it directly adjusts the population base for the unstable rate. In addition, this method allows the user to assess the stability of a rate, because the population base for every rate is known. This method is not likely to exhibit regional under-smoothing (like that with the constant-distance method). Nevertheless, since its buffer distance will be large in an area with low population density, it can exhibit over-smoothing in rural areas, and may conceal meaningful local geographical patterns in those areas. One way to overcome this potential difficulty is to perform smoothing using a series of population thresholds. This would allow the user to check rates at different locations using maps at different smoothing levels, which would illuminate the tradeoff between the presence of geographical patterns and the stability of rates in areas with small populations.

The subjectivity in determining the neighborhood for the unit under smoothing has been criticized as a major drawback of locally-weighted averaging methods. In this paper we propose a process to reduce the arbitrariness in the decision-making by testing a series of population sizes and identifying the threshold beyond which the variances in the rates become stable. This process is time consuming because it requires running the smoothing operations many times, but it is an option for the researcher who seeks a relatively objective way for selecting the minimum population size for smoothing.

In this study, we used a simplistic way to calculate the population within the neighborhood of the town under smoothing, based on the assumption that the population is evenly distributed across a town. This assumption does not reflect reality. When the population is unevenly distributed across a town, the weight of that town can be either over- or underestimated. However, the bias from our method should be smaller than when assuming that the entire population is concentrated into a single point. The bias caused by the “all or none” strategy of the point assumption must be on one of the 2 extremes, and the bias from the even-distribution assumption will be between the 2 extremes. Further, detailed information about the population distribution within geographic areas can improve the accuracy of the result from a polygon-based method. This would have no effect on the results from a point-based method.

To more accurately characterize the spatial distribution of the population within a geographic area, some researchers have used ancillary information, such as satellite imagery, topographic maps, and road network data, to redistribute the population within the unit.\(^{15,16}\) A polygon-based method is inherently compatible with these new technologies, and can be easily adapted to incorporate such data if they are available.

In conclusion, our findings suggest that a polygon-based method and the associated software tool are suitable for researchers and public health workers who have some knowledge of GIS, and who want to map smoothed disease rate for exploratory or surveillance purposes. In this paper, we used town-level lung cancer incidence data; the method is in principle applicable to any geographic unit (eg, zip code areas, census units, school districts, etc.) and any type of disease.

**ACKNOWLEDGMENTS**

*We thank Francis Magilligan, Margaret Karagas, and Linda Titus-Ernstoff, who kindly proofread the manuscript and provided highly valuable suggestions. We also appreciate the insightful comments from 2 anonymous reviewers. New Hampshire Department of Health and Human Services and the New Hampshire State Cancer Registry provided the data for this research.*

**REFERENCES**


© 2007 Lippincott Williams & Wilkins

To access the full article, please visit the Lippincott Williams & Wilkins website.
Estimating the Effects of Time-Varying Treatments
Incidence of Fractures Among Postmenopausal Japanese Women

Shiro Tanaka,* Yutaka Matsuyama,* Masataka Shiraki,† and Yasuo Ohashi*

Background: In the absence of unmeasured confounding, standard methods for estimating the effects of time-varying treatments on an event are biased when a time-dependent risk factor for the event also predicts subsequent treatments and when past treatment history predicts subsequent risk factor levels. In contrast, structural models provide unbiased estimates of effects when unmeasured confounding is absent.

Methods: We describe a multiplicative structural mean model and use it to estimate the effects of time-varying osteoporosis treatments on incidence of fractures among 1328 postmenopausal women over 40 years of age in a hospital-based cohort study in Japan. The parameters of the structural mean model are estimated by g-estimation.

Results: The number of vertebral fractures and bone mineral density levels predicted the selection of subsequent treatments and were affected by the previous treatments. Incidence rate ratios of bisphosphonates, active vitamin D3, and conjugated estrogen compared with no treatments on incidence of fractures among 1328 postmenopausal women over 40 years of age in a hospital-based cohort study in Japan. The parameters of the structural mean model are estimated by g-estimation.

Conclusions: Our analysis using the structural mean model showed that bisphosphonates, active vitamin D3, and conjugated estrogen all had preventive effects on the incidence of fractures by appropriate adjustments for time-dependent confounders. The results from standard Poisson-GEE analysis were the opposite of these results and of evidence from randomized trials. We consider our methods useful to estimate time-varying treatments within observational data.

(Epidemiology 2007:18: 529–536)

Adjustments for confounding such as stratification or regression analysis are commonly used to obtain an unbiased estimate of treatment effects in observational studies.1 However, if patients receive repeated treatments in the study and the main objective is to estimate the effects of time-varying treatments, these standard analyses may provide biased estimates under 2 conditions even in the absence of unmeasured confounders: (1) there exists a time-dependent risk factor that also predicts subsequent treatments, and (2) past treatment history predicts subsequent risk factor levels.2–6 These 2 conditions apply whenever there are time-dependent risk factors that are simultaneously confounders and intermediate variables. Factors that meet these conditions are known as time-dependent confounders.3

In clinical practice, outcomes of osteoporosis, such as incidence of fractures, back pain, bone mineral density, and other biomarkers are repeatedly measured, and osteoporosis is diagnosed based on a bone density of 2.5 SD (standard deviation) below the young adult mean.7–9 In this situation, time-dependent confounders arise when estimating the treatment effects on incidence of osteoporotic fractures based on observational data from daily clinical practice. The first condition (stated above) is met by a low bone density level because this level is a risk factor for fractures and is used to decide whether to initiate drug therapy. The second condition is met because prior treatments can change the bone density. Therefore, including bone mineral density as a time-dependent confounder in standard models such as generalized estimating equations (GEE)10 does not appropriately adjust confounding.3,5

In contrast to standard adjustment methods, structural models can be used to estimate treatment effects even if these 2 conditions hold.5,6,11 For a count outcome variable such as number of fractures, a multiplicative structural mean model5,11 can be used to estimate the effects of dynamic treatment plans in which treatment at a given time is decided partly on the patient’s evolving risk factor history.

In this article, we describe a multiplicative structural mean model to analyze daily clinical practice data involving time-dependent confounders. We applied the model to a hospital-based cohort study of 1328 postmenopausal women in Japan, evaluating the time-varying treatment effects on incidence of fractures of 3 categories of drugs used to treat osteoporosis (bisphosphonates, active vitamin D3, and conjugated estrogen).
METHODS

Subjects and Measurements

The Nagano cohort study is an observational study that has followed osteoporosis patients prospectively since December 1992 at one medical institute in Nagano, Japan. The protocol was reviewed by the ethical committee at the Research Institute and Practice for Involutional Diseases. Over 3500 Japanese women were enrolled from December 1992 through 2005 and we obtained informed consent from each subject. The current analysis is restricted to 1853 primary osteoporosis patients, that is, patients with secondary osteoporosis such as osteoporosis with hyperparathyroidism, hyperthyroidism, chronic renal failure or osteomalacia were excluded from our analyses. From these, we selected 1328 ambulatory postmenopausal women over 40-year-old whose osteoporotic fractures were assessed more than twice and for whom other covariates were not missing.

Women given drug treatment were on one of 3 regimens: (1) bisphosphonates (alendronate at 5 mg/d, risedronate at 2.5 mg/d, or etidronate at 200 or 400 mg/d for 14 days followed by 12 weeks’ cessation), (2) estrogen (conjugated estrogen at 0.3125 or 0.625 mg/d with or without progestin at 2.5 mg/d), or (3) alfacalcidol (active vitamin D₃ at 1 µg/d). Those regimens were adapted to the approved dosages of the drugs by the Japanese Minister of Health and Labor. Treatments may have been stopped or changed depending on change in bone density, occurrence of fractures, or other health conditions (including adverse events). The remaining subjects were followed without treatment for osteoporosis.

Outcome was assessed by the annual incidences of clinical fractures (hip fracture, surgical neck fracture of humerus, Colles’ fracture at the distal end of the radius, and the other sites of long-bone fractures) and of morphometrical vertebral fractures evaluated by radiographs of the thoracic and lumbar vertebral bodies (T4 to L4). The prevalent vertebral fractures were diagnosed in accordance with the Japanese criteria for vertebral fractures as reported elsewhere.9,12 Prevalent vertebral fracture was defined as having a ratio of bone mineral density (g/cm²) of the lumbar spine (L2–L4) was measured at baseline and then at 1- to 2-year intervals by dual-energy x-ray absorptiometry with the use of Lunar DPX-L or DPX-IQ (Lunar Corporation, Madison, WI). The young adult mean ± SD used for defining osteoporosis were 1.198 g/cm² ± 0.146 g/cm². The following values were also measured at baseline: other characteristics of study subjects (age, height, weight, and years after menopause); serum levels of calcium (Ca), phosphorus (P), and alkaline-phosphatase (AL-P); urine levels of Ca and creatinine (Ca/Cr ratio); and bone turnover markers (urinary deoxypyridinoline and N-telopeptides of type I collagen). The bone turnover markers deoxypyridinoline and N-telopeptides of type I collagen were standardized as T-scores that were calculated by the means and SDs of these parameters obtained from the young premenopausal Japanese women (n = 135). Appendix A1 shows a simple directed acyclic graph (DAG) representing our study and a simple numerical example illustrating a bias in standard analysis.

A Multiplicative Structural Mean Model

In this study, we followed N patients and each patient had measurement at discrete Tᵢ visits by year (t = 1, . . . , N) after registration. We diagnosed morphometrical vertebral fracture and bone density at each visit, and incidence of clinical fractures during the past one year was also reported. Let Yᵢ,t+1 be the count of incident clinical and morphometrical vertebral fractures observed during a 1-year interval (t, t + 1) (t = 0, 1, . . . , Tᵢ−1). Let Xᵢ,t = (Xᵢ,t,1, Xᵢ,t,2, Xᵢ,t,3)³ be the indicator variables for treatment during intervals [t, t + 1] where no drug therapy is the reference category (Xᵢ,t,3 = 1 for bisphosphonates, Xᵢ,t,2 = 1 for active vitamin D₃ and Xᵢ,t,1 = 1 for conjugated estrogen). Zᵢ,t is a vector of time-dependent covariates recording a bone density level, age, years after menopause, and visit time. Zᵢ,0 is time-independent characteristics at baseline, including the number of prevalent vertebral fractures. The treatment is selected at each visit, based on the patients past history involving the measurement at this visit. For example, at baseline, t = 0, Zᵢ,0 is measured and Xᵢ,0 is determined. Then, at t = 1, Yᵢ,1 and Zᵢ,1 are obtained and Xᵢ,1 is determined. For notational simplicity, we denote Hᵢ,t to be the past history used to select treatment at t, ie, Hᵢ,t = (Zᵢ,0, . . . , Zᵢ,t−1, Xᵢ,t−1, Yᵢ,t−1, . . . , Yᵢ,0) y for t = 1, . . . , Tᵢ−1 and Hᵢ,0 = (Zᵢ,0) y.

For a given patient, let Yᵢ,t+1,g(s) denote the potential outcomes at visit t + 1 in response to the hypothetical treatment g(s) = (Xᵢ,t+1,g(s), . . . , Xᵢ,t+1,0, 0, . . . , 0) s for s, t = 0, . . . , Tᵢ−1. That is, Yᵢ,t+1,g(1) represents the outcome we would have observed during (t, t + 1) if, possibly contrary to fact, the patient had her actual treatment history through visit s but was then switched to no drug therapy at visit s + 1 and did not receive any further drug therapy through the end of the study. Note that, at best, we observe Yᵢ,t+1,g(0) only for patients for whom the latter treatment sequence is what actually occurred, ie, for whom Xᵢ,s = (0, 0, 0), k = s + 1, . . . , Tᵢ−1. Our notation for the potential outcomes implicitly assumes Rubin’s stable unit treatment value assumption, which implies that potential outcomes of patient i do not depend on the treatment received by any other patient.14 We will also assume that the potential outcomes satisfy the consistency assumption5 that serves to link the potential outcomes with the observed outcomes. This assumption states that Yᵢ,t+1 = Yᵢ,t+1,g(0) for all t when Xᵢ,t+1 = . . . = Xᵢ,Tᵢ−1 = (0, . . . , 0) y. We assume a simple multiplicative structural mean model5,11 for the potential outcomes:

\[ \log \mathbb{E}[Y_{i,t+1,g(0)} | H_{i,t}, X_{i,t}] - \log \mathbb{E}[Y_{i,t+1,g(1)} | H_{i,t}, X_{i,t}] = \delta_0' X_{i,t} \]  

where \( \delta_0 = (\delta_{0,1}, \delta_{0,2}, \delta_{0,3})' \), t = 0, 1, . . . , Tᵢ−1 and g(−1) = (0, 0, . . . , 0). Under model, (1) \( \exp(\delta_0) \) can be
interpreted as the constant across $t$ and across covariates relative effect of a treatment $X_{i,t,k}$ at visit $t$ on the potential outcome $Y_{i,t+1,g(t-1)}$, following a patient’s actual treatment through visits $0, 1, \ldots, t-1$ and no treatment after $t$.

**G-estimation**

To estimate the treatment effects $\delta_0$ in model (1), we will assume sequential conditional independence$^5$ for any $t$ and $s$ with $s \leq t-1$:

$$Y_{i,t+1,g(t)}|X_{i,t}, H_{i,s} = Y_{i,t+1}(\delta) = Y_{i,t+1} \exp \left(-\sum_{k=1}^{t} \delta X_{i,t,k} \right),$$

which states that, when $s \leq t-1$, treatment $X_{i,t}$ is independent of the potential outcomes $Y_{i,t+1,g(t)}$ given the observed history up to visit $t$, $H_{i,s}$. It ensures that information in $H_{i,s}$ is sufficient to adjust for confounding to estimate the effect of $X_{i,t}$ on $Y_{i,t+1}$. Robins$^4$–$^6$ has referred to equation (2) as the assumption of no unmeasured confounders.

To estimate $\delta_0$, we define the “estimated” potential outcome at visit $t + 1$, $U_{i,t+1}(\delta) = Y_{i,t+1} \exp \left(-\sum_{k=1}^{t} \delta X_{i,t,k} \right)$, $U_{i,t+1}(\delta)$ is an estimate of the number of fractures during intervals $(t, t + 1]$ when the patient had her actual treatment history through visit $t-1$ but was then switched to no drug therapy at visit $t$ and did not receive any further drug therapy through the end of the study and $\text{E}[U_{i,t+1}(\delta_0)] = \text{E}[Y_{i,t+1,g(t-1)}]$ under model (1) and the consistency assumption. Therefore, under assumption (2), when $\delta = \delta_0$, $Y_{i,t+1,|X_{i,t}, H_{i,s}} = Y_{i,t+1}(\delta_0) = Y_{i,t+1} \exp \left(-\sum_{k=1}^{t} \delta X_{i,t,k} \right)$, $E[U_{i,t+1}(\delta_0) | X_{i,t}, H_{i,s}] = E[U_{i,t+1}(\delta_0) | H_{i,s}]$, where the estimated potential outcomes are mean independent of future treatments given the history. Robins$^1$ showed that equation (3) implies that the true values of $\theta = (\theta_1, \theta_2, \theta_3)$ in the following polytomous logistic model (4) are equal to 0:

$$\text{Pr}[X_{i,t,k} = 1 | H_{i,s}, U_{i,t+1}(\delta_0)] = \frac{\exp[\alpha_i + \theta_1 U_{i,t+1}(\delta_0)]}{1 + \sum_{i=0}^{n} \exp[\alpha_i + \theta_1 U_{i,t+1}(\delta_0)]}. \hspace{1cm} (4)$$

For any hypothesized vector values $\delta$, we can calculate the estimated potential outcome $U_{i,t+1}(\delta)$ from the data. The point estimate called g-estimate is the $\delta$ solving $S(\alpha, \theta = 0, \delta) = 0$, where $S(\alpha, \theta = 0, \delta)$ is a vector of score statistic for (4) with respect to $\theta$. Under the assumed multiplicative model (1), $\exp(\delta)$ is an estimate of IRR for treatment $k$ relative to no drug therapy. To construct the Wald type confidence intervals (CIs) of the IRR estimates, we estimated standard error of $\delta$ by numerical derivatives and used the multivariate delta method (see appendix A2).$^{15,16}$

**Data Analysis**

We analyzed Nagano cohort data using a multiplicative structural mean model by a program written in matrix language SAS/IML. The candidates of time-dependent covariates in model (4) were the following: drugs given in the previous year, number of vertebral fractures and bone density measured at each visit, age, years after menopause and visit time. The candidates of time-independent covariates in model (4) were drugs given at baseline, the number of prevalent fractures at baseline, bone density measured at baseline, height, weight, serum levels of Ca, P and AL-P, Ca/Cr ratio, and T-score converted bone turnover markers. We also used quadratic terms of the number of vertebral fractures and bone density. To reduce the number of covariates, we used the stepwise variable selection method with the inclusion and exclusion criterion being a $P$ value less than 0.1. The drug treatment variables, linear terms of the number of prevalent fractures, and bone density levels were not entered into the variable selection procedures. Consequently, the linear terms of age and height, and the quadratic terms of bone density and visit time remained, and we used these variables as covariates in model (4). We tried other variable selection procedures such as backward deletion or forward selection strategies, and confirmed that the final analysis results were not sensitive to the chosen variable selection procedures.

**RESULTS**

**Descriptive Characteristics**

Table 1 shows descriptive characteristics of 1328 patients according to their initial drug treatment status. Initial osteoporosis treatments were classified into 4 groups: bisphosphonates ($n = 138$; alendronate [$n = 60$], risedronate [$n = 6$], and etidronate [$n = 72$]), active vitamin D$_3$ (alfacalcidol; $n = 201$), conjugated estrogen (estrone, estriol, or estradiol; $n = 127$), and no drug therapy (observational; $n = 862$). Prevalent vertebral fractures at baseline were found in 22% of the women. Women who received bisphosphonates and active vitamin D$_3$ as the initial drug treatment had more prevalent vertebral fractures (44%) than women in the other 2 groups. The overall mean ± SD of bone density levels at baseline was 0.894 ± 0.193 which was lower than the young adult mean minus 2SD. Women who received no drug treatment at baseline had higher bone density levels (mean ± SD = 0.957 ± 0.187) than those of the other 3 treatment groups. The overall mean ± SD of age at baseline was 65.2 ± 9.4, and women who received conjugated estrogen at baseline were younger (mean ± SD = 58.3 ± 8.5) than women in the other 3 groups.

**Assessment of Time-Dependent Confounding**

We first assessed whether the bone density levels were time-dependent confounders or not. Three conditions are necessary for a variable to be both a confounder and an intermediate variable: first, the bone density levels are risk factors for incidence of future fractures, second, the 3 drug treatments (bisphosphonates, active vitamin D$_3$, and conjugated estrogen) affect bone density levels compared with no drug therapy, at least during the first year, and third, the selection of treatment depends on history, including bone density levels and previous treatment.

Table 2 shows the effects of the number of vertebral fractures and bone density on fracture incidence in the first year. Existing vertebral fractures, low lumbar bone density, and higher age are well-established risk factors for osteoporotic fractures.$^7$ As expected, the incidence of fractures in the
first year was increased with lower bone density levels (IRR of one SD increase was 0.75; 95% CI = 0.58–0.97). IRR of one SD increase of prevalent vertebral fractures was 1.19 (1.06–1.33). Furthermore, all 3 initial drug treatments (bisphosphonates, active vitamin D3, and conjugated estrogen) decreased the incidence of fractures in the first year.

Table 3 shows the results of the effects of initial treatment on subsequent bone density levels in the first year using linear regression analysis. The bone density levels of women who were not receiving drug therapy at baseline were decreased (intercept = −0.008), whereas those of women taking bisphosphonates, active vitamin D3, and conjugated estrogen were increased (0.039, 0.009, 0.034 g/cm², respectively) after adjustment of other covariates.

Table 4 shows the estimates of the odds ratio (OR) for receiving the 3 drug treatments compared with no drug therapy. In each treatment, the number of vertebral fractures and bone density levels just before the visit predicted the selection of future treatments. Furthermore, lower bone density levels just before the visit were stronger predictors of treatment with bisphosphonates than of treatment with active vitamin D3 and conjugated estrogen, which tended to be selected for younger patients.

Adjustment of Time-Dependent Confounding

Table 5 shows the IRRs for time-dependent drug treatments on the incidence of clinical and vertebral fractures estimated by the structural mean model. The IRRs for initial drug treatment are also shown. The IRRs for bisphosphonates, active vitamin D3, and conjugated estrogen estimated by the structural mean model were 0.58 (95% CI = 0.44–0.77), 0.82 (0.48–1.38), and 0.60 (0.47–0.76), respectively.
To compare the multiplicative structural mean model with standard analysis, we also show the results of the Poisson-GEE regression analysis. To estimate the effects of initial drug treatments, we included treatments at baseline as drug treatment variables, only the baseline confounders were adjusted, and intradividual correlation is accounted for by GEE methods using compound symmetry correlation structure. The IRRs for initial drug treatment obtained by the Poisson-GEE analysis were 1.61 (1.23–2.10), 1.16 (0.96–1.40), and 0.73 (0.52–1.02), respectively. Clearly, the findings from the standard GEE methods were in conflict with the results from the structural mean model.

**DISCUSSION**

We evaluated the effectiveness of 3 categories of the drugs for osteoporosis (bisphosphonates, active vitamin D3, and conjugated estrogen) on the incidence of fractures among 1328 postmenopausal women of a hospital-based cohort study in Japan. Using the structural mean model, we found the risk of clinical and vertebral fractures among women treated with these drugs was lower than that of women who were not treated. The bisphosphonates and conjugated estrogen were shown to be particularly effective.

An important limitation associated with this study is the potential for selection bias: about 28% (1 − [1328/1853] = 0.28) of primary osteoporosis patients were not included in the analyses, mainly due to inadequate measurements of osteoporosis or missing covariates during follow-up. Thus, it is possible that our analysis group may not be representative of the entire primary osteoporosis patients. However, various baseline variables were quite comparable between the analysis sample and the entire population; therefore, we believe this bias is minimal. Other limitations are the misspecification of the structural model and residual confounding, ie, violation of Equations (1) and (2) are also potential source of bias. For simplicity, we chose a multiplicative structural model (1), but the model can be extended to the model including interaction terms or some functions of past treatment histories. If we have sound background knowledge on treatment effects, such complicated models can be applied in the future. For residual confounding, many clinically important prognostic factors were measured and all of them were used as covariates; therefore, any departure from assumption2 will be small in our analysis.

Many randomized clinical trials of osteoporosis treatments have been reported.17–30 The efficacy of bisphosphonates has been studied intensively in large clinical trials with vertebral fractures as the primary end point,18,19 and the treatment effect of bisphosphonates estimated in our study was consistent with results of these trials. Very few reports on the effects of hormone-replacement treatment (HRT) for the prevention of bone fractures in osteoporosis (treatment study) were available.20–24 For instance, the Women’s Health Initiative study, the largest study to date, was performed in healthy postmenopausal women, and it showed the effectiveness of HRT on the prevention of hip fracture with some adverse events.20

Our study supports the efficacy of HRT to prevent fractures in osteoporosis. We used a low dosage of conjugated estrogen and we found no adverse events (eg, cardiovascular or cerebrovascular events or development of breast cancer). The most frequent reason for discontinuing HRT was vaginal bleeding.

In contrast to standard practice in Western countries, the active vitamin D3 analog, alfacalcidol, is the most commonly used drug to treat osteoporosis in Japan. Although there are several reports of the use of active vitamin D3 to prevent osteoporotic fractures, the efficacy of alfacalcidol in preventing fractures is controversial. A recent study reported by Ishida et al24 did not show a substantial preventive effect of active vitamin D3 on osteoporotic fractures. Kushida et al25 reported that alendronate was more effective in preventing vertebral fractures than alfacalcidol. Our current study agrees

### TABLE 4. Polytomous Logistic Regression Result for Selecting Treatments

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Bisphosphonates OR (95% CI)</th>
<th>Active Vitamin D₃ OR (95% CI)</th>
<th>Conjugated Estrogen OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of vertebral fractures at baseline</td>
<td>0.59 (0.45 to 0.72)</td>
<td>0.80 (0.63 to 1.01)</td>
<td>0.65 (0.48 to 0.90)</td>
</tr>
<tr>
<td>No. of vertebral fractures just before visit</td>
<td>1.61 (1.33 to 1.94)</td>
<td>1.28 (1.05 to 1.57)</td>
<td>1.88 (1.44 to 2.44)</td>
</tr>
<tr>
<td>BMD at baseline (±1 SD)</td>
<td>1.30 (0.92 to 1.86)</td>
<td>1.00 (0.69 to 1.49)</td>
<td>1.54 (0.99 to 2.38)</td>
</tr>
<tr>
<td>BMD just before visit (±1 SD)</td>
<td>0.27 (0.19 to 0.38)</td>
<td>0.50 (0.34 to 0.74)</td>
<td>0.43 (0.28 to 0.67)</td>
</tr>
<tr>
<td>Age just before visit (±10-yrs-old)</td>
<td>0.81 (0.68 to 0.97)</td>
<td>1.03 (0.87 to 1.21)</td>
<td>0.48 (0.38 to 0.61)</td>
</tr>
<tr>
<td>Height (±10 cm)</td>
<td>1.36 (1.07 to 1.73)</td>
<td>1.15 (0.92 to 1.43)</td>
<td>1.54 (1.15 to 2.06)</td>
</tr>
</tbody>
</table>

### TABLE 5. Estimates of Treatment Effects by the Structural Mean Model and the Poisson-GEE Analysis

<table>
<thead>
<tr>
<th>Treatment Variables</th>
<th>IRR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural mean model</td>
<td></td>
</tr>
<tr>
<td>Bisphosphonates</td>
<td>0.58 (0.44 to 0.77)</td>
</tr>
<tr>
<td>Active vitamin D₃</td>
<td>0.82 (0.48 to 1.38)</td>
</tr>
<tr>
<td>Conjugated estrogen</td>
<td>0.60 (0.47 to 0.76)</td>
</tr>
<tr>
<td>Poisson-GEE (initial treatments)</td>
<td></td>
</tr>
<tr>
<td>Bisphosphonates</td>
<td>1.61 (1.23 to 2.10)</td>
</tr>
<tr>
<td>Active vitamin D₃</td>
<td>1.16 (0.96 to 1.40)</td>
</tr>
<tr>
<td>Conjugated estrogen</td>
<td>0.73 (0.52 to 1.02)</td>
</tr>
</tbody>
</table>
with these previous reports; thus, it may be concluded that the effect of active vitamin D3 treatment on fracture prevention is weaker than that of bisphosphonates or HRT. In summary, it is meaningful that our study confirmed the efficacy of these 3 treatments compared with no drug therapy on fracture incidence, because many clinical trials evaluating drugs for osteoporosis treatment in Japan did not have a placebo control group, and changes in bone density levels were used as a primary end point.

Our study also suggests the usefulness of analysis by structural mean models that allow estimation of time-varying treatment effects even in the presence of time-dependent confounders. The estimates obtained from Poisson-GEE analysis were different from those of the Poisson regression using data in the first year (Table 2). Further, the effect of bisphosphonates was opposite to the results from previously published trials and common understanding that bisphosphonates are accepted to prevent osteoporosis-related fractures. Routine use of Poisson-GEE analysis should be avoided when time-dependent confounders exist. Our data suggested that well-known prognostic factors in osteoporosis, such as bone density level and the number of fractures, met the first 2 conditions stipulated above (bone mineral density was a risk factor for fractures, and the drug treatments affect bone mineral density) and were time-dependent confounders (Tables 2, 3, and 4).

In conclusion, our analysis using the structural mean model showed that the bisphosphonates, active vitamin D3, and conjugated estrogen all had a preventive effect on the incidence of fractures by appropriate adjustments for time-dependent confounders. The results from standard Poisson-GEE analysis were the opposite of these results and of evidence from randomized trials. We consider our methods useful to estimate time-varying treatments within observational data.

REFERENCES

2. Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. J R Statist Soc A. 1984;147:656–666.
APPENDIX

A.1. Numerical Example Illustrating a Bias in Standard Analysis

Figure 1 shows a simple DAG, which represents our study with $T = 2$. In Figure 1, for simplicity, $Z_t$ ($t = 0, 1, 2$) represents the value of the bone density level at visit $t$ ($Z_t = 0$ if normal bone density and $Z_t = 1$ if low level), $X_t$ represents the bisphosphonate treatment at visit $t$ ($X_t = 0$ if no drug, $X_t = 1$ if bisphosphonate), which is determined by the measured previous history, and $Y_t$ ($t = 1, 2$) represents the numbers of fracture at visit $t$, which are observed before the selection of treatment at visit $t$. As the DAG in Figure 1 shows, a treatment received at a particular visit affects the treatment, fracture, bone density at subsequent visit, and is affected by the previous treatment, bone density histories, and numbers of fracture. The assumption of no unmeasured confounders means that the selection of treatment at each visit can be totally explained by measured variables and there are no arrows from the unmeasured covariates into treatment variables like in Figure 1. The DAG in Figure 1 also shows that bone density is a time-dependent confounder. For example, $Z_1$ affects both treatment ($X_1$) and fracture ($Y_1$), but is an intermediate variable in the pathway from treatment ($X_0$) to fracture ($Y_1$).

Figure 2 shows a hypothetical cohort of 10,000 osteoporosis patients with $T = 2$. In this cohort, suppose that bone density levels at visit 0 is $Z_0 = 0$ for all patients and patients are randomized to $X_0 = 0$ or 1 with probability of 0.5 (this means that there is no arrow from $Z_0$ to $X_0$ in Fig. 1). We assume that the selection of treatment at visit 1 can be totally explained by previous treatment and bone density at that visit, through a logistic regression model, $\logit[Pr(X_1 = 1)] = 1.73*Z_t + 0.463*X_0$. This assumption means that 2 arrows from $Z_0$ and $Y_1$ into $X_1$ are removed in Figure 1. At visit 1, it is determined whether a bone density level of a patient is $Z_1 = 0$ or 1, in which 20% of patients with $X_0 = 0$ are remained $Z_1 = 0$ and 70% of patients with $X_0 = 1$ are remained $Z_1 = 0$. In Figure 2, the numbers in parenthesis at each visit are expected frequencies of patients classified by each branch of a bone density ($Z$) level. Selected treatment at visit 1 ($X_1$) is determined by the above logistic regression model, where the selection probabilities in each branch are shown in Figure 2. At visit 2, it is also determined whether a bone density level of a patient is $Z_2 = 0$ or 1, in which 20% of patients with $X_1 = 0$ are remained $Z_2 = 0$ and 70% of patients with $X_1 = 1$ were remained $Z_2 = 0$. The expected numbers of fracture observed at each visit are shown in bracket of the right side in Figure 2, which are determined by the model, $\log(\text{incidence rate of } Y_t) = -3.51 + 1.61*Z_t + 1.61*Z_{t-1} - 0.047*X_{t-1} (t = 1, 2)$.

Based on this hypothetical data, we show that the GEE analysis including the time-dependent confounder, bone density level, may falsely suggest that bisphosphonate treatment is harmful (we ignore the statistical error in the following analyses).

In Figure 2, the IRR of the initial treatment $X_0$ on $Y_1$ can be unbiasedly estimated by comparing the randomized group,

![FIGURE 1. A simple DAG of a hypothetical osteoporosis cohort study under the assumption of no unmeasured confounders.](image)

![FIGURE 2. Hypothetical data of osteoporosis cohort study with $T = 2$. Numbers in parenthesis at each visit are expected frequencies of patients classified by each branch of bone density ($Z$) level. Numbers on arrows from $Z$ to $X$ are selection probabilities of treatment in each branch.](image)
For this hypothetical data, we fitted the Poisson-GEE models with unstructured correlation structure, with and without adjustment of a time-dependent confounder $Z$. First, we used the following model for the unadjusted estimate: $\log E(Y_t | X_{t-1}) = \beta_0 + \beta_1 X_{t-1}$, where $t = 1, 2$. The resulting IRR estimate was exp($\beta_1$) = 0.92, which would be biased due to the confounding effect of $Z$. Second, we used the following model for the adjusted estimate: $\log E(Y_t | X_{t-1}) = \beta_0 + \beta_1 X_{t-1} + \beta_2 Z_t$ where $t = 1, 2$. The resulting IRR estimate was exp($\beta_1$) = 1.70, which would be biased due to the adjustment of intermediate variable $Z$. Thus, both results with and without adjustment of a time-dependent confounder were different from the unbiased result (IRR = 0.50), and the adjusted GEE analysis reversed even the direction of treatment effect.

### A.2. Computational Details of G-Test and Confidence Intervals

For any hypothesized vector values $\delta$, we can calculate the estimated potential outcome $U_{i,r+1}(\delta)$ from the data. Thus, not knowing the true values $\delta_0$, we can test the hypothesis that particular values of $\delta$ are equal to $\delta_0$ by the score test of the hypothesis that the true values of $\theta_k$ are 0 in model (4). This score test is called the g-test, and the test statistics can be written as follows:

$$Z(\delta) = S(\hat{\alpha}_k, \theta = 0, \delta)\hat{\Omega}(\delta)^{-1} S(\hat{\alpha}_k, \theta = 0, \delta),$$

where $\hat{\alpha}_k$ is the restricted maximum likelihood estimates of $\alpha$ when $\theta_k$’s are set to 0 in model (4) and $S(\hat{\alpha}_k, \theta = 0, \delta)$ is a vector of score statistic and $\hat{\Omega}(\delta)$ is its covariance matrix. Under the hypothesis $\delta = \delta_0$, $S(\hat{\alpha}_k, \theta = 0, \delta_0)$ has asymptotic normal distribution $N(0, \hat{\Omega}(\delta_0))$.

To construct the Wald type CIs of the IRR estimates, we estimated standard error of $\delta$ by numerical derivatives and used the multivariate delta method. Because we can calculate $S(\hat{\alpha}_k, \theta = 0, \delta)$ at any values of $\delta$, derivative of $\delta$ as a function of $S(\hat{\alpha}_k, \theta = 0, \delta)$ can be estimated by the slope of fitted line to the graph of $S(\hat{\alpha}_k, \theta = 0, \delta)$ versus $\delta$. Thus, let $D$ be a matrix of $dS/d\delta$ estimated numerically, and we get the following estimate of covariance matrix $\hat{V}$ of $\delta$ using the delta method, $\hat{V} = D\hat{\Omega}D$. Using this estimate of standard error, we constructed the approximate 95% CIs of logarithm of IRR by the Wald method.
Estimating the Longitudinal Prevalence of Diarrhea and Other Episodic Diseases

Continuous Versus Intermittent Surveillance

Wolf-Peter Schmidt,* Stephen P. Luby,† Bernd Genser,‡ Mauricio L. Barreto,‡ and Thomas Clasen*

Background: Longitudinal prevalence (ie, the proportion of time with the disease) is used to describe morbidity from diarrhea and other epidemic conditions. The aim of this analysis was to compare estimates of longitudinal prevalence based on intermittent sampling at regular intervals with 24- or 48-hour recall, with estimates based on continuous surveillance.

Methods: Based on 2 real datasets from Brazil and Guatemala, we developed a simulated dataset representing the diarrhea morbidity of 10,000 individuals followed over 365 days.

Results: Both the model and the real datasets showed that the standard deviation of the longitudinal prevalence increases with decreasing numbers of days sampled, so that a study sampling only a fraction of days would require a larger sample size. However, due to the correlation of diarrhea between consecutive days, sampling at 7- to 14-day intervals results in relatively small loss of precision and power compared with daily morbidity records, especially when the average diarrheal episode is long. A study based on morbidity data for every seventh day may require only a 5%-24% larger sample size than a study with daily records, depending on the average duration of episodes. Using a recall period of 48 hours instead of 24 hours increases power if the average episode is short.

Conclusions: The results question the necessity of continuous surveillance to estimate longitudinal prevalence. In addition to savings in cost and staff time, intermittent sampling of morbidity may improve validity by minimizing recall error and reducing the influence of surveillance on participants’ behavior.

(Epidemiology 2007;18: 537–543)
estimate of the longitudinal prevalence and the implications for the sample size.

**METHODS**

We addressed the study objectives by developing a simulated dataset based on the methods published by Morris and colleagues.\(^8\) The simulated dataset reflected the diarrhea occurrence in a population, with episodes being randomly distributed over time, but highly clustered in individuals, with the majority of individuals having a small number of episodes and few people experiencing many episodes. The model was parameterized based on 2 real datasets with daily morbidity data over the course of at least 1 year. The first dataset was based on 2 longitudinal cohort studies conducted in 1997–1999 and 2000–2002 in Salvador de Bahia in Brazil in children under the age of 5.\(^{14,18,19}\) Diarrhea was assessed by twice-weekly household visits, ie, recall periods of 3–4 days. We combined the 2 cohorts into one dataset, matched by calendar date to preserve the seasonal pattern of diarrhea occurrence. The dataset was reduced to 365 days of observation to enable comparability with the simulation. The study represented an open cohort. The number of days of observation varied among individuals. In the reduced dataset (n = 1839), the mean number of days under observation was 263 days (range, 2–365). The weighted mean number of episodes per individual in the reduced dataset (weighted by the number of days of observation) was 3.7. The mean episode duration was 2.5 days, whereas the weighted mean longitudinal prevalence (weighted by the number of days of observation) was 2.9%.

The second dataset was from a randomized-controlled trial studying the effect of different household water treatment techniques on diarrhea occurrence in rural Guatemala.\(^20\) This study was a closed cohort (n = 2982) with equal follow-up time of just over 365 days, with disease being assessed in all age groups by weekly visits (ie, recall periods of around 7 days). We reduced the dataset to 365 days of observation, again ensuring that each day in the dataset represented one calendar day. All 5 study arms were combined. The mean number of episodes per individual in the reduced dataset was 1.7. The mean episode duration was 5.1 day, and the mean longitudinal prevalence was 2.5%. Thus, the 2 real datasets differed in the distribution of the number of episodes and the episode duration (presumably due to differences in age range, study setting or procedures) while being similar with regard to the mean longitudinal prevalence.

**Development of the Simulated Dataset**

We generated a dataset representing the daily diarrhea experience of 10,000 individuals over a period of 365 days. First, we created a variable that defined the number of diarrhea episodes per individual following a gamma distribution (a distribution suitable to represent highly skewed data such as diarrhea incidence).\(^8\) The parameters of the gamma distribution were chosen so that the distribution of the number of episodes was within the bounds of the 2 real datasets (Fig. 1A); the parameters chosen for this default model were \(\alpha = 0.48\) (shape parameter) and \(\beta = 6\) (stretch parameter). This resulted in a mean number of 2.9 episodes per individual, which was less than assumed by Morris et al\(^8\) based on data from Peru (mean number of episodes 9.0, \(\alpha = 1.5, \beta = 6\)).

In the next step we distributed the episodes of each individual randomly over the period of 365 days. Each episode was then randomly allocated an episode duration following an exponential distribution \(y = \exp(\alpha x)\) with \(\kappa = -0.32\), which resulted in a distribution of episode durations between the 2 real datasets (Fig. 1B), with a mean duration of 3.8 days and a mean longitudinal prevalence of 2.7%.

Different episodes were allowed to overlap. Thus, the average "observed" number of episodes in the simulated dataset was slightly lower than the number of episodes as given by the gamma distribution (2.6 vs. 2.9). Likewise, the average "observed" episode duration was slightly longer then determined by the negative exponential distribution (3.8 vs. 3.6).
Choice of Outcome Measures

Unlike the conventional definition of prevalence, the longitudinal prevalence is a continuous variable. An individual can experience a longitudinal prevalence between 0% and 100%. For diarrhea, the distribution of the individual values of the longitudinal prevalence is highly skewed, with most individuals experiencing a longitudinal prevalence of up to one percent (Fig. 2, highest bars in each panel), and few individuals having a high longitudinal prevalence. Thus, the mean longitudinal prevalence may at first sight not seem to be an appropriate outcome measure to describe the longitudinal prevalence on population level. However, the mean longitudinal prevalence in a population is equal to the average diarrhea point prevalence over the study period (ie, the mean of the daily point prevalences) if each point prevalence is based on the same number of individuals. The mean longitudinal prevalence is also equal to the proportion of diarrhea days among all days observed in the study population, provided that all individuals were followed for the same number of days. If the follow-up time varies among participants, then the weighted mean longitudinal prevalence (weighted by the number of days of observation) is equal to the mean point prevalence over time weighted by the number of individuals contributing to the daily point prevalence. This again is equal to the proportion of days with diarrhea in the population, which certainly is an outcome measure of interest.

To describe precision, we used the standard deviation (SD), which allows straightforward sample size calculations for studies using the longitudinal prevalence as an outcome measure. With decreasing days of observation per individual, the SD of individual longitudinal prevalence values is expected to rise, indicating the loss of precision with fewer days of observation per individual. The increase of the SD with fewer numbers of visits is proportional to the increase of the standard error of the mean longitudinal prevalence (SD/√n), since the overall number of individuals remains unchanged.

The loss of precision of the longitudinal prevalence estimate has implications for the required sample size of a study using the longitudinal prevalence as an outcome measure. We applied a standard formula for the comparison of 2 means \( n = (0.84 + 1.96)^2 \left( \frac{2 \times SD^2}{\text{mean}_1 - \text{mean}_2} \right) \), because most diarrhea studies compare 2 or more groups. For illustration, we assumed a 30% reduction of the mean longitudinal prevalence, 80% power and \( \alpha = 0.05 \).

Simulation Procedures

We simulated the assessment of the longitudinal prevalence by assuming increasing intervals between visits over 365 days. At each visit, we simulated a 24-hour and a 48-hour recall period, which have been suggested as the optimum periods to achieve a high level of recall.\(^{10-12}\) For simplicity, we assumed perfect recall.

To determine the robustness of our findings to changes in the model assumptions, we varied the parameters of the default simulation by assuming plausible high and low values for number of episodes and duration of illness as suggested by the 2 datasets from Brazil and Guatemala and by the parameterization used by Morris et al.\(^8\) We also explored the effect of seasonal variation of illness incidence by relocating one-third of the episodes at random from the second half of the simulated time period to the first half, resulting in a diarrhea incidence that was twice as high during the first half of the year as in the second half. Simulations and analyses were performed with STATA 9.0 (StataCorp, College Station, TX).

RESULTS

Figure 3 shows the association between number of days of observation and SD of the longitudinal prevalence for the simulation model. It demonstrates the loss of precision and power that occurs in the default model when decreasing numbers of days of observation are sampled. As the whole
sampling period covers 365 days, a 7-day surveillance interval amounts to 53 visits overall, a 14-day interval to 27 visits, and so on. The left y-axis reflects the mean longitudinal prevalence, which remains constant, and the SD of the mean longitudinal prevalence, which rises with longer surveillance intervals (and decreasing number of visits). As the SD enters into the sample size formula for the comparison of 2 means to the square, the effect of reducing the number of days sampled on the relative increase of the sample size is more pronounced (right y-axis). For example, a study recording daily disease prevalence over 365 days would require a sample size of 390 persons per arm (baseline) in the default model simulation. Recording only every 28th day (14 visits) increases the SD by the factor 1.3 (from 0.041 to 0.054) and the sample size by the factor 1.8 (from 390 to 716 participants per arm). In this model there seems to be a slight benefit of applying a 48-hour recall period instead of a 24-hour period (sample size 659 vs. 716 for 28-day interval).

In the dataset from Brazil (Fig. 4A.) the rise of the SD (factor 1.5 for 28-day interval) and the sample size (factor 2.3 for 28-day interval) with decreasing numbers of days sampled is steeper than in the simulated dataset, although the increase in the sample size when relying on a 7-day interval is still limited. Using a recall period of 48 hours instead of 24 hours substantially reduces the need to increase the sample size.

Figure 4B shows the data from Guatemala, where the average duration of illness was much longer. For a sampling interval of 28 days, the SD increases by the factor 1.2, and the required sample size by a factor of just 1.6 compared with all days sampled. There is little benefit of sampling at intervals shorter than 10 days. Using a 48-hour recall offers a slight advantage only for long sampling intervals.

We explored the apparent association between illness duration and the loss of precision with decreasing numbers of visits by varying the input parameters of the simulated dataset. Figure 5A confirms that fitting the illness duration in the model to the distribution observed in Brazil (approximately an exponential curve with $k = -0.6$ and mean episode duration 2.4 days) results in a steep rise in the sample size that is reduced by using a 48-hour recall. For a 28-day interval, the sample size goes up by a factor of 2.6 (compared with sampling all 365 days), which is similar to the Brazil dataset (factor of 2.4). Likewise, increasing the illness duration to what was observed in Guatemala ($k = -0.25$; mean episode duration = 4.6 days) leads to a sample size increase equal to the Guatemala dataset (factor 1.6 for a 28-day interval). In contrast, varying the incidence of diarrhea by fitting the parameter for the gamma distribution to the data from Brazil (approximately $\alpha = 0.56$, $\beta = 6$) and Guatemala ($\alpha = 0.3$, $\beta = 6$), while leaving the illness duration constant, does not have a strong effect on the proportional increase of the sample size (Fig. 5B). Although the overall sample sizes are highly dependent on the disease incidence, the relative increase in the required sample size for decreasing numbers of surveillance visits changes little, even if assuming a very high disease incidence based on the Ghana/Peru model ($\alpha = 1.5$, $\beta = 6$, mean longitudinal prevalence = 8.3%). The main findings for weekly and fortnightly visits with the key parameters for illness duration and incidence are summarized in Table 1.

Assuming a 2-fold higher incidence in the first half year to simulate seasonality while keeping the number of episodes unchanged resulted in a slightly steeper curve compared with the default simulation model (factor 2.0 for 28 days interval instead of 1.8), indicating that seasonal variation has only a limited impact on our findings.

Finally, we explored the implications for the precision and the sample size if the number of visits remains constant, while varying the length of the sampling intervals between visits. In other words, we simulated a situation where a fixed number of visits per household is spread over different overall durations of a study. Figure 6 shows that for 25 visits in the default model and the long episodes model ($k = -0.25$) with 24-hour recall, the power of a study can be maximized by applying at least a 10–14 day interval (equivalent to a study duration of 241–337 days). If the average duration of illness is smaller, as in the Brazil dataset ($k = -0.6$), the surveillance intervals can be shorter (around 7 days) with a slight advantage for 48-hour recall compared with 24-hour recall when using a 14-day interval.
days, upper line). This was broadly confirmed by applying this approach to the 2 real datasets, although the marked fluctuations in diarrhea occurrence over the year made it difficult to achieve comparable estimates for different study durations (not shown). The curves for a 48-hour recall leveled off in a very similar way (not shown). Again, varying the number of episodes (as in Fig. 5B) had little impact on the slope, showing that the preferred interval does not depend on the incidence (not shown).

**DISCUSSION**

Our results suggest, that in many situations, sampling every seventh day yields only slightly less precise estimates of the mean longitudinal prevalence of diarrhea in a population than collecting disease records for every single day. To achieve the same precision, a study based on morbidity data for every seventh day may require a 5%–24% larger sample size than a study with daily records, depending on the average duration of episodes (Table 1). In settings with short epi-

---

**FIGURE 4.** Mean longitudinal prevalence, standard deviation, and sample size for different sampling intervals over the course of 365 days for the (A) Brazil and (B) Guatemala datasets. Mean LP and SD for Brazil weighted by the number of days of observation; relative sample size (right y-axis) indicates increase compared with baseline sample size if all days are sampled (Brazil: 415; Guatemala: 833 per arm); sample size calculation for comparison of 2 mean LP estimates.

**FIGURE 5.** Effect of changing the simulated duration and the number of episodes on the sample size. Sample size per arm (y-axis) for the comparison of 2 mean longitudinal prevalence estimates; k is the parameter of the exponential distribution \( y = \exp(kx) \) representing episode durations; \( \alpha \) is the shape parameter of the gamma distribution representing the number of episodes in an individual (stretch parameter kept constant at \( \beta = 6 \)); numbers on right side of both graphs indicate sample size increase relative to baseline (all days sampled).
sodes, the increase in the sample size can be reduced (in our simulated example from 24% to 14% for 7-day intervals) by applying a 48-hour recall period.

Recording daily disease occurrence (eg, by frequent visits or by relying on a long recall period) is resource-intensive, could affect the reporting and risk behavior of the study population, and could produce imprecise estimates. Unless the study requires measuring the incidence of diarrhea or close surveillance for other reasons (such as prompt treatment or collection of stool samples), intermittent sampling with a 24- or 48-hour recall period could improve reporting and reduce expenses. Compared with daily records, the increase in required sample size with sampling intervals of up to 14 days appears moderate. However, interval and sample size also depend on the expected average duration of illness, which affects the degree to which consecutive days are correlated. Likewise, depending on the average episode length, investigators may opt for either a 24-hour recall period, which may be simpler and more precise, or a 48-hour recall period, which still seems to yield valid data10 –12 but with little advantage if the average illness duration is long.

Often the logistical constraints lie not so much in the overall duration of the study as in the total number of visits performed. For a fixed number of visits to each household, spacing disease recordings to at least 7–14 days maximizes efficiency. Investigators may thus choose to employ a small number of well-trained field workers for a longer time, rather than a large group of field workers for a short and intensive period. A longer duration of a study has the additional advantage that it better captures seasonal variations in disease occurrence.

Our findings confirm some aspects of previous work by Morris and colleagues9 who used real datasets to show that sampling morbidity every 5 days can reliably classify study participants into longitudinal prevalence quintiles. By using simulated datasets and comparing with real data, we have identified illness duration as the key parameter for estimating the longitudinal prevalence on population level. Morris and colleagues had suggested that prevalence of disease is the main determinant of the required number of visits to estimate the longitudinal prevalence on individual level without specifically considering illness duration.

![Figure 6](image)

**FIGURE 6.** Effect of increasing the sampling intervals for a fixed number of visits (n = 25). Sample size per arm (y-axis) for the comparison of 2 mean longitudinal prevalence estimates; upper line (△) indicates simulation with short episodes as observed in Brazil; middle line (x) shows default simulation model; lower line (+) shows simulation with long episodes similar to Guatemala; k is the parameter of the exponential distribution y = exp(kx) representing episode durations.

### TABLE 1. Examples of Sample Sizes for a Hypothetical Diarrhea Intervention Trial Based on the Simulated Datasets and Different Sampling Strategies

<table>
<thead>
<tr>
<th>Surveillance Scheme</th>
<th>Daily Records</th>
<th>Visit Every Seventh Day</th>
<th>Visit Every 14th Day</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>24-Hour Recall</td>
<td>48-Hour Recall</td>
</tr>
<tr>
<td>Default dataset (as in Fig. 3; mean episode length 3.8 d)</td>
<td>390</td>
<td>428</td>
<td>415</td>
</tr>
<tr>
<td>Changing episode length</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short episode dataset (k = −0.6; mean length = 2.4 d)</td>
<td>391</td>
<td>486</td>
<td>446</td>
</tr>
<tr>
<td>Long episode dataset (k = −0.25; mean length = 4.6 d)</td>
<td>382</td>
<td>404</td>
<td>395</td>
</tr>
<tr>
<td>Changing incidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low incidence dataset (α = 0.3; mean number of episodes = 1.6)</td>
<td>636</td>
<td>698</td>
<td>673</td>
</tr>
<tr>
<td>High incidence dataset (α = 0.56; mean number of episodes = 3.0)</td>
<td>325</td>
<td>358</td>
<td>347</td>
</tr>
</tbody>
</table>

Sample size per group for the comparison of 2 mean longitudinal prevalence estimates assuming a 30% reduction in the intervention group, P = 0.05, 80% power.

* k is the parameter of the exponential distribution y = exp(kx) representing episode durations.

† α is the shape parameter of the gamma distribution representing the number of episodes in an individual (stretch parameter kept constant at β = 6).
colleagues proposed as a rule of thumb that visits should be separated by at least the average duration of episodes. Our analysis suggests that intervals between visits should be at least twice the average episode duration to maximize efficiency.

For this analysis, we made a number of simplifications that may affect the interpretation of the findings. We assumed perfect recall, without under- or overreporting of diarrhea occurrence within the chosen recall periods. A number of studies have previously suggested that applying either a 24- or 48-hour period results in a similarly high level of recall. There is also evidence that the recall process is complex, with illness more than 48 hours earlier being underreported, or remembered as having occurred more recently, possibly leading to over-reporting of diarrhea. However, although imprecise disease reporting may affect the size of the estimate, it is unlikely to affect the proportional loss of precision with decreasing number of visits, as identified in our analysis.

Further, the 2 real datasets were based on weekly/ twice-weekly household visits with recall periods of up to 7 days, whereas we assumed a shorter recall to allow comparison with the simulation. Also, the study populations in Guatemala and Brazil both displayed variations in the diarrhea incidence over time. However, we found that the findings from the real datasets were well reflected by our model (Fig. 5) and that the effect of seasonality was limited.

Finally, one may question the choice of mean longitudinal prevalence and its SD as our outcome measures. Although the mean is influenced by extreme values, it also represents, in the case of the longitudinal prevalence, the population prevalence for all days of observation. Also, as long as extreme values of longitudinal prevalence are plausible and not due to measurement error, they are of public health interest because they are associated with poor nutritional status and higher mortality. In this situation, presenting the mean longitudinal prevalence and comparing 2 groups with the t test may be a better approach than relying on nonparametric methods such as the median and the Wilcoxon test if the sample size is large.

Instead of the SD or the standard error of the mean, we could have used the standard error of illness-days as a proportion of all days observed, taking into account the clustering of illness days in individuals. We applied this approach to the data from Brazil and found that the standard errors with increasing intervals were almost exactly proportional to the SD and the standard error of the mean longitudinal prevalence. Likewise, we could have used sample size formulae commonly used for clustered data. Most of these formulae include a measure of the between-cluster variation, which enters the formula as square. Although these formulae might have resulted in slightly different overall sample sizes, the proportional changes would have been very similar, if not identical.

In conclusion, our analysis suggests that the longitudinal prevalence of diarrhea can be efficiently estimated by periodic sampling while minimizing expense and inconvenience to study participants. Our findings have implications for the longitudinal prevalence of other episodic conditions and symptoms, such as respiratory infections, cough or fever. Sampling only a fraction of days during a study period deserves to be tested in the field.

REFERENCES

Potential Confounding by Exposure History and Prior Outcomes
An Example From Perinatal Epidemiology
Penelope P. Howards,* Enrique F. Schisterman,* and Patrick J. Heagerty†

Abstract: Prior pregnancy outcomes, such as spontaneous abortion and preterm birth, are often predictive of future pregnancy outcomes. Therefore, many researchers adjust for reproductive history. Although this adjustment may be appropriate for a predictive model, it is not necessarily appropriate when the goal is to obtain an unbiased estimate of the effect of exposure on disease. Reproductive history may seem to meet the conventional criteria for confounding because it is unlikely to be on the causal pathway between exposure and current outcome, is often associated with current outcome, and may be associated with exposure as well. However, whether reproductive history is a confounder or not depends on the underlying reason for its associations with exposure and current outcome. Thus, conventional methods for assessing confounding are often inadequate. Directed acyclic graphs (DAGs) can be used to evaluate complex scenarios for confounding when the research question is clearly defined with respect to the exposure, the outcome, and the effect estimate of interest. Special care is required when reproductive history affects future exposure. We use 5 DAGs to illustrate possible relations between reproductive history and current outcome. We assess each DAG for confounding, and identify the appropriate analytic technique. We provide a numeric example using data from the Collaborative Perinatal Project. There is no single answer as to whether reproductive history should be included in the model; the decision depends on the research question and the underlying DAG.

Epidemiology 2007;18: 544–551

Reproductive outcomes, such as preterm birth, low birth weight (LBW), and spontaneous abortion, are often predictive of the outcomes of subsequent pregnancies,1–8 and therefore, reproductive history variables are often included as explanatory variables in analytic models. Although this inclusion may be appropriate for predictive models, there is controversy whether it is appropriate when the goal is to understand etiology.9–12 Analysts presumably include reproductive history in etiologic models in an attempt to control for confounding. Reproductive history may in fact seem to be a confounder in a given data set because it is likely to be associated with both the study outcome and the study exposure. (Reproductive history could even causally affect exposure: a poor pregnancy outcome could motivate women to change their behavior during subsequent pregnancies.) Because reproductive history is unlikely to be on the causal path between current exposure and the outcome of the index pregnancy, it may seem to meet the conventional criteria of a confounder. However, whether it actually is a confounder depends on the underlying causes of the associations observed in the data.

In this paper, we use directed acyclic graphs (DAGs)13–16 to explore when it is appropriate to adjust for reproductive history in order to get an unbiased estimate of an effect measure. DAGs are graphic devices that help to identify sources of bias, such as confounding. Confounding is created by noncausal, “backdoor paths” from the exposure of interest to the outcome through other variables. This article is not intended as an introduction to the use of DAGs; we merely use DAGs as a tool to help explore the question of when we should adjust for prior outcome or exposure history. More instructional papers ranging from sophisticated methodologic reports to descriptions of the use of DAGs in specific scenarios are available.13–15

In the following sections, we start with a discussion of how the research question affects the identification of potential confounders. We assume a situation where each woman has 2 pregnancies, and each pregnancy has an exposure that occurs before the outcome. We treat the second pregnancy as the “current” pregnancy and the first pregnancy as the “prior” or “past” pregnancy. When we are generally interested in the effect of exposure on the outcome for any pregnancy, we refer to the effect of the “proximal” exposure on the corresponding pregnancy.
Specifically, we consider 3 research questions. First, what is the effect of the current exposure on the current outcome? Then more generally, what is the effect of the proximal exposure on the corresponding outcome? Finally, what are the effects of the past and current exposure on the current outcome? For each question, we examine DAGs that represent 5 different underlying causal relations of interest to determine whether confounding is present, and if so, how it can be addressed through modeling. Finally, we use data from the Collaborative Perinatal Project to illustrate how an effect estimate changes based on what our research question is, which DAG is selected, and which model is fit.

DEFINING THE RESEARCH QUESTION

In this paper, we address the goal of obtaining an unbiased estimate of the average causal effect of the exposure on the outcome, in contrast to determining the best predictive model. For convenience, we assume that the data are not misclassified or missing, and that there is no correlation among women. Even if these conditions are met, measures of effect in observational studies cannot be interpreted as causal unless the exposed and the unexposed are exchangeable conditional on the measured variables. Such an assumption should not be made lightly. Clearly defining the research question with respect to the outcome, the exposure, and the specific effect of interest is a prerequisite to identifying variables that confound the effect estimate. Specifically, the appropriate treatment of the reproductive history variables may change depending on the research question. Although this statement may seem obvious, it is easy to have a research question that is too vague. For example, “What is the effect of smoking on LBW?” is not sufficiently clear. The timing of exposure status has an effect on the current outcome, or it could be whether the proximal exposure has an effect on the corresponding outcome (ie, multiple observations per person). The latter situation commonly occurs in pregnancy studies where data are available on current and prior outcomes, the researcher may be interested in all outcomes (ie, including multiple outcomes per person) or only 1 outcome per person. The latter situation commonly occurs in pregnancy studies where data are collected prospectively for a single pregnancy and more limited data are collected about prior pregnancies. Defining the exposure can be even more complex. For this paper, we will focus on scenarios where the researcher is interested in exposure status during the index pregnancy or exposure status during both the index pregnancy and the prior pregnancy. Once the outcome and exposure are defined, the type of confounding, and, therefore, the effect of E1 on O1 could be estimated using an unadjusted model. The link function for the model would be determined by the type of outcome. For example, the logistic model would be:

\[
\text{Logit } P(O_1 = 1) = \beta_0 + \beta_1 E_1
\]  

(1)

Adding O0 to the model would increase the model fit (because O0 is predictive of O1) but would likely also affect the effect estimate for E1 due to noncollapsibility of the odds ratio (OR). In other words, even though O0 is not a confounder, the stratum-specific ORs for the effect of E1 on O1 would not equal the pooled OR if E1 and O0 are associated within both strata of O1 (which could occur even if E1 was randomized and there was no association between E1 and O0 in the pooled data). However, the difference between the unadjusted OR and the OR adjusted for O0 would be small if the outcome was rare.

DAG 1—No Confounding

We start with a simple DAG with no confounding of the exposure’s effect on the outcome (Fig. 1, DAG 1). The lack of a backdoor path from E1 to O1 indicates the absence of confounding, and, therefore, the effect of E1 on O1 could be estimated using an unadjusted model. The link function for the model would be determined by the type of outcome. For example, the logistic model would be:

\[
\text{Logit } P(O_1 = 1) = \beta_0 + \beta_1 E_1
\]

Adding O0 to the model would increase the model fit (because O0 is predictive of O1) but would likely also affect the effect estimate for E1 due to noncollapsibility of the odds ratio (OR). In other words, even though O0 is not a confounder, the stratum-specific ORs for the effect of E1 on O1 would not equal the pooled OR if E1 and O0 are associated within both strata of O1 (which could occur even if E1 was randomized and there was no association between E1 and O0 in the pooled data). However, the difference between the unadjusted OR and the OR adjusted for O0 would be small if the outcome was rare.

DAG 2—Simple Confounding

DAG 2 (Fig. 1) builds on DAG 1 through the inclusion of an arrow from E0 to E1. This addition causes the effect of E1 on O1 to be confounded by the backdoor path
E₁ ← E₀ → O₀ → O₁. Theoretically, that path could be blocked by including either E₀ or O₀ in the model:

\[
\text{Logit } P(O₁ = 1) = \beta₀ + \beta₁E₁ + \beta₂E₀
\]  

\[
\text{Logit } P(O₁ = 1) = \beta₀ + \beta₁E₁ + \beta₂O₀
\]  

Thus, even if E₀ was unknown, an unbiased effect of E₁ on O₁ could be estimated if O₀ was available.

**DAG 3—Confounding by O₀**

DAG 3 (Fig. 1) builds on DAG 2 by including an arrow from O₀ to E₁. In this case, the effect of E₁ on O₁ is confounded by 2 backdoor paths: E₁ ← E₀ → O₀ → O₁ and E₁ ← O₀ → O₁. The simplest model would include O₀ as a confounder because doing so would block both backdoor paths:

\[
\text{Logit } P(O₁ = 1) = \beta₀ + \beta₁E₁ + \beta₂O₀
\]  

Controlling for E₀ alone would not produce an unbiased estimate of the effect of E₁ on O₁.

**DAG 4—Complex Confounding**

In DAG 4 (Fig. 1), we replace the arrow between E₀ and E₁ in DAG 3 with a covariate (C₀) that causes exposure at both times. In addition, we replace the arrow between O₀ and O₁ in DAG 3 with a different covariate (C₁) that causes both outcomes. Here, the effect of E₁ on O₁ is confounded by the backdoor path E₁ ← O₀ ← C₁ → O₁. The simplest solution to this problem would be to include C₁ in the model:

\[
\text{Logit } P(O₁ = 1) = \beta₀ + \beta₁E₁ + \beta₂C₁
\]  

If C₁ was unavailable or unknown, it would be tempting to include O₁ in the model because it would be associated with both E₁ and O₁ in our data and it is not an intermediary. However, adding O₀ to the model without any other covariates would not yield an unbiased effect estimate because O₀ is a collider.

A collider is a variable with 2 arrows going into it from 2 parent variables. In DAG 4, E₀ and C₀ are parents of O₀. Any path through a collider is blocked, but when a collider is controlled for (Fig. 1, DAG 4b), it induces an association between the collider’s parents in at least 1 stratum.¹³,¹⁴ Thus, the path E₁ ← C₀ → E₀ → O₀ ← C₁ → O₁ is blocked by O₀ if O₀ is not included in the model, but adding O₀ to the model opens the previously nonexistent backdoor path E₁ ← C₀ → E₀ – C₁ → O₁. Therefore, the effect of E₁ on O₁ would be biased if O₀ was the only other covariate in the model. However, if C₀, E₀, or C₁ were added as well (Equations 6–8), the effect estimate would be unbiased:

\[
\text{Logit } P(O₁ = 1) = \beta₀ + \beta₁E₁ + \beta₂O₀ + \beta₃C₂
\]  

\[
\text{Logit } P(O₁ = 1) = \beta₀ + \beta₁E₁ + \beta₂O₀ + \beta₃E₀
\]  

\[
\text{Logit } P(O₁ = 1) = \beta₀ + \beta₁E₁ + \beta₂O₀ + \beta₃C₁
\]
The most efficient model would include only E₁ and Cₙ (Equation 5). If Cₙ and C₁ₙ were unknown or unavailable, the effect of E₁ on O₁ could still be estimated without bias if both E₀ and O₁ were available, but not if O₀ was the only available covariate.

It is worth emphasizing that if either DAG 3 or DAG 4 were true, an association between O₀ and E₁ and an association between O₁ and O₀ would be observed in the data. However, the latter association would be causal only if DAG 3 was true. In addition, including only O₀ in the model to control for confounding would be appropriate for DAG 3, but would not be sufficient for DAG 4.

**DAG 5—More Complex Confounding**

In DAG 5 (Fig. 1), we allow E₀ to have a direct effect on O₁. This relation allows for 2 backdoor paths from E₁ to O₁ (E₁ ← O₀ ← Cₙ → O₁ and E₁ ← Cₙ → E₀ → O₁), and once again, O₀ is a collider. There is no single variable that can block both paths, but there are several pairs of variables that could control for confounding, including Cₙ and C₁ₙ (Equation 9), E₀ and Cₙ (Equation 10), O₀ and Cₙ (Equation 6), or O₀ and E₀ (Equation 7).

\[
\text{Logit } P(O_1 = 1) = \beta_0 + \beta_1 E_1 + \beta_2 C_n + \beta_3 C_b \quad (9)
\]

\[
\text{Logit } P(O_1 = 1) = \beta_0 + \beta_1 E_1 + \beta_2 E_0 + \beta_3 C_b \quad (10)
\]

Failure to add one of those pairs of variables would result in a biased effect estimate.

**Effect of Proximal Exposure on Outcome**

For our second research question, we want to identify more generally the effect of the proximal exposure (Eᵢ, where i refers to the ith pregnancy) on the corresponding outcome (Oᵢ). In other words, we are interested in the short-term effect of the exposure on the outcome, regardless of pregnancy order. We could do this by including both pregnancies in our model and adjusting for the lack of independence of the outcomes using generalized estimating equations with robust standard errors, mixed models, or a comparable method.²⁴,²⁵

For the 5 DAGs chosen for this paper, confounding is present only for the second pregnancy, if at all. Therefore, the same confounders would need to be adjusted for as described in the previous sections. This adjustment could be done by including an indicator variable to distinguish the first pregnancy and setting the reproductive history variables equal to zero for the first pregnancy. When there is no confounding, no indicator variable is necessary, and so for DAG 1, the generalized estimating equations corresponding to Equation 1 are appropriate. For DAG 2, Equation 2 would be modified to:

\[
\text{Logit } P(O_i = 1) = \beta_0 + \beta_1 E_i + \beta_2 E_{i-1} + \beta_1 J \quad (11)
\]

where J equals 1 for the first pregnancy and 0 for the second pregnancy and E₁₋₁ is set to 0. Equations 3 through 10 could be modified similarly and fit using generalized estimating equations.

**The Effects of E₀ and E₁ on O₁**

Finally, we are interested in the effects of E₀ and E₁ on O₁. We have chosen to examine the effects of E₀ and E₁ separately; in actual studies, the cumulative or joint effect might be of interest.

**DAG 1—No Confounding**

The effects of E₀ and E₁ on O₁ are not confounded in DAG 1, so an unadjusted model including both exposures (Equation 2) can be used.

**DAG 2—Confounding of E₁ Addressed by E₀**

The backdoor path from E₁ to O₁ in DAG 2 (E₁ ← E₀ → O₀ → O₁) goes through E₀, which is part of our exposure of interest. Thus, although the effect of E₁ on O₁ is confounded by the backdoor path through the reproductive history variables, this confounding is addressed by the inclusion of E₀ in the model (Equation 2). Even though including O₀ blocked the backdoor path for the research questions above, it would introduce bias here because O₀ is an intermediary between E₀ and O₁. Therefore, O₀ cannot be included as a covariate in this case.

**DAG 3—Complex Confounding by O₀**

In DAG 3, there are 2 backdoor paths from E₁ to O₁ (E₁ ← E₀ → O₀ → O₁ and E₁ ← O₀ → O₁) including one that goes through O₀ but not E₀. Therefore, unlike DAG 2, the fact that E₀ is already in the model does not prevent confounding. Although including O₀ in the model would prevent confounding of the effect of E₁ on O₁, it cannot be included in the model because it is also an intermediate between E₀ and O₁. As a result, the effects of E₀ and E₁ on O₁ cannot be estimated without bias using conventional methods. Instead, more sophisticated models, such as marginal structural models, are required.

Marginal structural models are an alternative method to adjust for confounding. They could be used in place of standard methods in any of the analyses described above, but they are particularly appropriate when there is a confounder that is affected by prior exposure (eg, an intermediate variable). For the rest of this paper, we will refer to such variables as time-dependent confounders; technically, we should use the phrase “time-dependent confounders affected by prior exposure,” but because there are no other time-dependent confounders in our examples, we will shorten the phrase. There are a number of accessible papers on marginal structural models as well as a wealth of more sophisticated references.²⁶–³¹ We defer to these sources for an in-depth description of marginal structural models but provide a brief overview here.

Assuming no unmeasured or time-dependent confounders, conventional adjusted models produce unbiased effect estimates. However, in the presence of time-dependent confounders affected by prior exposure, both adjusted and unadjusted standard models are biased for causal effects. Marginal structural models resolve this issue through the use of inverse-probability weights that break the association between the exposure and the time-dependent confounder. Adjusting
for confounding through marginal structural models is a 2-stage process: first the weights must be estimated and then the effect of the exposure is estimated. Each observation is assigned a weight equal to the inverse probability of getting the observed exposure conditional on confounder status. The true weights are unknown, but they may be estimated using the inverse of the predicted probabilities from a model where exposure status is predicted by the measured confounders. Stabilized weights are recommended to improve the estimation precision. They are calculated by multiplying the inverse-probability weights by the probability of having the observed exposure conditional on any non–time-dependent confounders.

In the second stage, the effect estimate model is fit without the confounders but with each woman weighted by her stabilized inverse-probability weights. Essentially, the weights create a pseudopopulation with copies of each woman corresponding to the value of her weight. In the pseudopopulation, there is no association between the exposure and the time-dependent confounders, so the association between the exposure and the outcome is unconfounded. Assuming no unmeasured confounders (as well as other standard assumptions), the effect estimate from the marginal structural models is equal to the causal effect estimate.

For DAG 3, the first-stage model would use the history variables to predict exposure:

\[ \text{Logit } P(E_i = 1) = \beta_0 + \beta_1 E_{i-1} + \beta_2 O_{i-1} + \beta_3 J \]  

(12)

where \( E_{-1} \) and \( O_{-1} \) are set to 0. In addition, for stabilized inverse-probability weights, the predicted probability of the woman having the exposure status she has would be calculated using:

\[ \text{Logit } P(E_i = 1) = \beta_0 + \beta_1 E_{i-1} + \beta_2 J \]  

(13)

Then, each woman's stabilized inverse-probability weight could be calculated using the predicted probabilities from both equations and the following formula:

\[ \text{sIPW} = \frac{P(E_0)P(E_1 | E_0)}{P(E_0)P(E_1 | E_0, O_0)} \]  

(14)

In this case, there are no covariates that predict \( E_0 \), but if there were covariates that predicted \( E_0 \) or other covariates that predicted \( E_1 \), they would be added to equations 12 and 13. The resulting predicted probabilities would be used in the denominator of equation 14. The second stage model would take the form of equation 2 with each woman weighted by her stabilized inverse-probability weight from equation 14.

**DAG 4—Simpler Confounding**

If we are interested in the effects of \( E_0 \) and \( E_1 \) on \( O_1 \), DAG 4 is actually simpler to solve than DAG 3. \( O_0 \) continues to be an intermediary between \( E_0 \) and \( O_1 \) and part of the only backdoor path from \( E_1 \) to \( O_1 \) (\( E_1 \rightarrow O_0 \leftarrow C_b \rightarrow O_1 \)), but the backdoor path also includes \( C_b \). Therefore, that path can be blocked by including \( C_b \) in the model (equation 10). If \( C_b \) was unknown or unavailable, then marginal structural models could be used instead with \( O_0 \) predicting \( E_1 \) in the first stage model as in DAG 3.

**DAG 5—Also Simple Confounding**

There are 2 backdoor paths from \( E_1 \) to \( O_1 \) in DAG 5 (\( E_1 \leftarrow O_0 \leftarrow C_b \rightarrow O_1 \) and \( E_1 \leftarrow C_a \rightarrow E_0 \rightarrow O_1 \)), but when the effects of \( E_0 \) and \( E_1 \) on \( O_1 \) are of interest, \( E_0 \) blocks one of the paths. As with DAG 4, \( C_b \) blocks the second path (equation 10), or marginal structural models could be fit if \( C_b \) was not available but \( O_0 \) was.

**A CASE STUDY USING DATA FROM THE COLLABORATIVE PERINATAL PROJECT**

For each research question, we examined how the effect estimate changed under each of the 5 DAGs using a subset of data from the Collaborative Perinatal Project. Our purpose is purely illustrative and we make no substantive conclusions; we therefore limit our description of the Collaborative Perinatal Project to features relevant to the data set we used. This restricted data set included 2211 women who each contributed 2 consecutive, prospectively observed pregnancies. Each woman reported no pregnancy before her first observed pregnancy.

The outcome was LBW (<2500 g vs. 2500+ g). Our exposure was the use of cigarettes during pregnancy (categories: nonsmoker; smoker of <1 cigarette/d, 1 cigarette/d to <1 pack/d, and 1+ pack/d). In addition, we had data on the following covariates: maternal age (<20, 20–24, >29 vs. 20–24 years), maternal race (nonwhite vs. white), prepregnancy maternal weight (kg), family income (unknown, <$2000, $2000–$3999, $4000–$5999 vs. >$6000), infant sex, clinical site (12 categories), and pregnancy order (1st or 2nd). For the second pregnancy, we created pregnancy history variables indicating the outcome of the first pregnancy and the smoking status during the first pregnancy. The pregnancy history variables were set to zero for the first pregnancy.

As expected, there is a strong association between prior and current birth weight outcomes [unadjusted OR = 4.1; 95% confidence interval (CI) = 3.1–5.4]. We do not know if the association is causal (DAGs 1–3) or due to a common cause (DAGs 4 and 5). A common cause seems the most likely. There is also a strong association between past and current smoking (38.8; 30.4–49.5 for smoking as a dichotomous variable). Again, it is likely that this association is due to a common cause (DAGs 4 and 5) but we cannot test whether the association is causal. The unadjusted OR for prior LBW and dichotomous current smoking is 1.5 (1.1–2.0), which persists even when adjusting for prior smoking (1.5; 1.0–2.3). Prior LBW could conceivably affect smoking status during the second pregnancy (DAGs 3–5) if women modify their behaviors based on prior outcomes; however, one would expect an OR <1.0 if women with LBW births stopped smoking.

For each research question, we ran the appropriate analysis for each DAG. The ORs and 95% confidence intervals for the effect of current (2nd pregnancy) smoking status...
Aside from small changes among women who smoked >1 cigarette per day during the second pregnancy, the ORs do not change across models. There seems to be a null effect of prior smoking on current outcome, which suggests that prior smoking status does not confound this research question, according to conventional criteria for evaluating confounding, despite the strong association between prior and current smoking status. Further, based on a simple evaluation of confounding, outcome history seems to be a weak confounder. Both DAG 3 and DAG 4 would be consistent with an observed association between prior LBW and current LBW in the data, but the solutions to the 2 DAGs differ. If smoking history is included in the model (as required by DAG 4) or not (DAG 3), the ORs do not change, which implies that DAG 4 is incorrect because the OR from the model excluding prior smoking should be biased if DAG 4 is correct.

One challenging characteristic of DAGs is that they do not provide quantitative insight into confounding. This limitation is highlighted by the inability to quantify the extent of confounding through DAGs. Therefore, additional methods, such as sensitivity analysis, may be necessary to fully assess the extent of confounding in these scenarios.

### TABLE 1. Effect of Current Smoking (E₁) on Current Low Birth Weight Outcome (O₁), Unadjusted and Adjusted for Prior Smoking (E₀) and Prior Low Birth Weight Outcome (O₀), Using Data From 2211 Women in the Collaborative Perinatal Project Who Each Had 2 Prospectively Observed Pregnancies

<table>
<thead>
<tr>
<th>Variable</th>
<th>DAG 1</th>
<th>DAG 2</th>
<th>DAGs 2 and 3</th>
<th>DAGs 4 and 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current smoking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsmoker*</td>
<td>1.0 (1.0–1.0)</td>
<td>1.0 (1.0–1.0)</td>
<td>1.0 (1.0–1.0)</td>
<td>1.0 (1.0–1.0)</td>
</tr>
<tr>
<td>&lt;1 cigarette/d</td>
<td>0.9 (0.3–2.9)</td>
<td>0.8 (0.2–2.8)</td>
<td>0.8 (0.2–2.8)</td>
<td>0.8 (0.2–2.7)</td>
</tr>
<tr>
<td>≥1 cigarette/d to &lt;1 pack/d</td>
<td>1.5 (1.0–2.0)</td>
<td>1.4 (0.9–2.3)</td>
<td>1.4 (1.0–2.0)</td>
<td>1.4 (0.9–2.2)</td>
</tr>
<tr>
<td>≥1 pack/d</td>
<td>1.7 (1.2–2.6)</td>
<td>1.9 (1.1–3.5)</td>
<td>1.5 (1.0–2.2)</td>
<td>1.5 (0.8–2.9)</td>
</tr>
<tr>
<td>Smoking during prior pregnancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsmoker*</td>
<td>—</td>
<td>1.0</td>
<td>—</td>
<td>1.0</td>
</tr>
<tr>
<td>&lt;1 cigarette/d</td>
<td>—</td>
<td>1.3 (0.5–3.6)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>≥1 cigarette/d to &lt;1 pack/d</td>
<td>—</td>
<td>1.0 (0.6–1.6)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>≥1 pack/d</td>
<td>—</td>
<td>0.8 (0.4–1.6)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Outcome history</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not low birth weight*</td>
<td>—</td>
<td>—</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Low birth weight</td>
<td>—</td>
<td>—</td>
<td>5.2 (3.6–7.5)</td>
<td>5.2 (3.6–7.5)</td>
</tr>
</tbody>
</table>

*Reference group.

### TABLE 2. Effect of Proximal Smoking Status (Eᵢ) on the Corresponding Low Birth Weight Outcome (Oᵢ), Unadjusted and Adjusted for Prior Smoking (Eᵢ₋₁) and Prior Low Birth Weight Outcome (Oᵢ₋₁)

<table>
<thead>
<tr>
<th>Variable</th>
<th>DAG 1</th>
<th>DAG 2</th>
<th>DAGs 2 and 3</th>
<th>DAGs 4 and 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current smoking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsmoker*</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>&lt;1 cigarette/d</td>
<td>0.7 (0.3–1.8)</td>
<td>0.7 (0.3–1.7)</td>
<td>0.7 (0.3–1.8)</td>
<td>0.7 (0.3–1.7)</td>
</tr>
<tr>
<td>≥1 cigarette/d to &lt;1 pack/d</td>
<td>1.4 (1.1–1.8)</td>
<td>1.4 (1.1–1.8)</td>
<td>1.4 (1.1–1.8)</td>
<td>1.4 (1.1–1.8)</td>
</tr>
<tr>
<td>≥1 pack/d</td>
<td>1.6 (1.2–2.2)</td>
<td>1.6 (1.1–2.3)</td>
<td>1.5 (1.1–2.0)</td>
<td>1.5 (1.0–2.1)</td>
</tr>
<tr>
<td>Smoking during prior pregnancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsmoker*</td>
<td>—</td>
<td>1.0</td>
<td>—</td>
<td>1.0</td>
</tr>
<tr>
<td>&lt;1 cigarette/d</td>
<td>—</td>
<td>1.4 (0.6–3.5)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>≥1 cigarette/d to &lt;1 pack/d</td>
<td>—</td>
<td>1.1 (0.8–1.5)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>≥1 pack/d</td>
<td>—</td>
<td>0.6 (0.6–1.6)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Outcome history</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior pregnancy not low birth weight*</td>
<td>—</td>
<td>—</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Prior pregnancy of any birth weight*</td>
<td>1.0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>No prior pregnancy</td>
<td>—</td>
<td>—</td>
<td>1.3 (1.0–1.6)</td>
<td>1.3 (1.0–1.7)</td>
</tr>
<tr>
<td>Prior pregnancy low birth weight</td>
<td>—</td>
<td>1.0 (0.8–1.3)</td>
<td>5.2 (3.6–7.5)</td>
<td>5.2 (3.6–7.5)</td>
</tr>
</tbody>
</table>

*Reference group.

Generalized estimating equations with robust standard errors were used to adjust for the lack of independence between pregnancy outcomes within a woman.
tation is highlighted in DAG 2. In theory, smoking history or LBW history should control for confounding if DAG 2 is true. However, the ORs for smokers in the highest category change depending on which reproductive history variable is included in the model. This change may be because DAG 2 is incorrect, or because of sampling variability.

Each woman contributed 2 observations in the analysis of the general effect of proximal exposure on outcome. We used generalized estimating equations with robust standard errors (independence working correlation) to account for the correlation of pregnancies within a woman. An indicator variable was included to distinguish first pregnancies from second pregnancies, and the value of reproductive history variables was set to zero for all first pregnancies. The ORs for the effect of the proximal smoking status on the corresponding LBW outcome was consistent across models (Table 2). These results had narrower confidence intervals than Table 1, even accounting for the lack of independence within a woman, because there were twice as many pregnancies. As with the prior hypothesis, previous smoking history seems to have no association with the later outcome, and prior LBW seems to be strongly associated with current outcome. The ORs are comparable to Table 1 because the DAGs had the same interpretation and therefore the same rules for adjusting for confounding.

The analyses for the effect of past and current smoking status on current LBW outcome differed from the previous 2 research questions. For DAGs 1 and 2, there was no confounding after both exposure variables were included in the model, but DAGs 3 to 5 required marginal structural models. Nevertheless, the actual ORs did not vary substantially across models (Table 3). This lack of variation may be because prior LBW only weakly predicts current smoking status.

**DISCUSSION**

There is no single answer as to whether reproductive history variables should be included as covariates in a regression model. Such a decision depends on the research question, and on the unknown causal structure. Even when data on the reproductive history variables are not available, it is important to consider whether the effect estimate would be biased with or without adjustment. This can be done by carefully defining the research question with particular attention to the definition of the exposure, the outcome, and the effect estimate of interest. Next, possible causal structures can be represented by DAGs, which should then be evaluated for potential confounders. DAGs remind us that an association in the data is not necessarily a causal association, and that different scenarios may have different solutions. DAGs also provide the opportunity to identify colliders, which require special consideration. If the research question addresses past and current exposures, DAGs can help identify time-dependent confounders affected by prior exposure as well. Once the research question is clear and the likely DAG identified, the appropriate analytic technique can be determined. Inappropriate treatment of reproductive history variables, particularly those strongly associated with the exposure and the outcome, can introduce substantial bias.

**ACKNOWLEDGMENTS**

We thank Robert Platt and Jennifer Hutchence for comments on an earlier draft.

**REFERENCES**

The Importance and Role of Intracluster Correlations in Planning Cluster Trials

John S. Preisser,* Beth A. Reboussin,† Eun-Young Song,‡ and Mark Wolfson‡

Abstract: There is increasing recognition of the critical role of intracluster correlations of health behavior outcomes in cluster intervention trials. This study examines the estimation, reporting, and use of intracluster correlations in planning cluster trials. We use an estimating equations approach to estimate the intracluster correlations corresponding to the multiple-time-point nested cross-sectional design. Sample size formulae incorporating 2 types of intracluster correlations are examined for the purpose of planning future trials. The traditional intracluster correlation is the correlation among individuals within the same community at a specific time point. A second type is the correlation among individuals within the same community at different time points. For a “time × condition” analysis of a pretest–posttest nested cross-sectional trial design, we show that statistical power considerations based upon a posttest-only design generally are not an adequate substitute for sample size calculations that incorporate both types of intracluster correlations. Estimation, reporting, and use of intracluster correlations are illustrated for several dichotomous measures related to underage drinking collected as part of a large nonrandomized trial to enforce underage drinking laws in the United States from 1998 to 2004.

(Epidemiology 2007;18: 552–560)

Cluster trials evaluate interventions delivered to intact social groups or clusters, such as communities, churches, schools, workplaces, and medical practices, whereas outcomes are measured on members of those groups.1,2 The distinctive feature of cluster trials is the presence of intracluster (or intraclass) correlation among members within groups that arises from restricting assignment of the interventions to groups instead of to individuals. Failure to account for the intracluster correlation within clusters will likely lead to 2 shortcomings: an underpowered study and inflated Type I error rate of hypothesis tests relating to the intervention.1,2 Proper planning of cluster trials is based upon sample size formulae that use hypothesized intracluster correlation values, often based on estimates from earlier trials.3–6

Although sample size formulae based upon the intracluster correlation for a posttest-only cluster trial design are fairly well established, many cluster trials have multiple time points for which more than one type of intracluster correlation arises. In the case of a pretest–posttest nested cross-sectional design, characterized by different groups of individuals within clusters sampled at 2 time points, 2 intracluster correlations may be defined. The traditional intracluster correlation is the correlation among individuals within the same community at a specific time point. A second intracluster correlation is the correlation among individuals within the same community at different points in time. Extensions of the posttest-only sample size formulae are needed to refine sample size determination for multiple time point designs.

The sample size formula adopted in planning a cluster trial depends upon the statistical analysis to be used and, more specifically, the test statistic for the null hypothesis of “no intervention effect.” The proposed analysis will depend upon the trial’s experimental design, including whether the intervention condition is randomized to groups. For example, in a randomized pretest–posttest nested cross-sectional design, the trial planners are concerned with how the statistical analysis should handle the pretest data to maximize statistical power for the test of intervention, given that a sufficiently large number of clusters are enrolled to ensure baseline balance among covariates and outcome. In short, the focus is on the comparison of posttest means. In contrast, potential bias from baseline imbalance is an additional concern in planning a nonrandomized cluster trial. In this case, a “time × condition” analysis may be appropriate because its test statistic is the difference in change in the mean outcome over time between intervention and control conditions. If the pretest intervention and control condition means are equal, a time × condition analysis will tend to have lower statistical power than a posttest only analysis because the former uses a test statistic that is a function of 4 means, resulting in greater variance than a test statistic that is a function of only 2 (posttest) means. In sum, the manner of accounting for pretest information in the analysis of a randomized cluster trial has implications for statistical power, whereas the choice may affect both power and bias for nonrandomized trials.
This article discusses the estimation of intracluster correlations and particularly their use in sample size formulae in the design of cluster trials. Although statistical analysis and power have received attention for modeling change in Gaussian outcomes, insufficient consideration has been given to binary outcomes. Focusing on large cluster trials with binary outcomes, a generalized estimating equations (GEE) approach is applied for the estimation of the 2 types of intracluster correlation. In planning a nonrandomized cluster trial with a pretest–posttest nested cross-sectional design, the paper discusses a sample size formula for a GEE time × condition analysis that incorporates the 2 types of intracluster correlations. Finally, the paper reports intracluster correlation estimates and estimates of their precision for several dichotomous measures of underage drinking from a large nonrandomized cluster trial to enforce underage drinking laws.

**METHODS**

The Enforcing Underage Drinking Laws Program

The Enforcing Underage Drinking Laws Program, launched by the United States Office of Juvenile Justice and Delinquency Prevention in 1998, is the largest federal initiative focused on reducing underage drinking in United States history. Each of the 50 states and the District of Columbia received significant funding and technical assistance to support state and local efforts to enforce laws related to alcohol use by underage persons and to prevent underage drinking. A major component of the program involved discretionary grants provided to states on a competitive basis. Selected states awarded subgrants to communities (cities or counties) according to criteria that varied across states. Randomization was not employed, and the evaluation team had no control over the choice of intervention communities. However, control communities with characteristics similar to the intervention communities were selected by identifying those with propensity scores similar to the intervention communities.

The propensity score was a scalar summary of several community characteristics captured by federal census and other external sources, and identified a priori as being likely related to the grantee selection process and the major outcomes measured by the youth survey.

The evaluation of the impact of the discretionary grants, or “national evaluation,” was conducted with a nested cross-sectional design. Data were collected using a variety of surveys; the focus in this work is on 3 annual telephone surveys using distinct samples of between 15 and 20 youths (age 16–20 years) in each selected community. For each community, data collection was conducted preintervention (or early in the intervention period), 1 year later, and 2 years later. A nested cross-sectional design was chosen over a nested cohort design because the long intervention period might result in substantial dropout in a cohort design. Also, interest focused on the change in the population over time as opposed to within-individual change.

The national evaluation was conducted using data collected during 3 funding cycles. The first began in 1999 with 52 intervention and 52 control communities in 9 states. The second began in 2000 with 16 intervention and 16 control communities from 7 states. The final cycle began in 2002 with 34 intervention and 34 control communities from 8 states. All but 2 communities participated in exactly one funding cycle. A total of 10,865 observations from 202 communities were used in the analysis of underage drinking outcomes.

Nine dichotomous measures of underage drinking use, alcohol risk behaviors, and negative consequences of alcohol use from the survey of youth were analyzed (Table 1). The following individual and contextual community-level variables were examined as covariates because they may partly explain the magnitude of intraclass correlations for alcohol use behaviors: age, sex, and community-level variables that are possibly characteristic of disadvantaged communities, namely, percentage of households with female head with no husband, percentage foreign born, and median income.

Statistical Analysis

The sample was predominantly white and well-balanced with respect to sex (Table 2). Sixteen and seventeen-year-olds were over-represented in the sample compared with 18, 19, and 20-year-olds. Observed prevalence for each outcome is reported in Table 3. Analyses of the intervention, reported elsewhere, used GEE with the simple “exchangeable” working correlation matrix to fit time × condition logistic regression models; use of the “robust” empirical covariance estimator provided valid large sample inference even if the correlation structure was misspecified.

The estimation of intracluster correlations in this article employs an extension of GEE that jointly specifies one set of estimating equations for the parameters in the logistic model for the probability that an individual reports the behavior, and a second set of estimating equations to estimate the parameters in the correlation model. In this approach, applied separately for the various behavior outcomes, a correlation model based upon 2 intracluster correlations is specified, the “within-time” correlation between outcomes from different youths at the same time (α0), and the “between-time” correlation between 2 outcomes from different youths at different times (α1). The approach produces estimates of the standard error for both intracluster correlations as well as for their covariance. As in the ordinary GEE, the extended GEE approach provides valid inference for assessing intervention effects, assuming the marginal model for the probability of behavior is correctly specified, even if the correlation model is misspecified.

Three sets of logistic models are fit for the probability that a youth reports the behavior. An initial set of models includes the design variables of condition (ie, control versus intervention), time, and funding cycle, as well as their pairwise interactions. A second set of models includes these terms in addition to the individual characteristics of age and sex. A third and final set of models adds both individual and the community characteristics, which serves 2 purposes. First, it may help to address any postsample selection imbalances among covariates in this nonrandomized study. Second, it may partly explain the magnitude of within-cluster corre-
TABLE 1. Description of Alcohol Use Measures From the Youth Survey of the Enforcing Underage Drinking Laws Program, 1998–2004

<table>
<thead>
<tr>
<th>Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binge drinking</td>
<td>“Think back over the last 2 weeks. How many times have you had 5 or more drinks in a row?” Respondents who answered zero were coded “0” whereas those who answered one or more occasions were coded “1.”</td>
</tr>
<tr>
<td>DWI driving</td>
<td>Among ever-drinkers who reported ever driving, “During the last 30 days, how many times (if any) have you driven after drinking 2 or more drinks in an hour or less?” Those reporting zero, never drinkers, and ever drinkers who had never driven were coded “0”; those replying one or more occasions were coded “1.”</td>
</tr>
<tr>
<td>Past 30-day alcohol use</td>
<td>“When was the last time you drank alcohol?” respondents who answered “in the last 30 days” were coded “1.” Respondents who answered that they had consumed alcohol but not in the last 30 days, as well as never drinkers, were coded “0.”</td>
</tr>
<tr>
<td>Past 7-day use</td>
<td>Similar to “Past 30-day alcohol use”; response “in the last 7 days” coded “1.”</td>
</tr>
<tr>
<td>Attempt to purchase alcohol</td>
<td>“In the last 30 days, how many times did you try to buy alcohol from a bar, restaurant, or store (whether you were successful or not)?” Respondents who answered zero were coded “0”; those who answered one or more occasions were coded “1.”</td>
</tr>
<tr>
<td>Nonviolent consequences due to alcohol use</td>
<td>Experienced any of the following after they had been drinking: being cited or arrested for drinking, possessing alcohol, trying to buy alcohol, being cited or arrested for driving under the influence of alcohol; missing any school due to drinking; being warned by a friend about your drinking; passing out; being unable to remember what happened while drinking; breaking or damaging something; having a headache or hangover; being punished by your parents or guardian; and having sex without using birth control. Respondents who reported experiencing any these in the past year were coded “1”; others were coded “0.”</td>
</tr>
<tr>
<td>Perception of alcohol use among peers</td>
<td>“How many of your friends do you think have had any alcohol to drink in the past 30 days? Would you say none, a few, some, most, or all?” Respondents who reported that most of all or their friends had consumed alcohol in the past 30 days were coded “1”, while all others were coded “0.”</td>
</tr>
<tr>
<td>Perception of getting caught by police</td>
<td>“If you had been drinking, how likely would it be for the police to catch you? Very likely, somewhat likely, not very likely, or not at all likely?”, coded “1” if respondent answered “very likely” or “somewhat likely”; else coded “0.”</td>
</tr>
<tr>
<td>Commercial source of alcohol</td>
<td>The last time you drank any alcohol, how did you get the alcohol?” those who answered that they obtained alcohol from a “commercial source” (businesses such as alcohol outlets, restaurants, and bars) were coded as “1.” Those who answered that they were given alcohol by their friends, family members, coworkers, acquaintances, or strangers at home or at events were coded “0.”</td>
</tr>
</tbody>
</table>

TABLE 2. Sex, Race, and Age of Participants in the Youth Survey of the Enforcing Underage Drinking Laws Program for Control and Intervention Communities (All Rounds), 1998–2004

<table>
<thead>
<tr>
<th></th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (n = 1835)*</td>
<td>Intervention (n = 1784)*</td>
<td>Control (n = 1825)*</td>
</tr>
<tr>
<td>Male; %</td>
<td>49.1</td>
<td>50.0</td>
<td>50.3</td>
</tr>
<tr>
<td>Race; %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>85.4</td>
<td>87.5</td>
<td>86.6</td>
</tr>
<tr>
<td>Black</td>
<td>6.6</td>
<td>5.2</td>
<td>6.4</td>
</tr>
<tr>
<td>Other</td>
<td>8.0</td>
<td>7.3</td>
<td>7.0</td>
</tr>
<tr>
<td>Age (yrs); %</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>32.0</td>
<td>30.8</td>
<td>31.1</td>
</tr>
<tr>
<td>17</td>
<td>31.1</td>
<td>32.0</td>
<td>31.8</td>
</tr>
<tr>
<td>18</td>
<td>17.6</td>
<td>18.0</td>
<td>17.1</td>
</tr>
<tr>
<td>19</td>
<td>11.6</td>
<td>10.9</td>
<td>11.3</td>
</tr>
<tr>
<td>20</td>
<td>7.7</td>
<td>8.4</td>
<td>8.7</td>
</tr>
<tr>
<td>Community sample; mean†</td>
<td>18.0</td>
<td>17.8</td>
<td>17.9</td>
</tr>
</tbody>
</table>

*Total number of youth sampled across communities.
†Mean number of youth sampled per community.
Role of Intracluster Correlations

Sample Size Determination

A general approach to power calculations for cluster trials is based upon an analysis of community summary statistics according to the study design. The general set-up is to specify the hypothesis of interest as $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \mu_1 - \mu_2$ is the intervention effect and $\mu_h = E(S_{ih})$ is the expected value of the summary statistic, $S_{ih}$, for the $i$th community receiving the intervention ($h = 1$) or control ($h = 2$) treatment condition. Specific examples of $S_{ih}$ relevant in the application of GEE to the national evaluation data are discussed below. As communities are assumed to be statistically independent, deriving an expression for the variance of the community summary statistic in each condition (intervention and control) is the critical step in determining sample size. For the $i$th community receiving the $h$th condition, let $\sigma_h^2 = \text{Var}(S_{ih})$, and let $m$ be equal to the number of subjects in each community at each time-point. Constant variance within condition results from assuming that $m$ is constant. The number of communities needed per condition ($n$) to test the intervention using a two-sided test with $\alpha$ significance level and power $1 - \beta$ is

$$n = \frac{\left(\sigma_1^2 + \sigma_2^2\right)(z_{1 - \alpha/2} + z_{1 - \beta})^2}{\delta^2} \tag{1}$$

where $z_c$ is the $(100 \times c)$th percentile of the standard normal distribution. For small $n$, Equation 1 may be refined using the t-distribution. Let $\Phi(\cdot)$ define the cumulative distribution function of the standard normal distribution, and

$$d = \sum_{i=1}^{n} S_{ih}/n - \sum_{i=1}^{n} S_{ih}/n,$$

an unbiased estimator of $\delta$. Power $(1 - \beta)$ is

$$1 - \beta = \Phi\left(\frac{\delta}{\sqrt{\text{var}(d)}} - z_{1-\alpha/2}\right) \tag{2}$$

where $\text{var}(d) = (\sigma_1^2 + \sigma_2^2)/n$.

Of specific interest are sample size formulae for pretest-posttest nested cross-sectional designs as they pertain to binary outcomes. In this design, the total number of individuals sampled per community, or cluster size, is $2m$. Let $\pi_{ih}$ be the probability of the outcome for an individual at time $t$ ($t = 0$ for pretest, $t = 1$ for posttest) from a community having condition $h$. Subscript $i$ for the individual is not needed because the probability of outcome is assumed to depend only upon the treatment status and time point.

In randomized cluster trials with a moderate to large number of communities, a planned GEE analysis need not adjust for pretest since groups may be expected to be balanced with respect to outcomes and covariates as a result of randomization. Statistical inference may be based on the posttest-only logistic model

$$\logit(\pi_{ih}) = \beta_0 + \delta_i x_{1i} \tag{3}$$

where $x_1 = 1$ if $h = 1$ (intervention) and $0$ if $h = 2$ (control); and $\delta_0 = \mu_1 - \mu_2$, where $\mu_1 = \logit(\pi_{i1})$ and $\mu_2 = \logit(\pi_{i2})$, is the log odds ratio comparing odds of response in the posttest period for subjects in intervention communities to the odds of response for subjects in control communities. Suppose $S_{ih}$ is the logit of the observed proportion reporting the behavior of those sampled at posttest from the $ih$th community having condition $h$. Then $d$ is an approximately unbiased estimator of $\delta_0$ whose large sample variance depends upon the approximate variance of $S_{ih}$ under $H_1 : \delta_0 \neq 0$:

$$\sigma_d^2 = \frac{\phi}{m \nu_{ih}} \tag{4}$$

where $\nu_{ih} = \pi_{ih} (1 - \pi_{ih})$ and $\phi = 1 + (m - 1)\alpha_0$ is the design effect. When $\alpha_0 > 0$, $\phi$ represents a multiplicative increase on the sample size required in a cluster trial to obtain a given level of power relative to the sample size required in a clinical trial with randomization of individuals. Inserting Equation 4 into Equation 1 and substituting $\delta_0$ for $\delta$ gives

$$n = \frac{\phi(1/\nu_{1i} + 1/\nu_{2i})(z_{1 - \alpha/2} + z_{1 - \beta})^2}{m[\logit(\pi_{i1}) - \logit(\pi_{i2})]^2} \tag{5}$$

the posttest-only cluster trial design sample size for binary outcomes. Similarly, inserting Equation 4 into Equation 2 gives the power formula. The equivalency of Formula 5 to a general GEE method of sample calculation for the Wald test statistic corresponding to $\delta_0$ in Equation 3 is shown in the Appendix. Alternatively, one can conduct sample size calculations for a planned linear model for the binary outcome (ie, use of identity link in Equation 3) using Equation 6 of Preisser et al.12

For a nonrandomized cluster trial with a nested pretest-posttest cross-section design, Formula 5 may be inadequate because adjustment for pretest response is needed in the analysis due to baseline imbalance. While, for the national evaluation, the estimation of intracluster correlations is based upon a model applied to data from 3 time points, sample size considerations for a future nonrandomized cluster trial address the anticipated effect at a single follow-up with respect to baseline. Thus, the comparison of intervention and control communities is operationalized as a one-degree-of-freedom contrast for the difference in the change in expected outcome from pretest to posttest. An appropriate sample size formula targets the contrast

$$\delta_i = [\logit(\pi_{i1}) - \logit(\pi_{i0})] - [\logit(\pi_{i2}) - \logit(\pi_{i0})],$$

the regression coefficient for the time \times condition interaction in the logistic regression model for a youth’s response at time $t$ under the $ht$th condition:

$$\logit(\pi_{ih}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \delta_i(x_1 \times x_2) \tag{6}$$
\[ \delta_i = \text{the difference in log odds ratios for pretest/posttest changes between intervention and control conditions.} \]

\[ \text{Equivalently, } \exp(\delta_i) = \text{the multiplicative factor by which the pretest/posttest odds ratio for communities under the intervention condition exceeds that of the control condition.} \]

For a community \( i \) having condition \( h \), define \( S_{ih} \) as the logit of observed proportion reporting the behavior at posttest minus the logit of the observed proportion reporting the behavior at pretest.

The corresponding mean over all communities within a condition is \( \sum_i S_{ih}/n \), an approximately unbiased estimator of \( \mu_{ih} = \logit(\pi_{ih}) - \logit(\pi_{ih0}) \). The approximate variance of \( S_{ih} \) is:

\[ \sigma_{ih}^2 = \frac{1}{m} \left( \frac{1}{\nu_{ih}} + \frac{1}{\nu_{i0}} - \frac{2m\alpha_i}{\nu_{ih}\nu_{i0}} \right) \tag{7} \]

Inserting Equation 7 into Equation 1, and setting \( \delta_i \) gives the pretest-posttest cluster trial design sample size formula for binary outcomes

\[ n = \frac{m(\logit(\pi_{i1}) - \logit(\pi_{i0})) - (\logit(\pi_{21}) - \logit(\pi_{20}))^2}{\sigma_{i1}^2} \tag{8} \]

Note that the resulting expression depends upon values of the 4 probabilities \( \pi_{ih} \). Often researchers are unwilling or unable to specify particular values of the \( \pi_{ih} \). Then values of \( \pi_{ih} \) may be chosen to give conservative values of \( \sigma_{ih}^2 \) in Equation 4 or Equation 7 accordingly. Finally, inserting Equation 7 into Equation 2 and setting \( \delta_i \) gives the corresponding formula for power. Alternatively, one can conduct sample size calculations for a planned linear model analysis of the proportions (ie, GEE identity link function) using Equation 9 of Preisser et al \(^{12} \) and \( \delta_i = (\pi_{i1} - \pi_{i0}) - (\pi_{21} - \pi_{20}) \).

**RESULTS**

**Intracluster Correlation Estimates**

Table 3 provides intracluster correlation estimates for several binary outcomes from the national evaluation from the model based upon 3 timepoints and 3 funding cycles. Estimates obtained from the extended GEE approach for the within-time (\( \alpha_0 \)) and between-time intracluster correlation (\( \alpha_1 \)) as well as their 95% large sample confidence intervals are reported. The upper confidence bound for each intracluster correlation (eg, for \( \alpha_0 \)) is given by UCB = \( \hat{\alpha}_0 + 1.96 \) se(\( \hat{\alpha}_0 \)) and the lower confidence bound is given by LCB = \( \hat{\alpha}_0 - 1.96 \) se (\( \hat{\alpha}_0 \)), where se (\( \hat{\alpha}_0 \)) is the standard error based upon the robust variance estimator.\(^{11} \) The estimated covariance between \( \hat{\alpha}_0 \) and \( \hat{\alpha}_1 \) is also reported for the purpose of conducting power for a range of \( (\alpha_0, \alpha_1) \) as illustrated in the next section. Three different sets of intracluster correlations are reported based upon the covariates included in the model for the probability a youth reports a behavior. For each measure, the first row gives the intracluster correlations based upon the model for the probability of the behavior that depends only upon study design variables: time point (year), cycle, intervention versus control condition, and their two-way interactions. The second row gives intracluster correlation estimates based upon a probability model that also adjusts for covariates age and gender. The third row gives intracluster correlations based upon the logistic model that additionally adjusts for the community level variables.

Intracluster correlation estimates range from 0.003 to 0.026 with modest differences in within-time and between-time intracluster correlations. Within-cluster correlations estimates for binge drinking, past 30-day alcohol use and attempt to purchase alcohol are larger than estimates for those outcomes reported by some authors\(^4,17 \) but similar in magnitude to those reported by others.\(^5,18,19 \) Generally, correlations adjusting for individual and community variables are smaller than design-adjusted correlations or those adjusting for age and gender in addition to design variables. SAS/IML software for applying the extended GEE is available.\(^{20} \)

**Sample Size Results**

Consider a future cluster trial to reduce underage drinking. Suppose the primary outcome is past 30-day alcohol use, and a pretest-posttest (2 time-points) nested cross-sectional design is planned with equal numbers of intervention and control communities and \( m = 15 \) youth to be surveyed from each community at each time point. Assume \( \pi_{i0} = \pi_{i20} = 0.40 \) and \( \pi_{i1} = 0.30 \) (ie, \( \delta_i = 0.44 \)) such that a 25% decline in underage drinking is anticipated. Applying Equations 1 and 7, and using the unadjusted intracluster correlation values of \( \alpha_0 = 0.0261 \) and \( \alpha_1 = 0.0219 \) from Table 3, results in \( n = 48 \) communities per condition needed to provide approximately 80% power to detect the desired effect based upon a 2-sided test at the 0.05 significance level.

A sensitivity analysis of power considers a range of values for \( \alpha_0 \) and \( \alpha_1 \) using information in Table 3. Because combinations of the 2 types of intracluster correlation are required, the univariate confidence intervals reported in Table 3 have limited utility, since the joint 95% confidence region is not rectangular, but rather is defined by the ellipse

\[ \{\alpha \in R^2 : (\alpha - \hat{\alpha})' [\text{Cov}(\hat{\alpha})]^{-1} (\alpha - \hat{\alpha}) \leq z_{0.025}^2 \} \tag{9} \]

where \( \alpha = (\alpha_0, \alpha_1)' \), \( z_{0.025}^2 = 3.842 \). Equivalently, the 95% joint confidence region consists of \( \alpha \) values that satisfy

\[ \frac{(\alpha_0 - \hat{\alpha}_0)^2 \text{var}(\hat{\alpha}_0)}{\text{var}(\hat{\alpha}_0)\text{var}(\hat{\alpha}_1)} - 2(\alpha_0 - \hat{\alpha}_0)(\alpha_1 - \hat{\alpha}_1)\text{cov}(\hat{\alpha}_0, \hat{\alpha}_1) + (\alpha_1 - \hat{\alpha}_1)^2 \text{var}(\hat{\alpha}_1) \leq z_{0.025}^2 \tag{10} \]

Estimated variances for outcomes can be determined from Table 3, ie, \( \text{var}(\hat{\alpha}_0) = [(UCB - \hat{\alpha}_0)/1.96]^2 \) or \( \text{var}(\hat{\alpha}_1) = [LCB - \hat{\alpha}_1)/1.96]^2 \), and similarly for \( \alpha_1 \). Using national evaluation estimates for past 30-day alcohol use (ie, \( \hat{\alpha}_0 = 0.0261 \), \( \hat{\alpha}_1 = 0.0219 \), \( \text{var}(\hat{\alpha}_0) = 0.0000246 \), \( \text{var}(\hat{\alpha}_1) = 0.0000128 \), and \( \text{var}(\hat{\alpha}_1) = 0.0000186 \) gives the ellipse plotted in Figure 1 along with contours of power to show how power varies over the joint confidence region for \( (\alpha_0, \alpha_1) \). First note the
The ellipse represents the boundary of the joint 95% confidence region for $(\hat{\alpha}_0, \hat{\alpha}_1)$. Second, the box is the intersection of the individual 95% confidence intervals, ignoring the covariance between the 2 correlation estimates. Third, the numbers on the plot represent power to reject $H_0 : \delta_1 = 0$ in favor of $H_1 : \delta_1 \neq 0$, for the assumed values of probabilities described above. Fourth, the 3 bands within the ellipse indicate constant levels of power (77%, 80%, 83%) for different combinations of $\hat{\alpha}_0$ and $\hat{\alpha}_1$. Within the ellipse, power attains its highest value (85%) for the boundary values of $(\hat{\alpha}_0 = 0.0210, \hat{\alpha}_1 = 0.0250)$, and its lowest value (76%) for the boundary values of $(\hat{\alpha}_0 = 0.0311, \hat{\alpha}_1 = 0.0187)$. Thus, for a sample design with $n = 48$ and $m = 15$, the 95% joint confidence region for the 2 intracluster correlations indicates a range of power from 76% to 85%. This sensitivity analysis for power is a considerable refinement over a naive approach, represented by the box in Figure 1, that gives a range of power that is artificially low (71%) or high (90%). Given knowledge of the covariance of the 2 intracluster correlations, the joint elliptical confidence region approach represents a clear improvement in cluster trial planning.

### TABLE 3. Within-Time ($\hat{\alpha}_0$) and Between-Time ($\hat{\alpha}_1$) Intracluster Correlation Estimates (ICC) of Youth Alcohol Use Measures and Their 95% Confidence Intervals (CI) From the Youth Survey of the Enforcing Underage Drinking Laws Program, 1998–2004

<table>
<thead>
<tr>
<th>Measure (Prevalence*) and Independent Variables†</th>
<th>Within-Time ICC (95% CI)</th>
<th>Between-Time ICC (95% CI)</th>
<th>Covariance‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binge drinking (0.1656)</td>
<td>0.0185 (0.0048 to 0.0322)</td>
<td>0.0169 (0.0088 to 0.0251)</td>
<td>0.0000200</td>
</tr>
<tr>
<td>Design variables</td>
<td>0.0193 (0.0060 to 0.0325)</td>
<td>0.0172 (0.0081 to 0.0262)</td>
<td>0.0000229</td>
</tr>
<tr>
<td>Age, sex</td>
<td>0.0192 (0.0053 to 0.0330)</td>
<td>0.0166 (0.0069 to 0.0262)</td>
<td>0.0000226</td>
</tr>
<tr>
<td>Community variables</td>
<td>0.0027 (−0.0037 to 0.0092)</td>
<td>0.0074 (0.0019 to 0.0129)</td>
<td>0.00000252</td>
</tr>
<tr>
<td>Design variables</td>
<td>0.0038 (−0.0032 to 0.0109)</td>
<td>0.0070 (0.0010 to 0.0131)</td>
<td>0.00000438</td>
</tr>
<tr>
<td>Age, sex</td>
<td>0.0038 (−0.0033 to 0.0109)</td>
<td>0.0071 (0.0010 to 0.0133)</td>
<td>0.00000452</td>
</tr>
<tr>
<td>Community variables</td>
<td>0.0261 (0.0164 to 0.0358)</td>
<td>0.0219 (0.0134 to 0.0303)</td>
<td>0.0001028</td>
</tr>
<tr>
<td>Design variables</td>
<td>0.0251 (0.0156 to 0.0346)</td>
<td>0.0217 (0.0133 to 0.0301)</td>
<td>0.0001017</td>
</tr>
<tr>
<td>Community variables</td>
<td>0.0234 (0.0139 to 0.0328)</td>
<td>0.0197 (0.0112 to 0.0282)</td>
<td>0.0001012</td>
</tr>
<tr>
<td>Past 30-day alcohol use (0.3035)</td>
<td>0.0190 (0.0092 to 0.0288)</td>
<td>0.0158 (0.0088 to 0.0228)</td>
<td>0.00000745</td>
</tr>
<tr>
<td>Design variables</td>
<td>0.0204 (0.0106 to 0.0301)</td>
<td>0.0156 (0.0082 to 0.0229)</td>
<td>0.00000785</td>
</tr>
<tr>
<td>Community variables</td>
<td>0.0197 (0.0099 to 0.0295)</td>
<td>0.0143 (0.0064 to 0.0222)</td>
<td>0.00000880</td>
</tr>
<tr>
<td>Past 7-day alcohol use (0.2174)</td>
<td>0.0178 (0.0069 to 0.0286)</td>
<td>0.0184 (0.0086 to 0.0281)</td>
<td>0.00001064</td>
</tr>
<tr>
<td>Design variables</td>
<td>0.0088 (−0.0003 to 0.0180)</td>
<td>0.0099 (0.0030 to 0.0165)</td>
<td>0.00000582</td>
</tr>
<tr>
<td>Nonviolent consequences to alcohol use (0.3609)</td>
<td>0.0112 (0.0039 to 0.0184)</td>
<td>0.0115 (0.0056 to 0.0173)</td>
<td>0.00000500</td>
</tr>
<tr>
<td>Design variables</td>
<td>0.0106 (0.0054 to 0.0169)</td>
<td>0.0111 (0.0054 to 0.0169)</td>
<td>0.00000451</td>
</tr>
<tr>
<td>Community variables</td>
<td>0.0087 (0.0022 to 0.0152)</td>
<td>0.0092 (0.0040 to 0.0145)</td>
<td>0.00000309</td>
</tr>
<tr>
<td>Perception of alcohol use among peers (0.5967)</td>
<td>0.0222 (0.0124 to 0.0319)</td>
<td>0.0200 (0.0116 to 0.0284)</td>
<td>0.0000131</td>
</tr>
<tr>
<td>Design variables</td>
<td>0.0208 (0.0115 to 0.0300)</td>
<td>0.0196 (0.0115 to 0.0276)</td>
<td>0.00000116</td>
</tr>
<tr>
<td>Community variables</td>
<td>0.0204 (0.0112 to 0.0296)</td>
<td>0.0192 (0.0110 to 0.0273)</td>
<td>0.00000309</td>
</tr>
<tr>
<td>Perception of getting caught by police (0.3988)</td>
<td>0.0114 (0.0033 to 0.0195)</td>
<td>0.0125 (0.0061 to 0.0190)</td>
<td>0.00000672</td>
</tr>
<tr>
<td>Design variables</td>
<td>0.0118 (0.0035 to 0.0200)</td>
<td>0.0127 (0.0062 to 0.0192)</td>
<td>0.00000694</td>
</tr>
<tr>
<td>Community variables</td>
<td>0.0087 (0.0010 to 0.0164)</td>
<td>0.0097 (0.0038 to 0.0157)</td>
<td>0.00000492</td>
</tr>
<tr>
<td>Commercial source of alcohol (0.0730)</td>
<td>0.0163 (0.0036 to 0.0290)</td>
<td>0.0157 (0.0049 to 0.0265)</td>
<td>0.0000206</td>
</tr>
<tr>
<td>Design variables</td>
<td>0.0164 (0.0040 to 0.0288)</td>
<td>0.0159 (0.0058 to 0.0262)</td>
<td>0.0000190</td>
</tr>
<tr>
<td>Community variables</td>
<td>0.0069 (−0.0026 to 0.0165)</td>
<td>0.0064 (−0.0004 to 0.0132)</td>
<td>0.0000339</td>
</tr>
</tbody>
</table>

*Observed overall prevalence.  
†Design variables are condition, time, round and their pairwise interactions. Community variables are percent of household with female head with no husband, percent foreign born, and median income.  
‡Covariance between $\hat{\alpha}_0$ and $\hat{\alpha}_1$. 

© 2007 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.
Providing further rationale for use of the proposed 2 intracluster correlation sample size formula is the strong possibility that use of the posttest-only sample size formula when planning a pretest-posttest nested cross-sectional trial will underestimate the sample size needed to obtain a desired level of statistical power in a GEE time \times condition analysis. Consider the ratio \( r \) of variances under the respective designs, \( \text{var}(\theta) \) based upon Equation 7 divided by \( \text{var}(\theta) \) based upon Equation 4. A ratio near 1 indicates that the formula for the posttest-only design is a good substitute for the pretest-posttest design formula. Figure 2 shows that generally \( r > 1 \) indicating that use of the posttest-only formula underestimates the required sample size. Using results for past 30-day alcohol use and \( m = 15 \), \( r \) ranges from 1.37 (corresponding to the highest value of power in the ellipse of Fig. 1) to 1.47 (based upon GEE estimates of intracluster correlations) to 1.56 (corresponding to lowest power). Figure 2 indicates that, in the case of \( (\alpha_0 = 0.0210, \alpha_1 = 0.0250) \), \( r < 1 \) only when \( m \geq 30 \); note \( m = 30 \) and \( \alpha_0 = 0.0210 \) gives \( \phi = 1.61 \), a relatively large design effect. Comparatively, the observed design effect for past 30-day use, calculated as \( \phi = 1 + \alpha_0 (\bar{m} - 1) \), where \( \bar{m} \) is the average number of youth from a community sampled at a timepoint, was 1.46. The fact that past 30-day use had the largest design effect among the 9 measures in the national evaluation suggests that observing a sufficiently large design effect in any cluster trial such that \( r \leq 1 \) appears unlikely.

**DISCUSSION**

The utility of any sample size formulae for cluster unit trials depends upon the availability of intracluster correlation estimates for various outcomes. Given their typically high cost, many cluster trials enroll a small number of clusters. However, in many government and foundation-sponsored programs, such as the national evaluation, the resources available for the evaluation of the program may be distinct from funds for implementing the intervention, so a moderately large number of clusters may be studied. These larger studies offer a unique opportunity to report on the magnitude and precision of intracluster correlation estimates for health behavior outcomes.

Relative to equivalent general power methods for GEE,\(^{21}\) the sample size formulae for a time \times condition analysis of binary outcomes presented in this paper are easy to apply. This article emphasized sample size formulae based upon a planned GEE analysis using a logit link, whereas Preisser et al\(^{12}\) emphasized similar formulae based upon an identity link. As a rule, the choice of sample size formula should be based upon the planned statistical analysis. However, because the 4-parameter interaction model in Equation 6 places no structure on the 4 time \times condition probabilities, correlation estimates obtained with the logit link model (eg, those in Table 3) may be used in sample size formula in equation 8 for a planned logit analysis or in Formulae 3 and 9 of Preisser et al\(^{12}\) for an analysis using the identity link. The difference in the respective formulae pertains to a difference in the definition of the effect of interest as a contrast of logits, or a contrast of proportions. Applying the formula based upon the identity link would have given \( n = 50 \) communities per condition in the previous section instead of \( n = 48 \); it is not always the case that larger sample sizes will be required for the identity link.

One practical obstacle in applying the proposed sample size methods is that information regarding variances of in-
cluster correlations (usually, in the form of confidence intervals or standard errors) are only occasionally published, and it seems less realistic that covariances of the 2 types of intracluster correlation will be available. If the 2 intracluster correlations are approximately equal, the problem may be circumvented by conducting a sensitivity analysis of power using the pretest-posttest formula under the assumption that \( \alpha = \alpha_0 = \alpha_i \); in this case, knowledge of a single intracluster correlation and its standard error are sufficient.

This article addressed the question of whether Formula 5 of the posttest-only design may be substituted for Formula 8 in planning a nonrandomized pretest-posttest nested cross-sectional cluster trial. Direct comparison of variances under the 2 designs showed that the posttest-only formula is generally not an appropriate substitute for the pretest-posttest design formula and will likely underestimate the required sample size and lead to an underpowered cluster trial. In other words, for nonrandomized cluster trials, pretest adjustment with a time \( \times \) condition analysis is undertaken to address potential bias, at the price of loss of power. This is in contrast to the design and analysis of randomized cluster trials, where adjustment for pretest using a time \( \times \) condition analysis may not be preferred because of the increased variance associated with Equation 7 relative to Equation 4. Rather, because communities tend to be similar across conditions due to randomization, covariate adjustment of cluster baseline response means may be undertaken to increase power.\(^{22}\)

The Enforcing Underage Drinking Laws Program that provided the intracluster correlation estimates in Table 3 had limitations. The sample under-represented 19- and 20-year-old subjects compared with 16-, 17-, and 18-year-old subjects. Such selection bias may lead to biased intracluster correlation estimates. However, for the national evaluation, stratified analyses of intracluster correlation estimates (not reported) were similar across age groups. Another limitation is that the random digit dialing methodology used to conduct the survey is known to underrepresent ethnic/racial minorities and lower socioeconomic status individuals.\(^{23}\) These limitations should be considered when deciding whether to use the reported intracluster correlations in planning a future cluster trial.

Finally, there are limitations with respect to the extended GEE and proposed sample size methods. The simple sample size formulae applied to the national evaluation data are applicable only to cluster trial study designs without matching or stratification. For these more complicated designs, use of the general GEE power method\(^{21}\) is recommended. Finally, using the extended GEE to produce confidence intervals for intracluster correlations requires a large number of clusters (eg, 80 or more); otherwise, confidence intervals may suffer from severe undercoverage.\(^{24}\) Small-sample bias adjustments to estimation of intracluster correlations,\(^{25}\) and to their estimated variances and covariances by extension of corrections for ordinary GEE,\(^{26,27}\) may broaden the applicability of these methods. However, because the validity of estimating equation methods depends upon the assumption of asymptotic normality of parameter estimates, construction of (possibly asymmetric) confidence intervals based upon resampling strategies may be a better choice for small samples. In articul, bootstrap methods,\(^{28}\) though more computationally intensive than the GEE approach, are often easy to implement.\(^{29}\)

ACKNOWLEDGMENTS

We are grateful to 2 anonymous referees for providing helpful comments.

REFERENCES

The following arguments show that the GEE Wald χ² statistic, QW, in Expression 5 of Rochon,²¹ is equal to Z² thus demonstrating equivalency of the 2 power analysis approaches. This first step deduces the GEE estimator $\hat{\beta}$ in Equation 2 of Rochon for the simple logistic model in Equation 3, binomial variance function $\psi_h$, and “exchangeable” working correlation matrix $R = \phi(J/m) + (1 - \alpha_0) \times (I - J/m)$, where $I$ is the $m \times m$ identity matrix and $J$ is the $m \times m$ matrix of 1’s. Define 0 and 1 as $m \times 1$ vectors of 0’s and 1’s, respectively. Following Equation 2 of Rochon,²¹ the GEE estimator for $\beta = (\beta_0, \delta_0)$ is

$$
\hat{\beta} = \left[ \sum_{h=1}^{2} X'_{h} W_h X_{h} \right]^{-1} \left[ \sum_{h=1}^{2} X'_{h} W_h (1g_0) \right]
$$

where $X_1 = [1, 1]$ and $X_2 = [1, 0]$, $W_h = \Delta_h R^{-1} \Delta_h$, $\Delta_h = \text{diag}(\psi_{h1})$, and $g_0 = \text{logit}(\pi_{h})$ for $h = 1, 2$. Matrix computations and the result $R^{-1} = J/(m\phi) + (I - J/m)/(1 - \alpha_0)$ lead to $\hat{\beta} = (g_2, g_1 - g_2)$ and, from Equation 3 of Rochon,²¹ its model-based variance estimator is

$$
cov_{MB}(\hat{\beta}) = \frac{\phi}{nm} \begin{bmatrix}
1/\nu_{21} & -1/\nu_{21} \\
-1/\nu_{21} & 1/\nu_{21} + 1/\nu_{11}
\end{bmatrix}.
$$

Finally, the hypothesis $H_0 : \delta_0 = 0$ can be expressed as $H_0 : H\beta = 0$, where $H = [0, 1]$. The Wald chi-square test statistic is

$$
Q_w = (H\hat{\beta})'[H\text{cov}_{MB}H']^{-1}(H\hat{\beta}) = Z^2
$$

proving the equivalency.
Abstract: It is possible to classify the types of causal relationships that can give rise to effect modification on the risk difference scale by expressing the conditional causal risk-difference as a sum of products of stratum-specific risk differences and conditional probabilities. Directed acyclic graphs clarify the causal relationships necessary for a particular variable to serve as an effect modifier for the causal risk difference involving 2 other variables. The directed acyclic graph causal framework thereby gives rise to a 4-fold classification for effect modification: direct effect modification, indirect effect modification, effect modification by proxy and effect modification by a common cause. We briefly discuss the case of multiple effect modification relationships and multiple effect modifiers as well as measures of effect other than that of the causal risk difference.

(Epidemiology 2007;18: 561–568)

Directed acyclic graphs have been used as causal diagrams in epidemiologic research for a variety of purposes. Directed acyclic graphs have been used to represent causal relations among variables; they have been used extensively to determine the variables for which it is necessary to control for confounding to estimate causal effects; more recently, they have been used by Hernán et al to provide a classification of the types of causal relationships that can give rise to selection bias. In this paper, we follow the work of Hernán et al by using directed acyclic graphs to provide a classification of the types of causal relationships that can give rise to effect modification. Specifically, we consider the possible relationships between an effect modifier variable, the variable constituting the cause, and the variable constituting the effect. Doing so yields a structural classification of effect modification; the classification is structural in that it makes reference to the structure of the causal directed acyclic graph.

We first provide some discussion of the various measures of effect used to assess effect modification. We will then focus on the causal risk difference as a measure of effect by which effect modification is assessed (though much of the discussion applies also to other measures of effect as well) and we use directed acyclic graphs to provide a classification of different types of effect modification. Extensions to conditioning on multiple items and to scales other than the risk difference are discussed at the paper’s conclusion and in Appendix 1.

**EFFECT MODIFICATION**

Epidemiologists apply the term “effect modification” to indicate that the effect of one variable on another varies across strata of a third. There are many different measures of effect and, thus many different measures by which a variable may be an effect modifier for the relationship between a cause and an effect. There has been considerable discussion as to which measure of effect one might most naturally use in assessing effect modification. The risk difference, the risk ratio and the odds ratio are all frequently used in assessing effect modification. In general, different measures of effect will be useful in different contexts. For example, the risk difference which, within the context of effect modification, measures departures from the additivity of effects, is arguably of greatest public health importance, whereas the odds ratio is the natural measure of choice for case-control studies.

Here we will focus primarily on the causal risk difference as our measure of choice. By using the causal risk difference, the various relationships between an effect modifier, the variable constituting the cause, and the variable constituting the effect are made particularly clear. Although we focus on the causal risk difference, much of the discussion applies also to other measures of effect. It should also be noted that effect modification, whether on the causal risk difference scale or any other scale, falls under the broader idea of “interaction.” The relationship between effect modification and different notions of interaction in a counterfactual framework has been developed elsewhere. These issues are outside the scope of this paper. Much of this literature concerns individual-level interaction. Directed acyclic graphs allow for the graphical representation of population-level causal relationships and thus the causal risk difference (or, alternatively, causal risk ratio or odds ratio) provides the most appropriate focus for our analysis.
CAUSAL DIRECTED ACYCLIC GRAPHS

A directed acyclic graph is composed of variables (nodes) and arrows between nodes (directed edges) such that it is not possible to start at any node, follow the directed edges in the arrowhead direction and end up back at the same node. A causal directed acyclic graph is one in which the arrows can be interpreted as causal relationships and in which all common causes of any pair of variables on the graph are also included on the graph. Thus, if a graph has nodes $V_1, \ldots, V_n$ and some variable $U$ is a common cause of say $V_1$ and $V_2$, then $U$ must be included on the graph for the graph to be a causal directed acyclic graph. If, on the other hand, $U$ were only a cause of $V_1$ or only a cause of $V_2$ but not both, then $U$ could be included on the graph or it could be left out. Its inclusion would not be necessary for the graph to be a causal directed acyclic graph. That the graph’s edges are directed ensures that causes precede effects; that the graph is acyclic ensures that no variable can be its own cause. If there is a directed edge from $A$ to $B$ then $A$ is said to be a parent of $B$ and $B$ is said to be a child of $A$. If there are a series of one or more directed edges such that it is possible to begin at node $A$, follow the directed edges in the arrowhead direction, and end at another node $B$, then $A$ is said to be an ancestor of $B$ and $B$ is said to be a descendant of $A$. If $A$ is a parent of $B$, then $A$ is a direct cause of $B$; if $A$ is an ancestor but not a parent of $B$, then $A$ is said to be an indirect cause of $B$ (through the intermediate variables between $A$ and $B$). Additional details can be found in the work of Greenland et al.\(^1\) Greater formalization is provided by Pearl.\(^1\)

Statistical associations on causal directed acyclic graphs can arise in a number of ways. Two variables, $A$ and $B$, may be statistically associated if $A$ is either a direct or indirect cause of $B$, or if $B$ is a direct or indirect cause of $A$. Even if neither is the cause of the other, the variables $A$ and $B$ may still be statistically associated if they have some common cause $C$. Finally, the variables $A$ and $B$ may be statistically associated if they have a common effect $K$ and the association is computed within the strata of $K$; that is to say, $A$ and $B$ will in general be statistically associated given $K$, if $K$ is a common effect of $A$ and $B$. We will graphically represent conditioning by placing a box around the variable on the graph upon which we are conditioning.\(^3\)

More formally, the statistical association between variables can be determined by blocked and unblocked paths. A path is a sequence of nodes connected by edges regardless of arrowhead direction. A directed path is a path that follows the edges in the direction of the graph’s arrows. A collider is a particular node on a path such that both the preceding and subsequent nodes on the path have directed edges going into that node (ie, both the edge to and the edge from that node have arrowheads into the node). A path between $A$ and $B$ is said to be blocked given some set of variables $Z$ if either there is a variable in $Z$ on the path that is not a collider or if there is a collider on the path such that neither the collider itself nor any of its descendants are in $Z$. It has been shown that if all paths between $A$ and $B$ are blocked given $Z$, then $A$ and $B$ are conditionally independent given $Z$.\(^18-20\)

We need one further result in the development of the structural classification of effect modification below. The backdoor path adjustment theorem states that if for intervention variable $E$ and outcome $D$, if a set of variables $Z$ in which no variable in $Z$ is a descendent of $E$ blocks all “back-door paths” from $E$ to $D$ (ie, all paths with directed edges into $E$), then conditioning on $Z$ suffices to control for confounding for the estimation of the causal effect of $E$ on $D$; this causal effect is given by $\mathbb{E}(D_{e=z}) = \sum_z \mathbb{E}(D | E = e, Z = z)P(Z = z).$\(^1\) Note that this is a graphical variant of Theorem 4 of Greenland and Rubin\(^21\) and Robin’s g-formula.\(^22-23\)
FIGURE 1. Direct effect modification: E, drug exposure; D, hypertension outcome; X, genotype, a direct effect modifier.

FIGURE 2. Indirect effect modification: E, drug exposure; D, hypertension outcome; X, genotype; C, mother’s genotype, an indirect effect modifier.

FIGURE 3. Effect modification by proxy: E, drug exposure; D, hypertension outcome; X, genotype; R, hair color, an effect modifier by proxy.

FIGURE 4. Effect modification by common cause: E, drug exposure; D, hypertension outcome; X, genotype; C, mother’s genotype; M, mother’s hair color, an effect modifier by common cause.

genotype C of the mothers of the study participants. The mothers’ genotype affects the genotype of the study participants but does not affect hypertension of the study participants directly. The causal relationships among these variables is represented by the causal directed acyclic graph in Figure 2. We will show more formally below that in Figure 2, C will likely serve as an effect modifier on the causal risk difference scale for the effect of E on D. This is essentially because C affects X, which serves as a direct effect modifier for the causal effect of E on D. We could thus say that C is an indirect effect modifier for the causal effect of E on D, since C affects D indirectly through X. Now suppose that genotype X also determined hair color R. The causal relationships among these variables could be represented by the causal directed acyclic graph in Figure 3. Here, R will also likely serve as an effect modifier on the causal risk difference scale for the effect of E on D because conditioning on R gives information on X (which serves as a direct effect modifier for the causal effect of E on D). However, because R is not a cause of D, we would say that R is an effect modifier by proxy. Finally, suppose that information is available on the mothers’ hair color, which we will denote by M. The causal relationships among the variables could then be represented by the causal directed acyclic graph in Figure 4. It will be seen below that M also will likely serve as an effect modifier of the causal risk difference of E on D because conditioning on M gives information on C, which affects X, which serves as a direct effect modifier. Because C is a common cause of X (which is a direct cause of D) and M (the variable we are conditioning on), we might refer to M as an effect modifier by common cause of the effect of E on D.

We now generalize this simple example and show that all instances of effect modification can be classified as falling into one of the 4 categories indicated above: direct effect modification, indirect effect modification, effect modification by proxy and effect modification by common cause. The classification is carried out by expressing the conditional causal risk difference as a sum of products of stratum-specific risk differences and conditional probabilities as given in Theorem 1. In Theorem 1, we assume that there are no intermediate variables between E and D on the directed acyclic graph. In Appendix 1 the result is generalized and this assumption is dropped. Proofs of all theorems are given in Appendix 2.

**Theorem 1:** Suppose that E is a parent of D and that there are no intermediate variables between E and D. Let X denote the parents of D other than E. Let Q be some set of nondescendants of E and D then

$$E[D_{e_1} \mid Q = q] - E[D_{e_0} \mid Q = q]$$

$$= \sum_x (E[D \mid X = x, E = e_1] - E[D \mid X = x, E = e_0]) P(X = x \mid Q = q).$$

(1)

Theorem 1 states that the causal risk difference for D comparing 2 levels of E, e1 and e0, within a particular stratum of Q, is given by the sum of the expected risk differences in D conditional on X and Q weighted by the probability of X given Q where X denotes the parents of D other than E. Equation 1 allows us to provide a structural classification of effect modification on the causal risk-difference scale. For Q to be an effect modifier of the causal effect of E on D, it is
necessary that the function $G(q) = \mathbb{E}[D_{E=e_1} \mid Q = q] - \mathbb{E}[D_{E=e_0} \mid Q = q] = \sum_i \mathbb{E}[D \mid X = q, E = e_i] - \mathbb{E}[D \mid X = q, E = e_0] P(X=x \mid Q=q)$ is not constant in $q$. In other words, it is necessary that the expected risk difference in $D$ conditional on $X$ and $Q = q$ weighted by the probability of $X$ given $Q = q$ is not constant in $q$. This latter expression depends on $Q$ only through $P(X=x \mid Q=q)$ and so it is necessary that $P(X=x \mid Q=q)$ is not constant in $q$. The requirement that $P(X=x \mid Q=q)$ is not constant in $Q$ is simply the requirement that $X$ and $Q$ are statistically associated. In the introductory material on directed acyclic graphs, we discussed the various structures that can give rise to association between 2 variables: cause and effect, common causes, and conditioning on a common effect. We will therefore now use directed acyclic graphs to consider various cases for which a potential effect modifier $Q$ will be associated with one or more of the variables in $X$. This will allow us to classify the type of effect modification for any potential effect modifier $Q$ on the graph. Our analysis will follow the hypothetical example given above. First, the conditioning variable may be among the variables in $X$ (ie, it may be a parent of $D$). This gives rise to what we will call direct effect modification (Fig. 1, with $Q = X$). Second, the conditioning variable may be an ancestor of one or more of the variables in $X$, which gives rise to what we will call indirect effect modification by proxy (Fig. 2, with $Q = C$). Third, $Q$ may be a descendent of one or more of the variables in $X$, which gives rise to what we will call effect modification by proxy (Fig. 3, with $Q = R$). Finally, $Q$ and one or more of the variables in $X$ may have a common cause, which gives rise to what we will call effect modification by a common cause (Fig. 4, with $Q = M$). Theorem 1 allowed us to transform the condition for effect modification on the causal risk difference scale into the necessary condition for effect modification that $Q$ and $X$ are statistically associated. Our knowledge of association structures on causal directed acyclic graphs then allowed us to classify types of effect modification.

The 4 types of effect modification can be distinguished in a number of ways. First, as is clear from Figures 3 and 4, an effect modifier for the effect of some exposure on a particular outcome might not itself have a causal effect on that outcome. In the cases of direct and indirect effect modification, the effect modifier does have a causal effect on the outcome; in the cases of effect modification by proxy and by common cause, the effect modifier does not. This is because the unblocked path from $Q$ to $X$ (which gives rise to the required association between $Q$ and $X$) will be a frontdoor path from $Q$ to $X$ in the cases of direct or indirect effect modification, and a backdoor path from $Q$ to $X$ in the cases of effect modification by proxy or by common cause. Second, direct effect modification may be distinguished from the other 3 types in an important way. If one is conditioning on multiple variables that include all the direct effect modifiers $X$, then no other variable on the graph will continue to serve as an effect modifier for the causal effect of $E$ on $D$ while conditioning on $X$. This is essentially because $X$ blocks all paths from any other potential effect modifier $Q$ to $D$. In a sense, direct effect modifiers take precedence over all other types. The case of conditioning on multiple variables is considered further in the Discussion and in Appendix 1.

One additional comment is necessary. For the function $G(q) = \sum_i \mathbb{E}[D \mid X = q, E = e_i] - \mathbb{E}[D \mid X = q, E = e_0] P(X=x \mid Q=q)$ not to be constant in $q$, it is also necessary that the function $\mathbb{E}[D \mid X = q, E = e_1] - \mathbb{E}[D \mid X = q, E = e_0] P(X=x \mid Q=q)$ not be constant in $q$. That is to say it is necessary that $X$ be an effect modifier for the relationship between $E$ and $D$. This will often, but not always, be the case. In the context of binary $E$ and $X$, the expression $\mathbb{E}[D \mid X = x, E = e_1] - \mathbb{E}[D \mid X = x, E = e_0]$ will often not be constant in $X$ if $E$ and $X$ exhibit synergism. Exceptions to the condition that $\mathbb{E}[D \mid X = x, E = e_1] - \mathbb{E}[D \mid X = x, E = e_0]$ is not constant in $x$ will occur whenever all individual response types that Greenland and Poole classify as exhibiting “causal independence” are absent. Further discussion of some of these issues can be found elsewhere. Here it suffices to note that if none of the variables in $X$ serve as an effect modifier for the causal effect of $E$ on $D$, then no other variable on the graph will serve as an effect modifier because $\mathbb{E}[D \mid X = x, E = e_1] - \mathbb{E}[D \mid X = x, E = e_0]$ is constant in $x$. Theorem 1 thus allows us to classify instances of effect modification but not to identify effect modification. It is possible for $P(X=x \mid Q=q)$ not to be constant in $q$ and for $\mathbb{E}[D \mid X = x, E = e_1] - \mathbb{E}[D \mid X = x, E = e_0]$ not to be constant in $x$, but still to have $G(q) = \sum_i \mathbb{E}[D \mid X = q, E = e_i] - \mathbb{E}[D \mid X = q, E = e_0] P(X=x \mid Q=q)$ constant in $q$ (in which case $Q$ would not be an effect modifier for the causal risk difference). This possibility arises because the differences $\mathbb{E}[D \mid X = x, E = e_1] - \mathbb{E}[D \mid X = x, E = e_0]$ may cancel each other out perfectly. These are exceptional cases, however; in general, whenever $P(X=x \mid Q=q)$ is not constant in $q$, there will be effect modification on the risk difference scale. Theorem 1 made reference to a set of variables $X$ constituted by the parents of $D$ other than $E$. If $E$ is the only parent of $D$, then this set is empty. It easily follows that there can then be no effect modifier on the graph for the causal effect of $E$ on $D$. This is stated formally in Theorem 2.

Theorem 2. Suppose a node $D$ on a causal directed acyclic graph has only one parent, $E$; there then exists no variable on the directed acyclic graph that is an effect modifier for the causal effect of $E$ on $D$.

Essentially, Theorem 2 states that if the exposure $E$ is the only variable on the directed acyclic graph that is a direct cause of $D$, then there can be no variable on the directed acyclic graph that acts as an effect modifier for the relationship between $E$ and $D$. This is because any other variable that could have an effect on $D$ must do so through $E$, but intervening on $E$ will supersede any effect another variable might otherwise have had. Note that the theorem does not state the absence of any effect modifier for the causal effect of $E$ on $D$, but only the absence of an effect modifier on the directed acyclic graph. A causal directed acyclic graph that included all the variables on the original graph plus others might have an effect modifier if at least one of the additional variables were a direct cause of $D$. (Note that there may be causes of $D$ that were not on the original causal directed acyclic graph if these variables are not also causes of another
variable on the graph and thus not common causes of 2 or more variables on the graph). We see then that for there to be an effect modifier for the causal effect of $E$ on $D$ on a causal directed acyclic graph, the node $D$ must have more than one parent.

**DISCUSSION**

Regarding possible limitations and extensions of these results, it should first be acknowledged that, although effect modification is an important phenomenon, it is sometimes of insufficient magnitude to be of clinical or public health relevance. The theory developed in this paper does not allow for the representation of the magnitude of effect modification and thus cannot distinguish between cases in which the magnitude is or is not of substantive importance. Directed acyclic graphs are a useful conceptual tool, but cannot take the place of empirical analysis and presentation of numerical results.

Second, effect modification relationships may be considerably more complicated than the examples given in Figures 1 to 4. The causal diagram may, for example, involve a number of intermediate variables. Furthermore, a variable in $Q$ may be associated with more than one variable in $X$, giving rise to multiple effect modification relationships. Additionally, the set $Q$ itself may contain more than one variable, thereby giving rise to multiple effect modifiers. These and other more complex causal structures and effect modification relationships are discussed in Appendix 1.

Third, although the analysis has been restricted to the use of the causal risk difference as a measure by which to assess effect modification, most of the remarks hold true for other measures of effect. For example, the causal risk ratio for the causal effect of $E$ on $D$ comparing 2 levels of $E$, $e_0$ and $e_1$, in stratum $Q = q$ is given by $\frac{\sum [P(D = \text{success}|X = x, E = e_1)P(X = x|Q = q)]}{\sum [P(D = \text{success}|X = x, E = e_0)P(X = x|Q = q)]}$. If $X$ and $Q$ are not independent, then the expression $\frac{\sum [P(D = \text{success}|X = x, E = e_0, P(X = x|Q = q)]}{\sum [P(D = \text{success}|X = x, E = e_1, P(X = x|Q = q)]}$ will most cases vary in $Q$ although exceptions can occur (just as in the case of the causal risk difference above). Similarly, the causal odds ratio for the effect of $E$ on $D$ will generally vary in levels of $Q$ if $Q$ is not independent of $X$. The analysis, however, is most simple for the causal risk difference.

Fourth, for an exposure $E$, an outcome $D$, and a potential effect modifier $Q$, several different causal directed acyclic graphs of varying complexity may represent the causal relationships among these variables. One graph with $E, D$, and $Q$ may have additional variables; another may have only $E, D$ and $Q$. The one requirement that must be satisfied with regard to the presence or absence of other variables on the graph is that any common cause of 2 variables on the graph must also be on the graph. Thus, for example the variable $X$, in Figure 2, could have been excluded from the graph with an arrow from $C$ directly into $D$. Even though $X$ is a cause of $D$, it is not a cause of any other variable on the graph, and its inclusion is therefore optional. In our system of classification, excluding $X$ would make $C$ a direct modifier of the causal effect of $E$ on $D$. The classification of effect modifiers will thus sometimes be relative to the particular variables included on the graph under consideration. However, Figure 5 illustrates that indirect effect modification cannot always be reduced to direct effect modification by excluding variables from the graph. In Figure 5, $Q_1$ serves as an indirect effect modifier for the causal effect of $E$ on $D$. However, the variable $X_1$ cannot be removed from the graph because it is a common cause of $E$ and $D$. Similarly, in Figure 4, $M$ was an effect modifier by common cause which could have been reduced to an effect modifier by proxy if $X$ had been excluded from the graph with an arrow from $C$ directly into $D$. However, in Figure 5, $Q_2$ serves as an effect modifier by common cause through $X_3$, but cannot be reduced to an effect modifier by proxy because $X_2$ cannot be removed from the graph because it is a common cause of $E$ and $D$. How an effect modifier is classified can be, but is not always, relative to the particular variables represented on the graph under consideration.

The results developed in this paper have allowed us to classify any given instance of effect modification into one of 4 types: direct effect modification, indirect effect modification, effect modification by proxy or effect modification by common cause. The theory does not pertain to identifying effect modification, or to graphically representing interactions or effect modification; rather, merely to classifying instances of effect modification according to the structure of a directed acyclic graph. Theory concerning the graphical representation of synergistic interactions for binary variables is developed in related work. The results in this paper classify and clarify the necessary causal relationships an effect modifier variable must exhibit in relation to the variable constituting the cause, and the variable constituting the effect.

**REFERENCES**

3. Greenland S, Brumback B. An overview of relations among causal

---

**FIGURE 5.** Multiple effect modifiers: $E$, exposure; $D$, outcome; $X_1$, $X_2$, $X_3$, other direct causes of $D$; $Q_1$, an indirect effect modifier; $Q_2$, an effect modifier by common cause; $Q_3$, an effect modifier by proxy; $C$, a common cause.

APPENDICES

**Appendix 1: More Complex Effect Modification Structures**

As noted in the Discussion section, effect modification relationships may be considerably more complicated than the examples given in Figures 1 to 4. Any number of variables might lie between $Q$ and $X$ in Figure 2 or between $X$ and $Q$ in Figure 3 or between $C$ and $X$ or $C$ and $Q$ in Figure 4; the set $X$ may contain more than one variable; and, as discussed below, there may be intermediate variables between $E$ and $D$. However, when one considers the relation between the potential effect modifier $Q$ and some particular member of the set $X$ with which it is associated, their association will arise from one of the 4 alternatives presented above. Note however that these 4 alternatives are not mutually exclusive. A variable in $Q$ may be associated with more than one variable in $X$ and may exhibit any one of these 4 relations to (or be independent of) each particular variable in $X$. An example in which multiple effect modification relationships are present is given in Figure 6.

In the rather complicated example given in this figure, $Q$ is an effect modifier for the causal effect of $E$ on $D$ indirectly through $X_1$, by proxy through $X_2$, and by common cause through $X_3$. The four-way classification above is sufficient if attention is restricted to conditioning on a single item. But it is also possible that a researcher is interested in 2 variables considered jointly as an effect modifier for the causal effect of $E$ on $D$. One can then provide a classification of the effect modification structure for each variable in the set $Q$. An example of multiple variables serving as effect modifiers is given in Figure 5.

In Figure 5, the set $X$ (the parents of $D$ other than $E$) consists of the variables $X_1$, $X_2$, and $X_3$. The variable $Q_1$ serves as an indirect effect modifier through $X_1$; the variable $Q_2$ is an effect modifier by common cause through $X_2$; the variable $Q_3$ is an effect modifier by proxy through $X_3$. The variables $X_1$, $X_2$, and $X_3$, if they had been conditioned upon (instead of $Q_1$, $Q_2$, and $Q_3$) would all have served as direct effect modifiers of the causal effect of $E$ on $D$. Of course, more complicated arrangements are also possible in which each of the variables in the set $Q$ exhibits multiple effect modification relationships to the variables in $X$ and in which additional intermediate variables are present.

We have seen that $X$ and $Q$ may be associated either because of a cause and effect relationship ($Q$ directly causes $X$, $Q$ indirectly causes $X$, or $X$ causes $Q$) or through a common cause of $X$ and $Q$; this led to the fold-four classification of effect modification given above. However, as noted in the introductory material on directed acyclic graphs, $X$ and $Q$ may also be associated by conditioning on a common effect of these 2 groups of variables. This might occur in practice if,

![FIGURE 6. Multiple effect modification relationships: $E$, exposure; $D$, outcome; $X_1$, $X_2$, $X_3$, other direct causes of $D$; $Q$, an effect modifier; $C$, a common cause.](image-url)
for example, one were interested in 2 variables considered jointly as an effect modifier. Alternatively this might occur if one were interested in only one variable as an effect modifier but, due to the sampling procedure, one intentionally or inadvertently conditioned on a particular subset of subjects which restricted the sample to a particular stratum of one of the causal directed acyclic graph’s variables. In such cases, the common effect being conditioned upon, say $K$, might open a previously blocked path between $Q$ and some variable in $X$. We might then still apply the 4-fold classification given above with $K$ now taking on the role of $Q$ for purposes of classification. It turns out, however, that when $Q$ is an effect modifier for the causal effect of $E$ on $D$ because of conditioning on a common effect of $X$ and $Q$, this will rule out the cases of direct and indirect effect modification; effect modification will always be either by proxy or by common cause. If the effect modification were direct or indirect (rather than by proxy or common cause) then the unblocked path from $K$ to $X$ would be a frontdoor path and conditioning on $K$ would then block the path from $Q$ to $X$ (unless $Q$ were also associated with $X$ in ways other than paths through $K$). Thus the relationship between $K$ and $X$ must either be that of proxy or of common cause. The relationship between $Q$ and $K$ may either be that $Q$ is an ancestor of $K$ or that $Q$ and $K$ share a common cause. In summary, when conditioning on $K$, we must thus have either effect modification by proxy (conditioning on a common effect, with $Q$ and $K$ related by ancestry or by a common cause) or effect modification by common cause (conditioning on a common effect, with $Q$ and $K$ related by ancestry or by a common cause). Examples of each of these 4 cases are given in the directed acyclic graphs presented in Figures 7 to 10.

Conditioning on 2 or more items along with the effect modifier may complicate matters yet further but the same principles apply. It may be that $Q$ is associated with $X$ only by conditioning on several other variables. Classification may take place by considering the first conditioning variable on a particular unblocked path between $X$ and $Q$. Each consecutive pair of conditioning items or the final conditioning item and $Q$ may be related by ancestry or by common cause. As was the case in Figures 5 and 6, multiple effect modification relationships or multiple effect modifiers might also be present.

Our examples and the discussion thus far have assumed that there are no intermediate variables between $E$ and $D$ on the directed acyclic graph. The results and discussion, however, easily generalize to the setting in which there are intermediate variables between $E$ and $D$ on the directed acyclic graph. Theorem 3 restates Theorem 1 dropping the assumption of no intermediate variables between $E$ and $D$ on the directed acyclic graph. The conclusion of Theorem 2 is the same as that of Theorem 1 but the conditions under which this conclusion holds are slightly different.

**Theorem 3.** Let $D$ be some node on a causal directed acyclic graph with ancestor $E$ and let $X$ denote all nondescen-
The effect of rather all nondescendents of \( E \) on \( Q \) is available through \( X \), a direct cause of \( D \); \( K \), a conditioning variable; \( C_1, C_2 \), common causes; \( Q \), an effect modifier by common cause.

In the presence of intermediate variables between \( E \) and \( D \) on the directed acyclic graph, effect modification can be classified as before, according to their relationships with the set of variables \( X \). However, in the context of intermediate variables between \( E \) and \( D \) on the directed acyclic graph, the set \( X \) is no longer simply the parents of \( D \) other than \( E \) but rather all nondescendents of \( E \) which are either parents of \( D \) or parents of a node on a directed path between \( E \) and \( D \).

A final warning is necessary. Theorems 1 and 3 did not allow for \( Q \) to be a descendent of \( D \). Consider the example given in Figure 11 in which \( Q \) is a descendent of \( E \) or \( D \).

When \( Q \) is a descendent of \( D \), then the causal effect of \( E \) on \( D \) conditioning on \( Q \) is no longer given by Equation 1 because conditioning on \( Q \) provides information on \( D \) other than that which is available through \( X \) and \( E \). In such cases \( Q \) does not serve as a genuine effect modifier for the causal effect of \( E \) on \( D \) because it is a consequence of \( D \).

**Appendix 2: Proofs**

**Proof of Theorem 1**

Theorem 1 is a consequence of Theorem 3 below.

**Proof of Theorem 2**

Let \( Q \) be some nondescendent of \( E \) and \( D \) then

\[
E[D_{E=a}|Q = q] - E[D_{E=0}|Q = q] = E[D_{E=a}] - E[D_{E=0}]
\]

by Theorem 3 of Pearl\(^1\) since \((D \cup Q|E)_{G_E}\), where \( G_E \) denotes the graph obtained by deleting from the original directed acyclic graph all arrows pointing into \( E \). Furthermore, \( E[D_{E=a}] - E[D_{E=0}] = E[D = 1|E = 1] - E[D = 1|E = 0] \) by the back-door path adjustment theorem since there are no unblocked back-door paths from \( E \) to \( D \) as \( E \) is the only parent of \( D \). Thus

\[
E[D_{E=a}|Q = q] - E[D_{E=0}|Q = q] = E[D = 1|E = 1] - E[D = 1|E = 0] \text{ is independent of } q.
\]

**Proof of Theorem 3**

By the law of iterated expectations we have

\[
E[D_{E=a}|Q = q] - E[D_{E=0}|Q = q] = \sum_X E[D_{E=a}|X = x, Q = q] P(X = x|Q = q) - \sum_X E[D_{E=0}|X = x, Q = q] P(X = x|Q = q).
\]

We will show that this latter expression is equal to

\[
\sum_X E[D_{E=a}|X = x] P(X = x|Q = q) - \sum_X E[D_{E=0}|X = x] P(X = x|Q = q).
\]

By Theorem 3 of Pearl\(^1\) it suffices to show that \((D \cup Q|X, E)_{G_E}\), where \( G_E \) denotes the graph obtained by deleting from the original directed acyclic graph all arrows pointing into \( E \). Any front door path from \( D \) to \( Q \) in \( G_E \) will be blocked by a collider. Any backdoor path from \( D \) to \( Q \) in \( G_E \) will be blocked by \( X \).

We thus have that

\[
E[D_{E=a}|Q = q] - E[D_{E=0}|Q = q] = \sum_X E[D_{E=a}|X = x] P(X = x|Q = q) - \sum_X E[D_{E=0}|X = x] P(X = x|Q = q) = \sum_X E[D|X = x, E = e_1] P(X = x|Q = q) - \sum_X E[D|X = x, E = e_0] P(X = x|Q = q).
\]

Because \( X \) will block all backdoor paths from \( E \) to \( D \) we have by the backdoor path adjustment theorem

\[
\sum_X E[D|X = x, E = e_1] P(X = x|Q = q) - \sum_X E[D|X = x, E = e_0] P(X = x|Q = q) = \sum_X (E[D|X = x, E = e_1] - E[D|X = x, E = e_0]) P(X = x|Q = q).
\]
Can DAGs Clarify Effect Modification?

Clarice R. Weinberg

Abstract: The system proposed by VanderWeele and Robins for categorization of effect modifiers that are causal nodes in a directed acyclic graph (DAG) was not intended to empower DAGs to fully represent complex interactions among causes. However, once one has algebraically identified effect modifiers, the DAG implies a role for them. The limitations of epidemiologic definitions of “effect modification” are discussed, along with the implications of scale dependency for assessing interactions, where the scale can be either absolute risk, relative risk, or odds. My view is that probabilistic independence leads to the log-complement as a natural scale for interaction, but even that scale does not necessarily admit unambiguous inference. Any 2 direct causes of D are effect modifiers for each other on at least 2 scales, which can make a reasonable person question the utility of the concept. Still, etiologic models for joint effects are important, because most diseases arise through pathways involving multiple factors. I suggest an enhancement in construction of DAGs in epidemiology that includes arrow-on-arrow representations for effect modification. Examples are given, some of which depend on scale and some of which do not. An example illustrates possible biologic implications for such an effect modification DAG.

(Epidemiology 2007;18: 569–572)

The theory of directed acyclic graphs (DAGs), as extensively developed by Pearl in the setting of artificial intelligence and in the epidemiology setting in 1999, is producing growing pains for the field, even as it clarifies how we think about sampling biases and confounder adjustment in statistical models for causal relationships. I welcomed the paper by VanderWeele and Robins in the current issue, as a long-awaited and important step toward using DAGs to clarify the role of effect modifiers in causing disease. Most diseases are caused by multiple factors acting together and often through distinct pathways that can lead to a common final phenotype. Teasing apart the causal choreography will remain a prize worth the struggle; and the prize seems more attainable than ever, thanks to the rich array of molecular tools that are newly available to us.

VanderWeele and Robins propose a system for classifying effect modifiers in DAGs according to 4 categories: direct effect modifiers, indirect effect modifiers, effect modifiers by proxy, and effect modifiers by common cause. Their categorizations are intended to be DAG-specific, and not necessarily biologically meaningful: an indirect effect modifier might morph to become a “direct” effect modifier, simply by omitting an intermediate node from the DAG. While one might wish for a more biologic meaning, few cause—effect relationships in epidemiology or biology are ever “direct,” as one can typically think of additional proximal intermediates. For example, a genotype is not plausibly a “direct” cause of much (despite Fig. 1 in the paper by VanderWeele and Robins). A variant allele that influences risk by producing an aberrant protein product (as in sickle cell anemia) and one that influences the rate of transcription of some other gene can both be validly represented as “direct” in the nosology of VanderWeele and Robins, even though their effects are biologically indirect. As another example, these authors clarify that a factor categorized as an effect modifier by “common cause” can sometimes be transformed to be an effect modifier by proxy, simply by omitting an intermediate factor from the DAG—thereby transforming the shared indirect cause of D to a direct cause. Thus the 4-way categorization is telling us something about the somewhat arbitrary way we have drawn the DAG itself and may capture little about the nature of the causal factor or its role in causing disease.

Are these classifications helpful? Presumably, the direct and indirect categories of effect modification are the ones with potential implications for intervention, although VanderWeele and Robins do not comment on the utility of their classification scheme.

It has been frustrating to me that important kinds of causal relationships are not captured graphically by DAGs, so I was disappointed that the example DAGs given by VanderWeele and Robins provide no graphical representation to indicate that a factor is regarded as an effect modifier for some other risk factor. Consider their example of a genotype, X, that might influence response to treatment in a randomized clinical trial of an exposure E. Their Figure 1 shows the corresponding DAG. One might have hoped to see causal diagrams that were able to show more interaction than is suggested by the 2 separate, mutually aloof arrows—one from E to D and one from X to D. Perhaps one could instead show an arrow from X that ends at the E-to-D arrow itself, as in my Figure 1. (Refer to the figure legends for how I am defining “direct cause” mathematically.) Note that if both X and E are causes of D, and X is an effect modifier for E, then E is necessarily also an effect modifier for X.
Variations of this causal scenario can also be captured. Consider Figure 1 in VanderWeele and Robins and remove the arrow from X to D. X might influence either the uptake or the metabolism of E, or the biologic response to E, but have no effect on D by itself. Thus, it may only affect the arrow from E to D, ie, the causal process itself. We would not think of X as a “direct cause” of D (its “main effect” would be 0), and we could show this by omitting its direct arrow to D, although the standard DAG might still demand that arrow. In environmental health, some of the most plausible effect modifiers are those that influence absorption, specific metabolic detoxification pathways, immune responses, apoptosis, or DNA repair processes. Some effect-modifying cofactors may have little or no effect in the absence of exposure (although others may retain effects via unmeasured other exposures whose effects are also modified). This phenomenon could be shown graphically as in my Figure 2. Note that this effect modification is present regardless of the scale selected. As another example, the polio vaccine should have no effect on an individual’s risk of developing polio in the absence of exposure to the polio virus.

There are also scenarios where both the X-to-D arrow and the E-to-D arrow could be omitted, because neither produces D by itself. This possible category of “pure” effect modification could be represented by arrows that join together, as in my Figure 3. Retardation secondary to phenylketonuria is an example, because neither the genetic metabolic defect nor dietary phenylalanine produces retardation by itself. Note that this kind of scenario also reflects effect modification that is present regardless of the scale selected, and could also be seen as a 2-component, sufficient-cause scenario, as described by Rothman as a causal pie. More generally, however, I tend to see things in stochastic rather than deterministic terms, and would not expect causal processes that do not involve a highly penetrant mutation to be usefully represented by pies.

Another kind of scenario that may not be well captured by the usual DAGs is one in which biologic intermediates are included as “E” in the DAG and those intermediates can be phenotypically diverse in ways that have not been identified, but may to some extent depend on how they were caused. Thus, for example, if pesticide exposure in pregnancy causes gestational diabetes, a known risk factor for developing pre-eclampsia, the condition may sometimes retain a physiologic fingerprint that we have not characterized, but which reflects its causation in a way that has implications for sequelae, eg, the risk of pre-eclampsia. Suppose an exposure X is not causally related to D except through a path involving an intermediate E. Then the risk of D among those with E may still depend on X, as if E retains a memory of its parent, X. This is an interesting kind of effect modification, which could be represented by the graph of Figure 4. Notice that if we did not include the arrow-on-arrow, representing effect modification, we could mistakenly think that D does not depend on X once we have conditioned on E. This kind of DAG strongly suggests the existence of diverse subtypes that have been inappropriately lumped into a single intermediate phenotype, subtypes that themselves carry implications for risk. In the example, the phenotypic diversity may be subtle, or may be as simple as variation in severity of gestational diabetes.

Having proposed Figures 1 to 4, which seem to me to capture more about effect modification than do the usual DAGs, there may be some formal logic reason to avoid arrow-intersecting graphs; such graphs are probably not kosher within formal directed acyclic graph theory. It may also
of X values, x₁ and x₀, we have effect modification, ie., E conditional on X is a cause of E, ie, the distribution of E for some pair of E values, e₁ and e₀, and some pair of X values, x₁ and x₀, we have effect modification, ie., h(E|D| e₁, x₁) − h(E|D| e₀, x₀)

\[ h(E|D| e₁, x₁) − h(E|D| e₀, x₀) \]

be that, once one allows arrow-on-arrow effects and begins to think about representing overlapping sets of contributing causes, the resulting tangle of arrows begins to look too much like some kind of pasta, and the complexity may become daunting.

I consider the terminology itself to be a significant problem in thinking about models for joint effects: “effect modification” may be the most unfortunate jargon in epidemiology. The phrase strongly implies that a cofactor actually is acting to modify the causal effect we are studying, and investigators (and others) can be seduced by the words into presuming their finding has causal meaning. In practice, a finding of “interaction” or “effect modification” usually means little more than inequality of an estimated parameter across strata, which is a much more accurate but less sexy way to say the same thing.

VanderWeele and Robins step very carefully around the old issues related to the choice of scale one should use when defining and identifying effect modification. Readers who favor an additive model, because it is of more immediate public health relevance than the popular multiplicative model, will appreciate their choice of the risk-difference scale in which to assess effect modification. Although they develop their ideas in a context where effect modification is based on a risk-difference criterion, VanderWeele and Robins point out that one could alternatively use either the risk ratio (log risk) or the odds ratio (log odds) as the scale for defining effect modification. The richness of these choices for identifying “effect modifiers” becomes clear if one considers the fact that if E is a risk factor, then any second risk factor must be an “effect modifier” for the effect of E on at least 2 of those 3 scales. Basic algebra guarantees that effect modification is all around us, while all this plenty must also make some of us doubt its usefulness as a statistical and epidemiological construct.

Nonetheless, a scale I think should be more often considered because it may sometimes actually have causal meaning, but one not mentioned by VanderWeele and Robins, is the log-complement scale. This is the scale that attempts to capture probabilistic independence among causal processes. Toxicologists developed the concept of “simple independent action” for toxic effects of 2 different chemicals administered simultaneously. The idea, which has also been discussed in the context of epidemiology, is simple. Suppose that exposure A can cause D and exposure B can cause D by a truly independent pathway, while D can also occur in the absence of either A and B through an independent background cause. The paradigm would be 2 hunters who are independently aiming at the same duck. Of course, the duck could also drop dead from some unrelated and independent background cause, eg, lightning. If the 3 pathways are probabilistically independent, we have:

Pr[D | A, B] = Pr[avoid causal effect of background]Pr[avoid causal effect of A];
Pr[D | A, B] = Pr[avoid causal effect of background]Pr[avoid causal effect of B];
and Pr[D | A, B] = Pr[D | A, B] Pr[avoid causal effect of B].

Hence we have that:

Pr[D | A, B] = \frac{Pr[D | A, B] Pr[D | A, B]}{Pr[D | A, B]}.

Under probabilistically independent effects then, it follows that there is additivity on the log complement scale:

\[ \frac{Pr[D | A, B]}{Pr[D | A, B]} = 1; \]
that is,
\[
\ln(Pr[D | A, B]) - \ln(Pr[D | \bar{A}, B]) = [\ln(Pr[D | A, B]) - \ln(Pr[D | \bar{A}, B]) + \\
\ln(Pr[D | \bar{A}, B]) - \ln(Pr[D | A, B])]
\]
Thus there is additivity of the effect of A and the effect of B, on the log-complement scale. For a rare outcome, the analogous additive model will approximately hold on the absolute risk scale. However, when the outcome is not rare, additivity of risks is not equivalent to statistical independence (there being a non-negligible chance that both hunters will hit the duck), so is not equivalent to additivity on the log complement scale.

When we study more than one individual, variation across individuals in susceptibility (certain ducks may be easier for the independently shooting hunters to see) can cause violations of probabilistic independence, even if stochastic independence of causes holds for each individual at risk, so even this (to me) appealing formulation does not necessarily permit a biologically meaningful inference. Moreover, one must also think differently and in a more complicated way about protective factors, such as vaccines. Nonetheless, although the field may have grown weary of this debate, issues of scale will continue to matter when we try to draw useful inference from models for joint effects.

DAGs have met with a mixed reaction in epidemiology, with some of my colleagues recognizing their importance for analysis of etiologic factors, and others preferring to think about confounding in more classic terms. I have long been one of the boosters.

One practical challenge with DAGs in my experience, however, is the contentiousness of choosing a DAG. These choices can be especially problematic in reproductive epidemiology, where time-related factors become important. For example: Is long interpregnancy interval a cause of change of partner, or is change of partner a cause of long interpregnancy interval? But the hard thinking and very careful consideration of the etiologic context that are needed to decide which DAG is epidemiologically most plausible can be extremely useful, as we try to break through our academically enforced reluctance to think directly about causes.

A different, longstanding problem with DAGs has been that when synergistic effects may be involved in the etiology of the disease, DAGs are annoyingly noncommittal. I’m afraid the paper by VanderWeele and Robins, although providing useful categorizations for effect modifiers, has not rescued DAGs from this limitation. There may be more future in rethinking the basics of how we draw the DAGs.

ACKNOWLEDGMENTS
I thank Richard MacLehose and Anne Marie Jukic for their critical reading of the manuscript.

REFERENCES
Are Girls More Susceptible to the Effects of Prenatal Exposure to Tobacco Smoke on Asthma?

Jouni J. K. Jaakkola* and Mika Gissler†

Background: Prenatal exposure to tobacco smoke through mother’s smoking increases the risk of developing asthma later in life. A recent study suggested that this effect is present only in girls. We explored potential differences in susceptibility between boys and girls.

Methods: We followed all 58,841 Finnish singleton babies born in 1987 through 5 nationwide registries for 7 years and identified all cases of doctor-diagnosed asthma (ICD-9 code 493). The birth registry provided categorical information on the mother’s smoking during pregnancy: no smoking (reference), low exposure (<10 cigarettes per day), and high exposure (≥10 cigarettes per day).

Results: In girls the cumulative incidence of asthma was 0.0245 in the reference group, 0.0310 in the low maternal smoking group (risk difference = 0.0065; 95% CI = 0.0053–0.0076), and 0.0360 in the high maternal smoking group (0.0115; 0.0096–0.0133). The corresponding cumulative incidences for boys were 0.0405, 0.0501 (0.0096; 0.0089–0.0103), and 0.0522 (0.0117; 0.0091–0.0142). In logistic regression analysis adjusting for confounding, the combined effect of male sex and high maternal smoking (compared with female sex and no smoking) was 112% excess risk. This corresponded closely to what would be expected from the additive independent effects of male sex (67% excess risk) and high maternal smoking (44% excess risk).

Conclusions: Effects of maternal smoking during pregnancy on the risk of developing asthma are similar in boys and girls, with no interaction on an additive scale.

(Epidemiology 2007;18: 573–576)

Early-life exposure to environmental tobacco smoke is an important determinant of the development of asthma later in life.1 Recently, prenatal exposure to tobacco smoke (through mother’s smoking during pregnancy) has also been linked to the risk of asthma in childhood and early-life wheezing.2–13 Some studies even suggest that prenatal exposure is more harmful than postnatal exposure.10,14 A recent cohort study of 3915 Australian children reported that heavy maternal smoking was related to an increased risk of asthma in adolescence, although this effect was found only in girls.14 Other studies of the effects of prenatal exposure on asthma or wheezing have not provided information on sex differences.2–13

We explored potential differences in the susceptibility of boys and girls to the effects of maternal smoking on asthma, using data from a cohort of Finnish children born in 1987.

METHODS

Data Sources and Study Population

The source population comprised all children born in Finland in 1987 (n = 60,254). We focused on all 58,841 singleton births and followed them through 5 national administrative health registries for 7 years.15,16

Information on the child’s birth weight, gestational age, and maternal smoking habits during pregnancy were obtained from the Finnish Medical Birth Registry established in 1987 and run by the National Research and Development Center for Welfare and Health.

Exposure Assessment

Information on smoking during pregnancy is routinely collected by the medical staff at the delivery hospital and registered in a standardized form. The individual forms are transferred to the Finnish Birth Registry in Helsinki, where the data are entered into electronic files. Information on maternal smoking is categorical: none, less than 10 cigarettes per day, and more than 10 cigarettes per day. (These are the categories as given on the questionnaire; exactly 10 cigarettes was not included.)

The validity of smoking information was assessed as part of the Finnish Prenatal Environment and Health Project by comparing the registry-based information on maternal smoking during pregnancy with 2 other sources of information: medical records that record maternal smoking routinely, and questionnaire information collected after the delivery.17 The agreement was excellent in both comparisons (daily smoking during pregnancy yes/no) with a κ coefficient of 0.84 [95% confidence interval (CI) = 0.81–0.87] against questionnaire information and 0.89 (0.86–0.91) against medical records. There was no information available on paternal smoking during pregnancy or postnatal exposure to tobacco smoke.
Health Outcomes
The outcome of interest was asthma defined as at least 1 hospitalization due to asthma (ICD-9 code 493), at least 1 entitlement to free medication due to asthma, or at least 1 entitlement to special care support (which can be granted for families with a disabled child, or with a child who has a long-term illness needing continuous help or surveillance) due to asthma before the age of 7 years.

Covariates
The basic adjustment was made using the following core covariates: sex, birth order, maternal age, marital status, and index of socioeconomic status.

Statistical Methods
We compared the risks (cumulative incidence in 7 years) of asthma according to fetal exposure to tobacco smoke due to maternal smoking separately among boys and girls. Risk difference was used as the measure of absolute effect; risk ratio (RR) and odds ratio (OR) were measures of relative effect. We used logistic regression analysis to estimate adjusted odds ratios for the relations of interest. The adjustment was made using the covariates listed above.

Next, we calculated independent and joint effects of sex and fetal exposure to tobacco smoke products on an additive scale. We compared the risk of asthma in 6 exposure categories: 1) female sex and no smoking exposure (R00, reference category); 2) male sex and no smoking exposure (R10); 3) female sex and low smoking exposure (R01); 4) female sex and high smoking exposure (R02); 5) male sex and low smoking exposure (R11); and 6) male sex and high smoking exposure (R12). On an additive scale, the interaction (IA) for heavy smoking and sex was:

\[ IA = (R_{12} - R_{00}) - (R_{10} - R_{00}) - (R_{02} - R_{00}). \]

RESULTS

Study Population
The follow-up rate was 99.9%. The prevalence of smoking during pregnancy was 15.5%, which was similar among the mothers of boys and girls. Smoking during pregnancy was related to young age, not being married, and low education.

Effect of Fetal Tobacco Smoke Product Exposure in Boys and Girls
Table 1 presents the risk of developing asthma and the relations between maternal smoking during pregnancy separately in boys and girls. In boys the risk of asthma was 0.0405 among unexposed, 0.0501 when the mother smoked less than 10 cigarettes per day during pregnancy, and 0.0522 with over 10 cigarettes per day. The corresponding effect estimates expressed in risk differences were 0.0096 (95% CI = 0.0089–0.0103) and 0.0117 (0.0091–0.0142) and in odds ratios 1.25 (1.05–1.49) and 1.30 (1.04–1.64). The adjusted odds ratios were similar to the crude odds ratios. The cumulative incidence of asthma was lower in girls than in boys, with a risk of 0.0245 among unexposed girls, 0.0310 with prenatal exposure less than 10 cigarettes per day, and 0.0360 with more than 10 cigarettes per day. The effect estimates in risk differences were 0.0065 (0.0053–0.0076) and 0.0115 (0.0096–0.0133), slightly smaller than those among boys, but expressed in crude odds ratios slightly greater, 1.26

<table>
<thead>
<tr>
<th>Sex</th>
<th>Cigarettes/d</th>
<th>No. Cases</th>
<th>No. Children</th>
<th>Risk</th>
<th>Crude RR (95% CI)</th>
<th>Crude OR (95% CI)</th>
<th>Adjusted OR* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>0&lt;10</td>
<td>234</td>
<td>5729</td>
<td>0.0408</td>
<td>1.25 (1.09–1.43)</td>
<td>1.26 (1.09–1.45)</td>
<td>1.23 (1.07–1.42)</td>
</tr>
<tr>
<td>Girls</td>
<td>&gt;10</td>
<td>136</td>
<td>3055</td>
<td>0.0445</td>
<td>1.36 (1.15–1.62)</td>
<td>1.38 (1.15–1.65)</td>
<td>1.35 (1.13–1.62)</td>
</tr>
<tr>
<td>Boys</td>
<td>0&lt;10</td>
<td>86</td>
<td>2776</td>
<td>0.0310</td>
<td>1.26 (1.01–1.58)</td>
<td>1.26 (1.01–1.60)</td>
<td>1.22 (0.97–1.54)</td>
</tr>
<tr>
<td>&gt;10</td>
<td>52</td>
<td>1445</td>
<td>0.0360</td>
<td>1.47 (1.11–1.94)</td>
<td>1.49 (1.11–1.98)</td>
<td>1.43 (1.07–1.92)</td>
<td></td>
</tr>
</tbody>
</table>

*Logistic regression analysis: adjusted for birth order, maternal age, marital status, and index of socioeconomic status.

**Reference category**
Excluding 2209 children (with 2 cases) without information on maternal smoking (boys: 0/1141, girls 2/1068) + 4 cases of children with undetermined sex (for which one had no information on maternal smoking).
The independent effect of male sex on asthma risk was 0.0160 expressed in risk difference and 1.67 (1.51–1.85) expressed in adjusted odds ratio, which is approximately a 67% excess risk (Table 2). The independent effect of tobacco smoke exposure among girls was 0.0065 and 0.0115 expressed in risk differences and 1.22 (0.97–1.54) and 1.44 (1.08–1.92) expressed in adjusted odds ratios and representing an excess risk of approximately 22% and 44%. The difference between the joint effect of male sex and heavy maternal smoking and the sum of their separate effects was 0.0002 (0.027–0.0160–0.0115), ie, with no evidence of interaction.

**DISCUSSION**

In our large cohort study comprising approximately 409,500 person-years of follow-up, we addressed the question of whether girls through age 7 years are more susceptible to the effects of maternal smoking during pregnancy on the development of asthma. As has been shown before, boys had a higher risk of asthma than girls, and maternal smoking during pregnancy increased the risk. The magnitude of absolute effect with mothers’ smoking was similar among boys and girls. The combined effect of male sex and maternal smoking corresponded to what was expected from the additive independent effects of male sex and maternal smoking, with no evidence for interaction.

**Validity of Results**

The source population included all children registered at birth in Finland in 1987. The coverage of the Finnish Birth Registry is close to 100%. For the purposes of the study we focused on all singleton births. The registry-based follow-up of asthma was expected to identify almost all the diagnosed cases.8

The outcome was doctor-diagnosed asthma, based on the registries. This information is likely to be both complete and of high quality for the following reasons. The registry-based follow-up of asthma was expected to identify almost all the diagnosed cases of asthma through the use of registries recording subsidized drug and other treatment. The National Social Insurance Institute controls the quality of the diagnostic procedures.

The lack of information on smoking during pregnancy was a potential source of selection bias. The magnitude of bias even in the worst scenario would be small, because this information was missing for less than 4% of mothers. The possibility of information bias was minimized because information on smoking during pregnancy and other relevant information were collected before the onset of the outcome.

We were able to control for several potential confounders. However, the birth registry data did not include information on exposure to environmental tobacco smoke (ETS) during pregnancy, and the registry-based follow-up data included no information on family smoking habits after birth. Maternal smoking during pregnancy and postnatal exposure to ETS from maternal smoking are strongly correlated.


<table>
<thead>
<tr>
<th>Exposure Category</th>
<th>No. Cases</th>
<th>No. Children</th>
<th>Risk Difference*</th>
<th>Interaction</th>
<th>Crude RR (95% CI)</th>
<th>Crude OR† (95% CI)</th>
<th>Adjusted OR† (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Girl</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>23,403</td>
<td>574</td>
<td>0.0245</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>&lt;10</td>
<td>2776</td>
<td>86</td>
<td>+0.0065†</td>
<td></td>
<td>1.26 (1.01–1.58)</td>
<td>1.27 (1.01–1.60)</td>
<td>1.22 (0.97–1.54)</td>
</tr>
<tr>
<td>&gt;10</td>
<td>1445</td>
<td>52</td>
<td>+0.0115†</td>
<td></td>
<td>1.47 (1.11–1.94)</td>
<td>1.48 (1.11–1.98)</td>
<td>1.44 (1.08–1.92)</td>
</tr>
<tr>
<td><strong>Boy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>24,442</td>
<td>991</td>
<td>0.0405</td>
<td></td>
<td>1.65 (1.49–1.83)</td>
<td>1.68 (1.51–1.87)</td>
<td>1.67 (1.51–1.85)</td>
</tr>
<tr>
<td>&lt;10</td>
<td>2953</td>
<td>148</td>
<td>+0.0256†</td>
<td></td>
<td>2.04 (1.71–2.44)</td>
<td>2.10 (1.74–2.52)</td>
<td>2.02 (1.68–2.42)</td>
</tr>
<tr>
<td>&gt;10</td>
<td>1610</td>
<td>84</td>
<td>+0.0277††</td>
<td></td>
<td>2.13 (1.70–2.66)</td>
<td>2.19 (1.73–2.77)</td>
<td>2.12 (1.68–2.68)</td>
</tr>
</tbody>
</table>

*Risk in exposure category minus risk in reference category (R00).

†Logistic regression analysis: adjusted for birth order, maternal age, marital status, and index of socioeconomic status.

‡Crude RR = R10 – R00.

§Crude OR† = R10 / R00.

¶Crude RR = R10 – R00 – (R11 – R00).

**Crude RR = R10 – R00 – (R11 – R00) – (R12 – R00).

††Crude OR† = R10 – R00 – (R11 – R00) – (R12 – R00).

Excluding 2209 children (with 2 cases) without information on maternal smoking (boys: 0/1141, girls 2/1068) + 4 cases of children with undetermined sex (for which one had no information of maternal smoking).
Therefore, the effect estimates calculated for smoking during pregnancy include some of the potential effect of ETS during childhood. However, the distribution of these potential confounders is expected to be similar among boys and girls, and thus assessment of sex differences should be unconfounded.

**Previous Literature**

Several epidemiologic studies have assessed the relation between maternal smoking during pregnancy and the risk of asthma and wheezing in the child, but only one has explored whether the effect of prenatal tobacco product exposure differs in boys and girls. Alati and colleagues reported consistent sex differences in the relation between maternal heavy smoking and early-life exposure to tobacco smoke and the risk of asthma at the age of 14. Among the girls, the adjusted odds ratio of asthma was 1.98 (95% CI = 1.25–3.33) for maternal smoking at least 20 cigarettes per day, compared with 1.02 (0.62–1.70) in boys. The authors suggested that smoking exposure in utero may exert more serious damage to airway function in females, which could be explained by sex differences in lung size. There is evidence that maternal smoking during pregnancy influences lung function, but to our knowledge clear sex differences in lung function have not been shown. In addition, interplay between airway function and sex hormones at pubertal age was suggested to affect the risk of asthma.

Our results suggest that in the presence of maternal smoking during pregnancy, boys and girls have similarly increased risks of developing asthma during the first 7 years of life.

**REFERENCES**

Prenatal Ultrasound Scanning and the Risk of Schizophrenia and Other Psychoses

Karin Stålberg, * Bengt Haglund, *† Ove Axelson, * Sven Cnatingius, ‡ Christina M. Hultman, ‡§ and Helle Kieler* ‡

Background: Prenatal ultrasound exposure has been associated with increased prevalence of left-hand or mixed-hand preference, and has been suggested to affect the normal lateralization of the fetal brain. Atypical lateralization is more common in patients with schizophrenia. We evaluated possible associations of prenatal ultrasound with schizophrenia and other psychoses.

Methods: We identified a cohort of individuals born in Sweden 1973–1978. During this period, one Swedish hospital (Malmo University Hospital) performed prenatal ultrasound on a routine basis, and all individuals born at that hospital were considered exposed to ultrasound. Children born at hospitals where ultrasound was not used routinely or selectively were considered unexposed. We used Poisson regression analysis to estimate the effect of ultrasound exposure on the incidence of schizophrenia and other psychoses.

Results: In all, 370,945 individuals were included in the study, of whom 13,212 were exposed to ultrasound. The exposed group demonstrated a tendency toward a higher risk of schizophrenia (among men, crude incidence rate ratio = 1.58 [95% confidence interval = 0.99–2.51]; among women, 1.26 [0.62–2.55]). However, men and women born in several of the 7 tertiary level hospitals without ultrasound scanning also had higher risks of schizophrenia compared with those born in other hospitals. For other psychoses there were no differences between groups.

Conclusions: No clear associations between prenatal ultrasound exposure and schizophrenia or other psychoses were found. Other factors related to place of birth might have influenced the results.

(Epidemiology 2007;18: 577–582)
Register, the Hospital Discharge Register and the Cause of Death Register. Information about place of residence during follow-up time was gathered from the Register of Total Population, held by Statistics Sweden. Individual record linkage between the registers was possible through the unique personal identification number assigned to each Swedish resident.

**The Swedish Medical Birth Register**

The Swedish Medical Birth Register contains prospectively collected information on more than 99% of all births in Sweden since 1973.19 The register includes data on maternal demographics, reproductive history and complications during pregnancy, delivery and the neonatal period. The information is provided through antenatal, obstetrical and neonatal records, which are filled in by midwives and physicians.

**The Hospital Discharge Register**

The Swedish Hospital Discharge Register includes data on dates of each hospital admission, discharge, main discharge diagnosis and secondary diagnoses if any. The diagnoses are classified and recorded by the treating physician according to the International Classification of Diseases (ICD) at the time of discharge from hospital. The agreement of Swedish register diagnoses versus review of medical records is very high for psychotic disorders. Recorded diagnoses are forwarded by computer medium to the Hospital Discharge Register held by the National Board of Health and Welfare. The routines of recording and forwarding diagnoses are standardized across Sweden. The register provides nationwide coverage since 1987 for psychiatric diagnoses of inpatient care, and includes care in psychiatric as well as medical clinics.

**The Cause of Death Register**

The Cause of Death Register contains data on dates and causes of death for Swedish citizens since 1961. Coverage is more than 99.5% and data are updated yearly.

**Population Databases, Statistics Sweden**

Information about emigration dates and place of residence during follow-up were retrieved from population databases of the total population in Sweden. Information about all births from 1969 through 1972 was collected from Statistics Sweden’s Birth Register.

**Study Sample**

We included singletons born at hospitals with reliable information on ultrasound scanning. Of 97 hospitals, 49 had reliable information on their ultrasound program regarding the current time period (1973 to 1978). To increase the homogeneity of the study population, we included only children of mothers who themselves had been born in one of the Nordic countries (Sweden, Denmark, Norway, Finland, and Iceland). All included children were alive and living in Sweden at the age of twelve (data from the Cause of Death Register). Of 593,917 singleton live births in Sweden between 1973 and 1978, we could include 370,945 individuals (62.5%) with reliable exposure information, of whom 190,405 were men and 180,540 were women.

Diagnoses of schizophrenia and other psychoses were obtained from the Hospital Discharge Register, and all individuals eligible for inclusion were followed up after reaching 12 years of age. The follow-up period was 1987 to 2004. A schizophrenia diagnosis included the following ICD codes: ICD-9 codes 295A-295E, 295G, 295W, 295X and ICD-10 code F20. Other psychotic disorders included ICD-9 codes 295F, 296–298 and ICD-10 codes F21–31.

**Exposed Group**

The University Hospital in Malmö was the first hospital in Sweden to introduce ultrasound scanning as part of standard antenatal care. Since October 1972, 90% of all pregnant women living in the city of Malmö have been scanned. The use of ultrasound from this period is well documented. Consequently, children born in Malmö to mothers registered as citizens of Malmö were considered exposed to ultrasound. During the introduction period, 1973–1975, the examinations were performed at approximately 28 weeks of gestation (83% had their routine scan before gestational week 30). Around 50% had additional scans, mostly to confirm measurements (because of lack of experience during this introduction period). The risk pregnancies having additional scans were mainly performed to assess intrauterine growth on request from clinicians. These repeated scans were generally performed with very short intervals.

From October 1976 through December 1978, 2 examinations were performed (full-scale period). For most mothers (95%) the first scan was in gestational week 18 to 20 and the second in week 32. There were no additional scans performed between these examinations, except for patients with bleeding from a suspected placenta previa. Of the more than 5000 women scanned during this period, there were 17 with a clinical diagnosis of placenta previa. The ultrasound machines used were the Kretz-Technic 4100 MGS and Combison II echoscope, Philips Diagnost B, and ADR-Kranzbücher. Each examination was scheduled for 15 minutes.24

**Unexposed Group**

Before 1980, 48 hospitals did not practice ultrasound scanning. We considered children born at any of these hospitals to be unexposed to ultrasound unless their mothers were registered as residents of Malmö, in which case they were excluded from the study.

**Statistical Analysis**

We used Poisson regression analysis to estimate the effect of ultrasound exposure on the incidence of schizophrenia and other psychotic disorders. The results are presented as incidence rate ratios (IRRs) with 95% confidence intervals (CIs). The time at risk was calculated from the 12th birthday, but no earlier than 1 January 1987, to the first date of event, death, emigration or end of follow-up (31 December 2004). The incidence was calculated as cases per 100,000 person years. The following variables from the Medical Birth Register were included as potential confounders: maternal age, parity, gestational age, intrauterine growth (calculated as the ratio between birth weight and expected birth weight for gestational age) and Apgar score. In addition, by retrieving
information from the Hospital Discharge Register, we were able to adjust for the mother’s number of hospitalizations at a psychiatric clinic. We also controlled for subjects’ attained age during follow-up. Preliminary analyses indicated that gestational age, intrauterine growth, and Apgar score had practically no impact on the estimates of interest. Hence, to avoid problems with missing values, we excluded these covariates from analyses.

To evaluate whether unmeasured factors related to hospital level might affect the results, we performed analyses in which hospitals of tertiary level were compared with all other hospitals. Apart from Malmö there were 7 hospitals of tertiary level in Sweden during the study period. We adjusted for county of residence during the follow-up period to control for geographical variation in diagnostic and admittance routines.

Because other factors related to place of birth might affect the risk of being diagnosed with schizophrenia, we compared the incidence of schizophrenia according to place of birth before and after the introduction of ultrasound scanning in Malmö. It was not possible to obtain information on hospital of birth before establishment of the Swedish Medical Birth Register in 1973. However, for individuals born in the cities of Malmö, Umeå, Uppsala, and Örebro between 1967 and 1978, hospital of birth could be inferred from the fact that each of these 4 cities only had one hospital. Consequently, we compared incidence of schizophrenia for individuals born in each of these 4 cities with those born in other parts of Sweden. For these analyses, we included 362,315 singletons born 1967 to 1972 (before the start of ultrasound scanning) and 315,989 singletons born 1973 to 1978 (after introduction of ultrasound scanning). Only mothers living in communities that fulfilled the criteria for the main study were included.

To assess the impact of number and frequency of ultrasound examinations on the risk for schizophrenia, we divided the exposed cohort into 2 subcohorts born 1973 to 1975 and 1976 to 1978. These 2 time periods coincide approximately with the introduction phase and full-scale phase of the ultrasound scanning program in Malmö.

The male fetal brain is considered more vulnerable than the female brain and higher risks of mixed-handedness/left-handedness in connection with prenatal ultrasound have been found only in men. Hence, we included sex as an effect modification term in all analyses and we present the results according to sex. Statistical analyses were conducted with the SAS 9.1 software package (SAS Institute, Cary, NC). Ethics Committees at Uppsala University and at Karolinska Institutet approved the study.

RESULTS

In Table 1, maternal and infant characteristics and associated incidence rates for schizophrenia and other psychotic disorders are presented stratified by ultrasound exposure (ie, exposed vs. nonexposed). We found a higher incidence of schizophrenia among individuals born at Malmö University Hospital (exposed to ultrasound) when compared with individuals born at all other hospitals (unexposed). The difference appears to be more pronounced among men (17.2 versus 10.9/100,000 person-years) than women (7.5 versus 6.0/100,000 person-years). In almost all categories of variables included in Table 1, the incidence of schizophrenia was higher in the exposed compared with the unexposed cohort. The same pattern could not be seen for other psychoses.

Among exposed as well as nonexposed children, male sex, high maternal age, preterm births, and mother’s psychiatric care were associated with high incidence rates of schizophrenia and other psychotic disorders. The estimated crude IRR for schizophrenia when exposed to ultrasound was 1.58 (95% CI = 0.99–2.51) for men and 1.26 (0.62–2.55) for women. The same estimates for other psychiatric disorders were 1.12 (0.80–1.58) for men and 0.92 (0.62–1.37) for women.

We continued with detailed analyses of ultrasound exposure and risks of schizophrenia and other psychoses. However, because there was no association between ultrasound exposure and other psychoses, we limit the presentation to ultrasound exposure and risk of schizophrenia. Because we found practically no differences in risks for schizophrenia between the subcohorts born during the introduction and the full-scale period, we present data for the whole cohort.

To control for possible confounders we performed multivariate analyses adjusting for maternal age, parity and maternal psychiatric care and subjects attained age during follow-up (Table 2). In Model I, individuals born at Malmö University Hospital were compared with those born at all other included hospitals. Among men, we found higher risks of schizophrenia if born in Malmö. For women, the point estimate was slightly increased but the confidence interval did not support an association. In Model II, individuals born at Malmö University Hospital and the 7 other hospitals of tertiary level were compared with those born at primary and secondary hospitals. Risks of schizophrenia remained essentially unchanged for both men and women born in Malmö. The highest risks of schizophrenia were found among men born at the university hospitals in Malmö and Uppsala. For women, the highest risk of schizophrenia was found for those born in Uppsala. Similarly, when adjusted for county of residence during follow-up (Model III), the highest risk of schizophrenia was found among men born in Malmö (IRR = 1.60; CI = 0.90–2.83) and women born in Uppsala (1.66; 0.89–3.10).

Finally, we compared risk of schizophrenia for singletons born in 4 cities between 1967 and 1972, before the introduction of ultrasound scanning, with risk of schizophrenia for those born between 1973 and 1978. Men born in Malmö had the highest ratio of IRRs between cohorts born before and after the introduction of ultrasound scanning (Table 3).

DISCUSSION

Individuals born at the university hospital in Malmö and assumed to have been exposed to prenatal ultrasound had higher incidence rates of schizophrenia than those born in other hospitals without ultrasound. However, when we extended the analyses and compared individuals born at other hospitals of tertiary level with those born at primary and
secondary hospitals, some of the other tertiary level hospitals also had high incidence rates of schizophrenia. Accordingly, factors other than ultrasound (factors that are unknown to us and seem to be related to place of birth) probably affect the risk of being diagnosed with schizophrenia. For other psychoses, the differences between ultrasound exposure and nonexposure were smaller and more ambiguous, which is in agreement with the opinion that nonschizophrenic psychoses are less associated with preand perinatal characteristics than schizophrenia.26

The women in our study had lower incidence rates of psychotic diseases than the men, which was most obvious for schizophrenia. The lifetime risk of schizophrenia is thought to be equal in men and women, although onset generally occurs about 4 years later in women (mean 26.5 years in men and 30.6 years in women).27 In the present study we were able to follow all individuals until 2004, which limited the maximum age to 31 years. This truncated follow-up time may have influenced the risk estimates, especially among women, and may also explain our lower incidence rates of schizophrenia among women. Thus, our conclusions should be restricted to subjects with a relatively early age of onset of schizophrenia.

Because the male fetal brain is considered more vulnerable than the female brain,16 and since higher risks of left-hand or mixed-hand preference in connection with prenatal ultrasound have been found only in men, we analyzed data according to sex. For most of the hospitals there was a reasonable agreement in risks of schizophrenia between men and women, ie, both men and women at a specific tertiary

| TABLE 1. Maternal and Infant Characteristics and Incidence (Cases per 100,000 Person-Years) of Schizophrenia and Other Psychoses According to Ultrasound Exposure |
|-----------------------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                      | Schizophrenia |                  | Other Psychoses |                  |
|                                      | Exposed*      | Unexposed*      | Exposed*        | Unexposed*      |
|                                      | No. Cases     | Incidence       | No. Cases       | Incidence       |
| Sex                                  |               |                 |                 |                 |
| Male                                 | 19            | 17.2            | 329             | 10.9            | 35              | 31.7            | 850             | 28.1            |
| Female                               | 8             | 7.5             | 170             | 6.0             | 25              | 23.5            | 726             | 25.5            |
| Maternal age (yrs)                   |               |                 |                 |                 |
| 10–19                                | 4             | 25.8            | 46              | 11.4            | 3               | 19.3            | 122             | 30.3            |
| 20–34                                | 20            | 10.5            | 411             | 8.0             | 54              | 28.3            | 1,319           | 25.8            |
| 35–54                                | 3             | 27.7            | 42              | 12.0            | 3               | 27.7            | 135             | 38.7            |
| Birth order                          |               |                 |                 |                 |
| 1                                     | 19            | 16.5            | 224             | 8.6             | 32              | 27.9            | 713             | 27.5            |
| 2 or more                            | 8             | 7.8             | 275             | 8.4             | 28              | 27.4            | 863             | 26.4            |
| Gestational age at birth (wks)        |               |                 |                 |                 |
| 23–36                                | 4             | 31.1            | 27              | 10.1            | 7               | 54.6            | 98              | 36.8            |
| 37–41                                | 17            | 9.9             | 402             | 8.5             | 41              | 24.0            | 1,239           | 26.2            |
| 42–45                                | 6             | 18.0            | 65              | 7.7             | 12              | 36.0            | 227             | 26.8            |
| Missing                               | 0             | 0.0             | 5               | 17.4            | 0               | 0.0             | 12              | 42.0            |
| Birth weight ratio†                   |               |                 |                 |                 |
| <0.75                                 | 1             | 13.4            | 10              | 6.6             | 2               | 26.8            | 51              | 33.7            |
| 0.75–0.90                            | 7             | 15.0            | 96              | 8.7             | 19              | 40.8            | 316             | 28.8            |
| 0.90–1.10                            | 15            | 12.4            | 271             | 8.2             | 26              | 21.4            | 851             | 25.6            |
| 1.10+                                | 4             | 9.7             | 117             | 9.2             | 13              | 31.6            | 346             | 27.3            |
| Missing                               | 0             | 0.0             | 5               | 16.2            | 0               | 0.0             | 12              | 38.9            |
| Apgar score at 5 min                  |               |                 |                 |                 |
| 0–6                                   | 0             | 0.0             | 6               | 10.2            | 0               | 0.0             | 22              | 37.6            |
| 7–10                                 | 25            | 12.3            | 487             | 8.5             | 56              | 27.5            | 1,523           | 26.6            |
| Missing                               | 2             | 33.4            | 6               | 7.6             | 4               | 0.0             | 31              | 39.2            |
| Mother’s inpatient psychiatric care   |               |                 |                 |                 |
| No                                    | 20            | 9.9             | 404             | 7.4             | 48              | 23.8            | 1,277           | 23.3            |
| Yes                                   | 7             | 44.2            | 95              | 24.1            | 12              | 75.8            | 299             | 76.1            |
| Total                                 | 27            | 12.4            | 499             | 8.5             | 60              | 27.6            | 1,576           | 26.9            |

*Children born at the University hospital in Malmö, which had ultrasound scanning as part of standard antenatal care.
†The ratio between actual and expected birth weight with regard to sex and gestational age.
port an association with ultrasound. Similarly, the association with schizophrenia reported here, was not supported in the subanalyses on hospital level. Nevertheless, men born in Malmö had the highest risks of schizophrenia and poor intellectual performance. Though these may be unrelated to ultrasound exposure or result of chance, it is somewhat alarming that men born in Malmö and assumed to be exposed to ultrasound deviate in all 3 outcomes that we have assessed.

We adjusted for potential confounding factors, such as maternal age, parity and mother’s number of hospitalizations for psychiatric care. Immigration, especially by persons from economically and genetically distinct populations, has been shown to increase risk of developing schizophrenia. In the present study we included individuals with Swedish mothers or with mothers that had immigrated to Sweden from the other Nordic countries, as the Nordic population is genetically and socioeconomically similar. Socioeconomic factors might affect risks of developing psychoses, yet in our previous studies we saw only minor differences in socioeconomic status between families living in Malmö compared with the rest of Sweden. We were unable to control for maternal smoking, as information on maternal smoking was not included in the Medical Birth Register for the birth cohorts included in the present investigation. Maternal smoking is causally related to fetal growth, and pregnant smokers may therefore be more likely to be exposed to repeated ultrasound scans during pregnancy. We suggest that in future studies maternal smoking should be included as a confounding factor in the analyses.

We were limited to include only patients with inpatient psychiatric treatment, as the information on outpatient treatment in the Hospital Discharge Register is incomplete. Routines for admitting psychiatric patients might vary among

level hospital had either increased or decreased risks. Although women born in Malmö had increased risks of schizophrenia, women born at 3 other tertiary level hospitals had even higher risks. Accordingly, prenatal ultrasound does not seem to be associated with increased risks of schizophrenia in women.

The lowest risks of schizophrenia were found in individuals born in Umeå. As we have adjusted for county of residence during the follow-up period, the differences in risks between the tertiary hospitals should not reflect geographical variation in diagnostic and admittance routines. We believe that a certain degree of variation in incidence of schizophrenia between the hospitals is to be expected and that the variations found in this study may reflect differences in genetic, socio-demographic and other factors.

For the same cohort, we have previously reported an increased risk of left-handedness among men born in Malmö and exposed to ultrasound, when compared with unexposed. In that study all the subanalyses on hospital level supported the association. However, in a previous study on intellectual performance, the subanalyses on hospital level did not sup-

### TABLE 2. Adjusted Incidence Rate Ratio for Schizophrenia According to Ultrasound Exposure by Sex and Hospital of Birth (n = 357,733)

<table>
<thead>
<tr>
<th>Schizophrenia</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IRR (95% CI)</td>
<td>IRR (95% CI)</td>
</tr>
<tr>
<td>Model I*†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malmö‡</td>
<td>1.55 (0.97–2.46)</td>
<td>1.26 (0.62–2.55)</td>
</tr>
<tr>
<td>Danderyd</td>
<td>1.33 (0.84–2.10)</td>
<td>1.65 (0.91–3.00)</td>
</tr>
<tr>
<td>Karolinska</td>
<td>1.11 (0.62–1.98)</td>
<td>0.97 (0.40–2.38)</td>
</tr>
<tr>
<td>Örebro</td>
<td>0.87 (0.46–1.63)</td>
<td>0.89 (0.36–2.17)</td>
</tr>
<tr>
<td>Sahlgrenska</td>
<td>0.78 (0.40–1.51)</td>
<td>1.06 (0.47–2.42)</td>
</tr>
<tr>
<td>Södersjukhuset</td>
<td>0.84 (0.46–1.53)</td>
<td>1.60 (0.84–3.06)</td>
</tr>
<tr>
<td>Uppsala</td>
<td>1.60 (1.06–2.41)</td>
<td>1.84 (1.05–3.20)</td>
</tr>
<tr>
<td>Umeå</td>
<td>0.36 (0.12–1.14)</td>
<td>0.50 (0.12–2.00)</td>
</tr>
<tr>
<td>Model II‡¶</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malmö‡</td>
<td>1.60 (0.90–2.83)</td>
<td>1.39 (0.63–3.05)</td>
</tr>
<tr>
<td>Danderyd</td>
<td>1.25 (0.77–2.05)</td>
<td>1.56 (0.84–2.90)</td>
</tr>
<tr>
<td>Karolinska</td>
<td>1.04 (0.57–1.91)</td>
<td>0.92 (0.37–2.28)</td>
</tr>
<tr>
<td>Örebro</td>
<td>0.52 (0.25–1.10)</td>
<td>0.55 (0.21–1.44)</td>
</tr>
<tr>
<td>Sahlgrenska</td>
<td>0.83 (0.42–1.66)</td>
<td>1.14 (0.49–2.63)</td>
</tr>
<tr>
<td>Södersjukhuset</td>
<td>0.79 (0.42–1.48)</td>
<td>1.51 (0.77–2.95)</td>
</tr>
<tr>
<td>Uppsala</td>
<td>1.43 (0.86–2.39)</td>
<td>1.66 (0.89–3.10)</td>
</tr>
<tr>
<td>Umeå</td>
<td>0.39 (0.12–1.32)</td>
<td>0.53 (0.12–2.27)</td>
</tr>
</tbody>
</table>

* Malmö in relation to all other hospitals.
† Adjusted for maternal age at birth, parity, mother’s hospitalization at a psychiatric clinic and subjects’ attained age during follow-up.
‡ Children born at the University hospital in Malmö, which had ultrasound scanning as part of standard antenatal care.
¶ Malmö and other hospitals of tertiary level in relation to all other hospitals.
* Adjusted for factors in Model I and II and county of residence during the follow-up period.

### TABLE 3. Adjusted Incidence Rate Ratio for Schizophrenia for 4 Swedish Municipalities of Birth, Before and After Start of Ultrasound Scanning*

<table>
<thead>
<tr>
<th>Birth Cohort</th>
<th>Birth Cohort</th>
<th>Cohort Ratios†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malmö‡</td>
<td>1.05 (0.70–1.58)</td>
<td>1.33 (0.80–2.21)</td>
</tr>
<tr>
<td>Umeå</td>
<td>0.98 (0.49–1.97)</td>
<td>0.35 (0.08–1.42)</td>
</tr>
<tr>
<td>Uppsala</td>
<td>1.59 (1.08–2.34)</td>
<td>1.22 (0.71–2.12)</td>
</tr>
<tr>
<td>Örebro</td>
<td>0.94 (0.54–1.64)</td>
<td>1.05 (0.52–2.12)</td>
</tr>
<tr>
<td>Women</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malmö‡</td>
<td>1.61 (1.02–2.53)</td>
<td>1.21 (0.58–2.54)</td>
</tr>
<tr>
<td>Umeå</td>
<td>1.47 (0.67–3.20)</td>
<td>0.37 (0.05–2.70)</td>
</tr>
<tr>
<td>Uppsala</td>
<td>1.59 (0.95–2.67)</td>
<td>1.76 (0.90–3.42)</td>
</tr>
<tr>
<td>Örebro</td>
<td>0.85 (0.39–1.86)</td>
<td>1.00 (0.36–2.77)</td>
</tr>
</tbody>
</table>

* Reference group was all municipalities in the rest of Sweden. Adjusted for subjects’ attained age during follow-up, maternal age, parity, mother’s numbers of hospitalizations for psychiatric care and city of residence during the follow-up period.
† Calculated as ratio of column 2:1.
‡ Children born at the University hospital in Malmö, which had ultrasound scanning as part of standard antenatal care 1973 and onwards.
regions of Sweden. To minimize the effects of restricting inclusions to inpatients, we adjusted for county of residence during the follow-up period. The adjustments for living place had minor effects on the results. The psychiatric diagnoses in the Hospital Discharge Register are reliable as Swedish diagnostic practice is generally considered to be of high quality. Risk of misclassification in diagnosed cases is probably small, as the diagnosis of schizophrenia is most certainly used with caution.

Strengths of the present study include the population-based design, the large size, and the prospective collection of data on birth characteristics, which precludes recall bias. We considered children of mothers from Malmö and born in Malmö as exposed to ultrasound. To avoid risk of misclassification of exposure we included only hospitals with reliable documentation on introduction of ultrasound programs. More than 90% of women attending antenatal care in Malmö had prenatal ultrasound scans from 1973 and onwards. However, some of the women in Malmö might not have had ultrasound, and some of those considered as unexposed may have had an ultrasound examination done in another part of Sweden. This possible error might lead to an underestimation of the association and reduce the possibility of finding an existing relation between exposure and outcome. The ultrasound examinations were performed as part of a routine screening program and were not performed on clinical indication, which minimizes the risk of exposure bias.

Although we had sufficient power to detect even small increases in risk in the main analyses, the strata included in subanalyses were too small to demonstrate small changes in schizophrenia incidence when comparing each tertiary hospital with other hospitals.

The present findings do not provide clear evidence of an association between prenatal ultrasound exposure and schizophrenia. Men born in Malmö had the highest risks compared with other men, whereas women born in Malmö were similar to women born at several other tertiary level hospitals. It seems that factors related to place of birth, rather than ultrasound exposure, may have influenced the results for both men and women. To overcome obstacles regarding uncertainty of exposure, adjustments for confounders and case ascertainment, we suggest that a case-control study should be performed using all available information from medical records.

REFERENCES
Belated Concerns and Latent Effects

The Example of Schizophrenia

Michaeline Bresnahan*† and Ezra Susser*†‡

Abstract: In this issue of Epidemiology, Stalberg and colleagues report the lack of an association between prenatal ultrasound and risk of schizophrenia in adulthood. These findings contribute to the search for unintended effects of an intervention designed to improve prenatal care. Although no adverse effect of ultrasound was reported, other gestational exposures have been associated with increased risk of schizophrenia. By extending the causal time frame to include long-term latent effects we are confronted with a potential paradox: interventions beneficial in early life may have undetected adverse consequences in adulthood.

(Epidemiology 2007;18: 583–584)

Since the alarm was first sounded for the teratogenic consequences of diethylstilbestrol (DES), the potential for latent effects of prenatal exposure on adult health has been well recognized. DES became the archetypal cautionary tale. The adverse effects of a drug taken by an expectant mother might not be manifest in offspring during infancy or even childhood, but could appear in adulthood. Later, investigators became concerned that affected offspring might go on to have children with related health problems even though these children were never exposed to the drug. DES underscored the need for a revised time frame for “do no harm.” The article in this issue examining the association of ultrasound exposure in gestation to risk of schizophrenia in adulthood reflects these concerns.

The study of prenatal ultrasound and schizophrenia also falls squarely within the approach of lifelong epidemiology. The importance of early life exposures and the lengthened causal time frame are fundamental to lifelong epidemiology, and have been an important focus in schizophrenia research. In our own work, prenatal nutritional, infectious, and chemical exposures have been tied to schizophrenia in adulthood. In other work, a variety of prenatal and obstetric complications have been associated with later risk of schizophrenia (reviewed by Cannon et al.). In several areas the findings are strong but not definitive; among the best established is nutritional deficiency. Maternal starvation during pregnancy was first identified as posing increased risk of schizophrenia in the context of the Dutch Hunger Winter, a famine induced by a Nazi blockade during WWII. Offspring conceived at the height of the famine were found to be at a 2-fold increased risk of schizophrenia in adulthood. Recent research on the Chinese famine in 1959–1960 replicated this finding in a vastly different context. The Chinese famine was brought about during the Great Leap Forward, an attempt at rapid industrialization in China. As in the Dutch example, there was a 2-fold excess risk of schizophrenia among individuals conceived during the height of the famine in Anhui Province. A third study from another region in China also provides concordant results.

The notion of latent effects is not intuitive. The insight that comes with the lengthened causal time frame extends our notion of “harm” to include the wellbeing of the infant, child and adult. By extending our consideration of time we are confronted with the paradox that interventions known to be beneficial for infants and children may have undetected adverse consequences in adulthood and vice versa, as well as with the complexities that accompany consideration of more than one disease outcome.

The outcomes we should be investigating are not always obvious across large spans of time, or easily ascertained. The article in this issue by Stalberg and colleagues illustrates a valuable strategy to establish hypothesized exposure-latent outcome associations. Linking birth and psychiatric registries to hospital ultrasound records, the authors were able to explore the association between prenatal ultrasound exposures and risk of schizophrenia decades later. In this instance, the strategy was applicable because the exposure and outcome were available in treatment records and registries. This study also illustrates the power that national registries in Scandinavia and Israel have brought to the study of schizophrenia.

Investigating latent effects of gestational exposures can also begin in older cohorts, such as the National Collaborative Perinatal Study and the Child Health and Development Study. These pregnancy/birth cohorts were established in the mid-twentieth century. The initial intent of these projects included assessing the impact of infection, medications, cigarettes, alcohol, and other exposures during gestation on obstetric, infant and childhood outcomes. Some of the par-
participants in this research have been followed into middle age for a variety of outcomes, notably schizophrenia.\(^{19,20}\) The utility of these cohorts, however, is limited to examining those exposures in existence at the time the cohorts were assembled, and identifiable through information previously collected or detectable in stored biologic samples.

The scope of relevant early life experience is increasingly appreciated, and this research challenge is being taken up by investigators in many areas, including psychiatry.\(^{21}\) Thinking broadly, in modern societies millions of children are prenatally exposed to antiretroviral cocktails, maternal antidepressants, ultrasound, and in vitro fertilization procedures, and virtually all children are prenatally exposed to toxins (potential teratogens) in our environment.\(^{22}\) Increased understanding of the health impact of these intentional and unintentional exposures during gestation and childhood is a major motivation for establishing a large pregnancy/birth cohort of approximately 100,000 in the United States (The National Children’s Study),\(^{23}\) following the precedents of Norway (Norwegian Mother and Child Cohort)\(^{24}\) and Denmark (Danish National Birth Cohort).\(^{25}\) Although these longitudinal studies represent a significant step in tracking the effects of environment on early growth and development, their contributions to our understanding of latent effects expressed in mid- and late-life are a distant return on this investment. In the meantime substantial smaller cohorts that are more advanced in age will be contributing to our understanding of adult outcomes.\(^{26}\)

No single design will circumvent the need for foresight, ingenuity of design, passage of time, and good luck required to capture latent effects. The lifecourse framework, however, provides a broad context and appropriate mindset for designing effective strategies to examine these questions. Having accepted the notion that early events—such as prenatal exposures—can have significant latent effects, we must also move beyond the exposure-outcome association to articulate mechanisms. Understanding these will increase our capacity to project possible effects at different points in the life course and begin to realistically appreciate benefits and costs of a given exposure or intervention. Acting from the lifecourse perspective, we should aim to exceed the mandate of “do no harm,” and promote lifetime health.

### ABOUT THE AUTHORS

MICHAELINE BRESNAHAN is Assistant Professor in the Department of Epidemiology at Mailman School of Public Health, Columbia University. Dr. Bresnahan’s work has focused on neurodevelopmental schizophrenia and autism, and pregnancy/birth cohorts research. EZRA SUSSER is the Anna Cheskis Gelman and Murray Charles Gelman Professor and Chair of the Department of Epidemiology at Mailman School of Public Health, and Professor of Psychiatry at Columbia University. Dr. Sussers’s primary research has been on the epidemiology of mental disorders, and on the role of early life experience in health and disease throughout the lifecourse. He and colleagues recently completed a textbook Psychiatric Epidemiology, Searching for the Causes of Mental Disorders.
Ambient Air Pollution and Cardiac Arrhythmias in Patients With Implantable Defibrillators

Kristina B. Metzger,*† Mitchel Klein,*† W. Dana Flanders,* Jennifer L. Peel,‡ † James A. Mulholland,§ Jonathan J. Langberg,¶ and Paige E. Tolbert*†

Background: Previous studies of ambient air pollution and ventricular tachyarrhythmias in patients with implantable cardioverter defibrillators have yielded mixed results.

Methods: We examined this relationship in a study of 518 patients with 6287 tachyarrhythmic event-days over a 10-year period in Atlanta, Georgia. The air quality data included daily measurements of PM2.5 mass and oxygenated hydrocarbons for the final 4 years of the study. Our primary analyses utilized generalized estimating equations, controlling for long-term time trends and meteorologic conditions as well as residual correlation within subjects.

Results: Our primary modeling approach found no association; additional sensitivity analyses and alternative analytic approaches supported those findings. The most suggestive positive findings were for coarse particles.

Conclusions: The present study constitutes the largest study to date of ambient air pollution and tachyarrhythmic events in patients with implantable cardioverter defibrillators. Other than the suggestive association between ambient air pollution and tachyarrhythmic events in patients with implantable cardioverter defibrillators, the study provides little evidence of an association with ventricular tachycardia or fibrillation—rapid and potentially fatal heart rhythms—it delivers a high-energy shock (defibrillation) or provides rapid pacing to the ventricles to terminate the malignant tachyarrhythmia and restore normal rhythm.

Information regarding the date and time of each tachyarrhythmic event, as well as the type of therapy delivered, is recorded and stored in the device. These recordings can be used retrospectively to identify past tachyarrhythmic events. The devices are routinely interrogated, and the event information is downloaded and stored in clinic records.

A consistent link between cardiovascular morbidity and ambient air pollution has been demonstrated in numerous epidemiologic studies.1–3 The risk of adverse events seems to be increased for certain subpopulations, including those with underlying health conditions such as diabetes, chronic obstructive pulmonary disease, congestive heart failure, previous arrhythmia, and hypertension.4–8 Several studies suggest that air pollution is associated with adverse changes in cardiac autonomic function such as increased heart rate,9,10 decreased heart rate variability,11–14 and increased systolic blood pressure.15 Results from studies of ventricular tachyarrhythmia events in patients with implantable cardioverter defibrillators have been inconsistent.16–21

Because patients with implantable defibrillators often have underlying conditions that put them at high risk of ventricular tachycardia and sudden cardiac death,22 this population could be particularly susceptible to the potential adverse effects of ambient air pollution. The implanted device continuously monitors the heart rate for abnormal heart rhythms. When the implantable defibrillator detects ventricular tachycardia or fibrillation—rapid and potentially fatal heart rhythms—it delivers a high-energy shock (defibrillation) or provides rapid pacing to the ventricles to terminate the malignant tachyarrhythmia and restore normal rhythm.
Ambient Air Quality Data

For the period 1 January 1993 to 31 December 2002, we obtained ambient air quality data for 24-hour average PM$_{10}$ (particulate matter with an aerodynamic diameter less than 10 $\mu$m) mass, 8-hour maximum ozone, and 1-hour maximum nitrogen dioxide (NO$_2$), carbon monoxide (CO), and sulfur dioxide (SO$_2$) from several existing monitoring networks, including the Air Quality System, the Georgia Department of Natural Resources, and the Metro Atlanta Index. Daily measurements of these pollutants from a central monitor located in downtown Atlanta were used in the analyses. For days on which a measurement was not available from the central monitor, we imputed pollution levels using data from at least 1 secondary monitoring site and meteorologic factors. Ozone levels were not monitored during the winter months, when ozone levels in Atlanta are low; the remaining pollutants were measured year-round. Data from the Air Quality System have been described previously.$^{2,23}$

For the period 1 August 1998 through 31 December 2002, an extensive suite of pollutants, including PM components, were measured at the ARIES monitoring station. Included in these analyses were 24-hour averages of PM$_{2.5}$ (particulate matter with an aerodynamic diameter less than 2.5 $\mu$m) mass, coarse PM (particulate matter with an aerodynamic diameter between 2.5 and 10 $\mu$m), oxygenated hydrocarbons, and the following PM$_{2.5}$ mass components: sulfate, organic carbon, elemental carbon, and an index of water-soluble metals. Although ultrafine particle counts and PM$_{2.5}$ acidity were measured at the ARIES monitoring station for its first 25 months of operation, these analytes were not included in these analyses because missing data led to model instability. The ARIES air quality data have been described previously.$^{2,23}$ The spatial variation and correlation between monitors for many of the pollutants studied here [nitric oxides (NO$_x$), CO, SO$_2$, 24-hour PM$_{2.5}$, and the following 24-hour PM$_{2.5}$ components: sulfates, organic carbon, and elemental carbon] have been examined.$^{24}$ Measurements from the ARIES monitoring station were consistent with those from the Air Quality System and other monitoring stations.$^{24}$ As expected, the spatial variability was greater for primary pollutants than for secondary pollutants.$^{24}$

Temperature and dewpoint temperature, as well as additional meteorological data measured at Hartsfield-Atlanta International Airport, were obtained from the National Climatic Data Center network.

Statistical Methods

Analyses were carried out using SAS statistical software, version 9.1 (SAS Institute Inc., Cary, NC). Our primary analytic approach involved repeated-measures logistic regression utilizing generalized estimating equations (GEE) to account for residual autocorrelation within subjects.$^{25}$ A dichotomous outcome variable was defined to distinguish whether a patient experienced at least 1 ventricular tachyarrhythmic event on a given day. Three types of tachyarrhythmic events were considered for the analysis: 1) any ventricular tachyarrhythmic event recorded by the device, 2) any ventricular tachyarrhythmic event that resulted in electrical therapy (cardiac pacing or defibrillation), and 3) any ventricular tachyarrhythmic event that resulted in defibrillation (a subset of the second group, and considered to be the most serious events).

The primary model had the following form:

$$\logit(E(Y_{ij})) = \alpha + \beta \text{pollutants}_i + \sum_{k=1}^{3} \lambda_k \text{DOW}_{ijk} + \sum_{p} \delta_p \text{holiday}_{ijp}$$

$$+ \delta_1 \text{temperature}_{ij} + \delta_2 \text{temperature}_{ij}^2 + \delta_3 \text{temperature}_{ij}^3$$

$$+ \eta_1 \text{dewpoint}_{ij} + \eta_2 \text{dewpoint}_{ij}^2 + \eta_3 \text{dewpoint}_{ij}^3 + g(y_1, \ldots, y_{N_j}; t_{ij})$$

$Y_{ij}$ was an indicator for an event-day or nonevent-day for subject $i$ on day $j$. Models included indicator variables for day of week (DOW) and federal holidays (holiday). We chose this model based on a priori considerations. Our goal was to control for suspected confounding related to time trends, seasonality, DOW, temperature, and dewpoint. To reduce model assumptions, we controlled for long-term time trends and seasonality (time) with cubic splines, $g(y_1, \ldots, y_{N_j}; x)$ using seasonal knots. Cubic terms for daily maximum temperature (temperature) and mean dewpoint (dewpoint) were included in the model (lagged 0 days). The follow-up time for each subject was divided into 6-month intervals by warm (April 15 to October 14) and cold (October 15 to April 14) seasons. To allow for autocorrelation, a stationary 21-dependent correlation structure was specified as the working correlation matrix for each cluster by subject and season. Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated for an increase of approximately 1 standard deviation (SD) of pollutant levels. We chose to model pollutant levels on the same day as the event day (lagged 0 days) based on results from a previous study of cardiovascular emergency department visits in Atlanta, which reported the strongest associations with ambient air pollution levels on the same day.$^2$

We conducted several sensitivity analyses using the GEE framework in order to evaluate the robustness of our primary model. We considered alternative ways of modeling meteorology, such as controlling for daily minimum temperature and using indicator variables for 5-degree increments of daily maximum temperature. We explored models with alternative lag structures, including pollution levels on the previous 6 days (unconstrained distributed lag models) and a
moving average of pollution lagged 0 and 1 day and of pollution lagged 0, 1, and 2 days. Season-specific analyses were conducted for the warm and cold seasons. As an alternative to modeling pollution as a log-linear variable, we categorized pollutant levels into quintiles. We also considered alternative correlation structures for GEE models. Additionally, in some models, we included indicator terms for recent ventricular tachyarrhythmic events and evaluated the potential interaction between the pollutant and occurrence within 3 days of a prior event.

We compared 2 other modeling approaches to the primary GEE model. A case-crossover analysis provided an alternative approach to control for long-term temporal trends.26–29 In these analyses, we selected comparison days within the same calendar month as the event day (to limit long-term trends), and on the same DOW (to eliminate DOW trends). All strata were assumed independent. This method of defining strata, based on subject and month, avoids bias due to overlapping sampling of controls by selecting from mutually exclusive time windows, and reduces potential autocorrelation by selecting the days at 7-day intervals. Since temperature and dewpoint may vary within a month, we included linear terms for maximum temperature and mean dewpoint, along with product terms with monthly indicator variables, allowing interaction between month and meteorology.

A second sensitivity analysis stratified only on subject. This approach coupled with conditional logistic regression allowed for the control of subject-specific fixed effects. Unlike the case-crossover approach, this method does not inherently match on time, so we included cubic splines in the model for the control of time trends, and indicator variables for DOW and federal holidays, as well as cubic terms for daily maximum temperature and daily mean dewpoint.

RESULTS

Information on 884 patients with implantable defibrillators was collected from 3 clinics during the study period (1 January 1993 to 31 December 2002) (Table 1). The majority of patients were men (78%). The age at the start of follow-up ranged from 15 to 88 years, with a mean of 61 years. The age and sex distribution were similar for those whose ventricular tachyarrhythmias triggered events cardiac pacing or defibrillation and those whose arrhythmias did not. The 518 patients with at least 1 tachyarrhythmia contributed 491,181 days of follow-up time and experienced 13,108 events on 6287 event-days. Three-fourths (72%) of patients with arrhythmia experienced electrical therapy (cardiac pacing or defibrillation); 57% experienced defibrillation. Among those with arrhythmia, the number of days with arrhythmia ranged from 1 to 198, with a mean of 12. One-fourth (24%) of those patients experienced tachyarrhythmia for only 1 day. Among patients who had more than 1 day with arrhythmia, these days appeared to be clustered in time; 21% of event-days were on the day after an arrhythmia, and 50% were within a week of a previous event-day.

The number of patients being monitored and the number of event-days per year generally increased over time; however, the rate of event-days per follow-up time was highest in 1993, decreased through 1995, and then increased again through 2000 (results available with the online version of this article). A slight seasonal pattern in the rate of event-days was observed, with peaks in July and December (results not presented). The weekly pattern indicated a slight peak in the rate of event-days on Monday (results not presented).

Daily concentrations of the air quality analytes are described in Table 2. Measurements of PM10, ozone, NO2, CO, and SO2 were available for the entire study period, 1 January 1993 to 31 August 2002, whereas measurements of coarse PM, PM2.5, PM1.5 components, and oxygenated hydrocarbons were available from 1 August 1998 to 31 December 2002. Correlations among the pollutants are presented in an online supplement.

Our results provided little evidence of associations between ambient air quality measurements and ventricular tachyarrhythmic events (Table 3). The ORs were generally consistent with the null in our primary analyses using GEE models. All of the CIs for the ORs included the null value except for a negative association of PM2.5 water-soluble metals and tachyarrhythmic events that resulted in any electrical therapy (pacing or defibrillation). The associations involving pollutants from the ARIES monitoring station, measured for the subperiod 1998–2002, were less precise than those for criteria pollutants, measured from 1993 to 2002. Estimates involving events resulting in defibrillation (a subset of events in the other outcome groups) also had wider CIs due to the lower number of event-days.

Results from additional sensitivity analyses of the primary GEE model were similar, with little or no association
between air pollution levels and tachyarrhythmic events. ORs from models that controlled for minimum rather than for maximum temperature were systematically more positive (Table 4). Eleven of the 12 pollutant ORs for any tachyarrhythmic events were greater than 1, but the CIs for all pollutants except coarse PM included the null value. We also examined several alternative lag structures. Results from unconstrained distributed lag models (pollution lagged 0–6 days) were also consistent with the null (Table 4). Models that included pollution lagged 0 and 1 day and models with pollution lagged 0, 1, and 2 days also suggested no association (results not presented).

**TABLE 2. Daily Ambient Air Quality Measurements**

<table>
<thead>
<tr>
<th>Percent of Days Missing</th>
<th>Mean ± SD</th>
<th>Minimum</th>
<th>10%</th>
<th>Median</th>
<th>90%</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 January 1993 – 31 December 2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-h PM₁₀ (µg/m³)</td>
<td>6.1</td>
<td>28.0 ± 12.2</td>
<td>0.5</td>
<td>13.8</td>
<td>26.4</td>
<td>44.7</td>
</tr>
<tr>
<td>8-h ozone (ppb)*</td>
<td>33.1</td>
<td>53.9 ± 23.0</td>
<td>2.9</td>
<td>26.1</td>
<td>52.1</td>
<td>84</td>
</tr>
<tr>
<td>1-h NO₂ (ppb)</td>
<td>0.7</td>
<td>44.9 ± 17.7</td>
<td>7</td>
<td>24</td>
<td>43</td>
<td>68</td>
</tr>
<tr>
<td>1-h CO (ppm)</td>
<td>1.2</td>
<td>1.7 ± 1.1</td>
<td>0.1</td>
<td>0.5</td>
<td>1.4</td>
<td>3.2</td>
</tr>
<tr>
<td>1-h SO₂ (ppb)</td>
<td>0.7</td>
<td>15.5 ± 16.4</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>Maximum temperature (°F)</td>
<td>0</td>
<td>72.4 ± 15.3</td>
<td>18</td>
<td>51</td>
<td>74</td>
<td>91</td>
</tr>
<tr>
<td>Minimum temperature (°F)</td>
<td>0</td>
<td>53.3 ± 15.0</td>
<td>6</td>
<td>32</td>
<td>55</td>
<td>72</td>
</tr>
<tr>
<td>Dewpoint (°F)</td>
<td>0</td>
<td>50.7 ± 16.2</td>
<td>−2.4</td>
<td>27.5</td>
<td>53.8</td>
<td>69.5</td>
</tr>
<tr>
<td>1 August 1998 – 31 December 2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-h PM₁₀ (µg/m³)</td>
<td>2.7</td>
<td>17.8 ± 8.6</td>
<td>1.7</td>
<td>8.2</td>
<td>16.2</td>
<td>29.5</td>
</tr>
<tr>
<td>24-h coarse PM (µg/m³)</td>
<td>8.5</td>
<td>9.6 ± 5.4</td>
<td>0.5</td>
<td>3.9</td>
<td>8.7</td>
<td>16.7</td>
</tr>
<tr>
<td>24-h PM₁₀, water-soluble metals (µg/m³)</td>
<td>11.3</td>
<td>0.029 ± 0.024</td>
<td>0.003</td>
<td>0.009</td>
<td>0.022</td>
<td>0.058</td>
</tr>
<tr>
<td>24-h PM₁₀, sulfates (µg/m³)</td>
<td>11.5</td>
<td>5.0 ± 3.4</td>
<td>0.5</td>
<td>1.7</td>
<td>4.1</td>
<td>9.7</td>
</tr>
<tr>
<td>24-h PM₁₀, organic carbon (µg/m³)</td>
<td>4.7</td>
<td>4.4 ± 2.4</td>
<td>0.4</td>
<td>2</td>
<td>3.9</td>
<td>7.3</td>
</tr>
<tr>
<td>24-h PM₁₀, elemental carbon (µg/m³)</td>
<td>4.8</td>
<td>1.7 ± 1.2</td>
<td>0.1</td>
<td>0.6</td>
<td>1.4</td>
<td>3.2</td>
</tr>
<tr>
<td>24-h oxygenated hydrocarbons (ppb)</td>
<td>25.5</td>
<td>31.1 ± 15.3</td>
<td>0.7</td>
<td>13.3</td>
<td>29.1</td>
<td>51</td>
</tr>
</tbody>
</table>


**TABLE 3. Results of Primary GEE Models for the Associations of Daily Ambient Air Quality Measurements and Tachyarrhythmic Events in Patients With Implantable Cardioverter Defibrillators**

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Unit*</th>
<th>All Events (n = 6287 Event-Days)</th>
<th>Events Resulting in Cardiac Pacing or Defibrillation (n = 2539 Event-Days)</th>
<th>Events Resulting in Defibrillation (n = 821 Event-Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td>1 January 1993 – 31 December 2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-h PM₁₀</td>
<td>10 µg/m³</td>
<td>0.997 (0.970–1.025)</td>
<td>0.972 (0.932–1.014)</td>
<td>0.988 (0.909–1.075)</td>
</tr>
<tr>
<td>8-h ozone</td>
<td>25 ppb</td>
<td>0.997 (0.929–1.070)</td>
<td>0.982 (0.886–1.088)</td>
<td>1.081 (0.904–1.294)</td>
</tr>
<tr>
<td>1-h NO₂</td>
<td>20 ppb</td>
<td>0.999 (0.965–1.035)</td>
<td>1.009 (0.957–1.063)</td>
<td>1.048 (0.955–1.151)</td>
</tr>
<tr>
<td>1-h CO</td>
<td>1 ppm</td>
<td>0.999 (0.970–1.028)</td>
<td>1.008 (0.964–1.054)</td>
<td>1.012 (0.925–1.107)</td>
</tr>
<tr>
<td>1-h SO₂</td>
<td>20 ppb</td>
<td>1.002 (0.968–1.037)</td>
<td>0.988 (0.936–1.042)</td>
<td>1.004 (0.911–1.105)</td>
</tr>
<tr>
<td>1 August 1998 – 31 December 2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-h PM₁₀</td>
<td>10 µg/m³</td>
<td>0.995 (0.953–1.039)</td>
<td>0.982 (0.920–1.049)</td>
<td>0.969 (0.846–1.110)</td>
</tr>
<tr>
<td>24-h coarse PM</td>
<td>5 µg/m³</td>
<td>1.031 (0.997–1.066)</td>
<td>1.048 (0.995–1.104)</td>
<td>1.049 (0.943–1.166)</td>
</tr>
<tr>
<td>24-h PM₁₀, water-soluble metals</td>
<td>0.03 µg/m³</td>
<td>0.952 (0.904–1.003)</td>
<td>0.897 (0.821–0.979)</td>
<td>0.891 (0.741–1.071)</td>
</tr>
<tr>
<td>24-h PM₁₀, sulfate</td>
<td>5 µg/m³</td>
<td>0.994 (0.932–1.061)</td>
<td>0.969 (0.880–1.067)</td>
<td>0.996 (0.805–1.233)</td>
</tr>
<tr>
<td>24-h PM₁₀, organic carbon</td>
<td>2 µg/m³</td>
<td>1.005 (0.977–1.034)</td>
<td>0.988 (0.941–1.039)</td>
<td>0.967 (0.883–1.060)</td>
</tr>
<tr>
<td>24-h PM₁₀, elemental carbon</td>
<td>1 µg/m³</td>
<td>1.013 (0.982–1.045)</td>
<td>0.984 (0.936–1.055)</td>
<td>0.939 (0.860–1.025)</td>
</tr>
<tr>
<td>24-h oxygenated hydrocarbons</td>
<td>15 ppb</td>
<td>0.996 (0.954–1.040)</td>
<td>0.970 (0.912–1.031)</td>
<td>0.905 (0.797–1.028)</td>
</tr>
</tbody>
</table>

*Approximately 1 standard deviation.
Results from separate analyses for warm and cold seasons were also largely consistent with the null, with the exception of a positive association for coarse PM in the cold season (Table 5). For PM$_{10}$, PM$_{2.5}$, sulfate, elemental carbon, and organic carbon, the ORs were more positive during the cold season. For CO, SO$_2$, and oxygenated hydrocarbons, the ORs were more positive during the warm season. For PM$_{2.5}$ water-soluble metals.

Coarse particles, which like fine particles can penetrate into the thoracic region, may have health effects independent of fine particles. Particles in this size tend to have chemical composition and source contributions that are different from those of fine particles. Fine particles are predominantly by-products of combustion processes, whereas coarse particles include relatively more crustal material, resuspended road dust containing tire and brake residues, metals, and plant and animal matter. Recent studies have attempted to assess the relative contributions of the 2 size fractions in health studies. A review of health effects of coarse particles by Brunekreef finds support for an association of coarse PM and cardiovascular admissions in studies that included information on both size fractions. Gong et al reported increased heart rate and decreased heart-rate variability in volunteers exposed to concentrated ambient coarse particles. Additional evidence of reduced heart-rate variability comes from a study by Lipsett et al in relation to coarse particles with a high content of wind-blown sand. In addition, laboratory studies are beginning to provide some insights regarding biological responses to coarse particles. For example, in an in vitro study of normal human bronchial epithelial cells and ambient particles, Becker et al reported that coarse particles were more potent in inducing cytokines (IL-6 and IL-8) than fine and ultrafine particles, and had similar potency as the other size fractions in stimulating the production of reactive oxygen species.

Previous investigations of the association between ambient air pollution and tachyarrhythmia in patients with implantable defibrillators have reported mixed results. The first published pilot study of 33 patients suggested that
Tachyarrhythmic events were associated with increased levels of NO₂, CO, and PM₁₅. A larger study of 203 patients in the same population reported results generally consistent with the null, except for events after a recent ventricular tachyarrhythmia. Associations with particles, NO₂, CO, and SO₂ were reported in patients who had a tachyarrhythmic event within the previous 3 days, a potential interaction not observed in the present study. A case-crossover analysis on a subset of these patients reported positive associations of ventricular arrhythmias and PM₂₅ and ozone. Another study of ventricular arrhythmias in 50 patients in Vancouver, Canada, reported no association between events and ambient air pollution levels in primary analyses; associations were observed for SO₂ in a subset of 16 patients with more than 6 months of follow-up and more frequent events. An alternative analysis of a subset of that population using case-crossover methodology also yielded null results. Most recently, a study of 56 patients living in St. Louis reported an increase in events of ventricular arrhythmia associated with an increase in ambient levels of SO₂. Previous studies of patients with implantable defibrillators have not examined the association of coarse PM and ventricular tachyarrhythmias.

We conducted several sensitivity analyses in order to evaluate the robustness of our primary GEE model. Most of the results were consistent with no association. The results from the GEE models controlling for minimum temperature, rather than maximum temperature were systematically more positive, although still largely consistent with the null. This difference may suggest residual confounding by temperature in the model with minimum temperature. In preliminary models that excluded all pollutant terms, maximum temperature was more strongly associated with tachyarrhythmic events than minimum temperature. Additionally, the regression parameters from models that included terms for both maximum and minimum temperature were more similar to those from models controlling for only maximum temperature rather than those controlling for only minimum temperature.

Outcome misclassification in the present study is likely to be minimal. An electrophysiologist or trained technician reviewed the intracardiac electrograms for each event detected by the implantable defibrillator at the time the information was downloaded. Detected events that were not ventricular in origin (eg, supraventricular tachycardia, sinus tachycardia, lead fracture, electromagnetic interference, or oversensing—constituting approximately 25% of all detected events) were excluded from the analysis. We did not have an additional review of the electrograms by an electrophysiologist. Our collaborating electrophysiologist (J.J.L.) determined that such a review would have had limited usefulness given the routine review performed at each clinic, and would likely have led to reclassification of very few events. Any resulting bias to the null arising from outcome misclassification is likely to be small. Given the power of the study conferred by the large sample size, a strong association is not likely to be missed.

An additional issue in studies of this type is that personal exposure measurements are not available. In the current study involving 518 patients (far fewer than is typical of hospital admission or mortality studies), the longitudinal correlation of personal and ambient levels may be insufficient to support the use of the available ambient air quality data for etiologic investigation. Spatial heterogeneity of ambient air pollution could also have affected our results, possibly biasing them toward the null, because the centrally located monitor may not have accurately measured air pollutants for distant areas. The spatial variation and correlation between

### Table 5. Results of Season-Specific GEE Models for the Associations of Daily Ambient Air Quality Measurements and Tachyarrhythmic Events in Patients With Implantable Cardioverter Defibrillators

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Unit*</th>
<th>Warm Season (April 15 to October 14) OR (95% CI)</th>
<th>Cold Season (October 15 to April 14) OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>24-h PM₁₀</td>
<td>10 µg/m³</td>
<td>0.980 (0.939–1.022)</td>
<td>1.018 (0.981–1.057)</td>
</tr>
<tr>
<td>8-h ozone</td>
<td>25 ppb</td>
<td>0.998 (0.929–1.071)</td>
<td>0.917 (0.763–1.103)</td>
</tr>
<tr>
<td>1-h NO₂</td>
<td>20 ppb</td>
<td>1.003 (0.952–1.055)</td>
<td>1.005 (0.958–1.054)</td>
</tr>
<tr>
<td>1-h CO</td>
<td>1 ppm</td>
<td>1.051 (0.976–1.085)</td>
<td>0.989 (0.941–1.034)</td>
</tr>
<tr>
<td>1-h SO₂</td>
<td>20 ppb</td>
<td>1.029 (0.989–1.116)</td>
<td>0.986 (0.956–1.023)</td>
</tr>
</tbody>
</table>

*Approximately 1 standard deviation.
monitors for many of the pollutants studied here (NO$_x$, CO, SO$_2$, 24-hour PM$_{2.5}$, and the 24-hour PM$_{2.5}$ components of sulfates, organic carbon, and elemental carbon) have been examined.$^{28}$ Measurements from the ARIES monitoring stations were consistent with those from the Air Quality System and other monitoring stations. As expected, the spatial variability was greater for primary pollutants than for secondary pollutants.$^{24}$ These findings would tend to weaken but likely not eliminate associations with ambient air pollution levels for some pollutants. Thus, although the study provides evidence suggesting that a strong association of ventricular tachyarrhythmic events with the ambient pollutants assessed is unlikely, it does not eliminate the possibility of weak associations, nor does it contribute evidence regarding whether personal exposure may be a determinant of ventricular tachyarrhythmia. Additionally, behavior such as air conditioning use or time spent outdoors may affect personal exposure levels. This could affect the magnitude of the observed associations in comparison to other geographic locations.

Although the association between ambient air pollution and cardiovascular events has been noted in numerous epidemiologic studies, results from studies examining the association of air pollution and arrhythmic events in patients with implantable defibrillators have been less consistent. The present study does not support a strong association with ventricular tachyarrhythmic events in such patients. This study included more patients over a longer period of time with more fully characterized ambient air quality measures (including detailed measurements of PM$_{2.5}$ mass), than previous studies. Although we cannot exclude the possibility that potential misclassification of tachyarrhythmic events may have attenuated the magnitude of the results slightly, the likelihood of missing a strong association with ambient pollutant levels was low. We conducted several alternative analyses, all of which indicated a lack of an association; only coarse particulate matter provided any suggestion of a positive association. Future studies may provide further insight into these findings.

ACKNOWLEDGMENTS
We thank Keely Cheslack-Postava, Melissa Berkowitz, Amie May, Rebecca Gunter, Aimee Cunningham, and Laurie Gold for their assistance on the project. We are also grateful for the cooperation of the staff at the participating clinics.

REFERENCES


Short-Term Effects of Particulate Air Pollution on Male Smokers and Never-Smokers

Chit-Ming Wong, Chun-Quan Ou, Nga-Wing Lee, King-Pan Chan, Thuan-Quoc Thach, Yuen-Kwan Chau, Sai-Yin Ho, Anthony Johnson Hedley, and Tai-Hing Lam

Background: Numerous studies have shown that ambient air pollution and smoking are both associated with increased mortality, but until now there has been little evidence as to whether the effects of these 2 factors combined are greater than the sum of their individual effects. We assessed whether smokers are subject to additional mortality risk from air pollution relative to never-smokers.

Methods: This study included 10,833 Chinese men in Hong Kong who died at the age of 30 or above during the period 1 January to 31 December 1998. Relatives who registered for deceased persons were interviewed about the deceased’s smoking history and other personal lifestyle factors about 10 years before death. Poisson regression for daily number of deaths was fitted to estimate excess risks per 10 \( \mu g/m^3 \) increase in particulate matter with aerodynamic diameter <10 \( \mu m \) (PM\(_{10}\)) in male smokers and never-smokers in stratified data, and additional excess risk for smokers relative to never-smokers in combined data.

Results: In smokers there was a significant excess risk associated with PM\(_{10}\) for all natural causes and cardio-respiratory diseases for men age 30 years or older and men 65 or older. For all natural causes, greater excess risk associated with PM\(_{10}\) was observed for smokers relative to never-smokers: 1.9% (95% confidence interval = 0.3% to 3.6%) in men age 30 and older and 2.3% (0.4% to 4.3%) in those age 65 and older.

Conclusions: Ambient particulate air pollution is associated with greater excess mortality in male smokers compared with never-smokers.

(Epidemiology 2007;18: 593–598)
the design and conduct of the investigation had been published elsewhere.  

The present analysis included male never-smokers and daily smokers who smoked at least one cigarette per day 10 years before death. Only men were included, as regular daily smoking in women was less prevalent (13%) than in men (51%).

The unique death registration number was linked to the Hong Kong Department of Health databases to obtain the certified underlying cause of death. The date of death was linked to the Environmental Protection Department databases from 8 monitoring stations for daily 24-hour concentrations of particulate matter with aerodynamic diameter <10 μm (PM$_{10}$) and to the Hong Kong Observatory for daily temperature and relative humidity data. The daily PM$_{10}$ concentrations among the 8 monitoring stations are highly correlated (Spearman correlation coefficient ranging from 0.86 to 0.96), and the population density (6.8 millions in 1092 km$^2$) of Hong Kong is very high; it is therefore most likely valid to compute the overall arithmetic mean daily PM$_{10}$ concentrations of the 8 stations after eliminating discrepancies between stations, and use them as a proxy measure for environmental exposure. The Ethics Committee of the institute where this study was carried out has approved the study proposal.

**Major Outcomes**

The underlying cause of death was coded according to the *International Classification of Disease* 9th revision (ICD-9). The health outcomes were deaths from all natural causes (ICD-9 1-799) and for cardiorespiratory diseases (ICD-9 390-519). We did analysis for all ages (ie, age 30+) and for the 65 or older age (age 65+) group to show the results for the complete data and for the older population.

**Risk Assessment**

We defined 2 measures of mortality risks; excess risk (ER) of death per 10 μg/m$^3$ increase in pollutant for smokers and never-smokers, and difference in ER for air pollution effects in smokers relative to never-smokers. The 95% confidence intervals (CIs) for the ERs were also computed.

**Statistical Methods**

Time-series count data were stratified according to smoking status of the deceased 10 years before death. Generalized additive Poisson regression model for daily counts $Y_t$ at calendar day $t$ ($t = 1, 2 \ldots 365$) of each health outcome was constructed. The variables in the model were day of the week, holiday, influenza epidemic indicators, temperature and relative humidity for observable confounding variables $Z_t$, and a smoothing function of $t$ ($S(t, df)$, $t = 1, 2, 3 \ldots 365$) defined with certain degrees of freedom (df) for confounding variables that are not directly observable but may be related to long-term trends and seasonal variations of the data.  

Thus,

\[ \log(\mu_t) = A + gZ_t + hC + S(t, df) \]  

where $A$ and $g$ are coefficients for the constant term and $Z_t$ the confounding variables of the regression model.

We then combined the stratified data with a dummy variable defined for smoking status (C, 1 for smokers and 0 for never-smokers). Poisson regression model for the combined data was specified as

\[ \log(\mu_t) = A + gZ_t + hC + S(t, df) \]  

where $h$ is the coefficient for the variable $C$ of the regression model.

Poisson regression Model 1 for stratified data and Model 2 for the combined data were fitted with a certain degrees of freedom for temperature, humidity and time until coefficients of partial autocorrelation function were less than around 0.1 and free from discernible patterns. Then, PM$_{10}$ concentration variable ($X$) at concurrent and previous 1–3 days was separately entered into Equation 1, and the effect of PM$_{10}$ in smokers and never-smokers was assessed by the coefficient associated with $X$. A product of $C$ and $X$ ($CX$) and the individual variable $X$ were then entered into Equation 2, from which the additional effect (ie, difference in effects of air pollution for smokers relative to never-smokers) was assessed by the coefficient of the CX term. Additionally, we fitted an unconstrained distributed lag model, simultaneously with inclusion of PM$_{10}$ concentrations at concurrent and previous 1–3 days and their interaction terms with smoking, to estimate the overall interaction between smoking and air pollution.

Two types of sensitivity analyses were performed. Using case-crossover analysis, we compared PM$_{10}$ levels at the day of individual death (ie, case) with the levels 7 days before and 7 days after death (ie, 2 controls). We also evaluated excess risk of mortality associated with PM$_{10}$ and additional ER in smokers compared with never-smokers using conditional logistic regression. Case-only logistic regression was also applied, with smoking status for individual deaths (no controls) as the dependent variable and PM$_{10}$ concentration as the independent variable, to detect the interaction between PM$_{10}$ and smoking status. The case-only approach does not depend on the restrictive assumption for other time-varying factors, or on the development of the core model required in Poisson regression.  

Finally we assessed copollutant effects on estimates for the interaction effects between smoking and PM$_{10}$ by entering the terms for main effect and the product between the smoking variable and the concentration for each of the 3 copollutants. All analysis was performed using R 2.0.1 program and Stata 8.2 statistical package (StataCorp, College Station, TX).

**RESULTS**

A total of 4182 male never-smokers and 6901 male smokers were included in the study. Compared with smokers, never-smokers were more likely to be older, locally born, better educated, and living in self-owned housing (Table 1).

During the study period, the mean daily number of all deaths attributable to natural causes was 37.4, of which 16.8
were from cardiorespiratory diseases. The mean PM$_{10}$ concentration was $48.1 \mu g/m^3$, temperature 24.0°C, and relative humidity 79.2% (Table 2).

Table 3 presents separate estimates of PM$_{10}$ effects for smokers and never-smokers. In smokers, the most significant effects of PM$_{10}$ were associated with exposures at zero or 2 days before death. Among smokers age 30 or more, the excess risks for exposures 2 days before death were 1.8% per 10 $\mu g/m^3$ increase in PM$_{10}$ (95% CI 0.5% to 3.1%) for all natural causes, and 2.3% (0.2% to 4.4%) for cardio-respiratory diseases. The same excess risks for smokers of age 65 were 2.4% (0.7% to 4.1%) and 2.6% (0.3% to 5.0%). In never-smokers, no excess risks were observed.

Table 4 shows the results for interaction between smoking and PM$_{10}$, namely, the additional ER of death
associated with PM\textsubscript{10} in smokers compared with never-smokers. The additional ER of death for exposures 2 days before death from all natural causes in smokers compared with never-smokers, were 1.9% (0.3% to 3.6%) for ages 30+ and 2.2% (0.2% to 4.2%) for ages 65+. The corresponding additional ER caused by cardiorespiratory diseases were 2.2% (−0.4% to 4.8%) and 2.4% (−0.2% to 5.2%; Table 4). The estimates of additional ER from the unconstrained distributed lag model were generally larger than those at individual lag days (Table 4).

The estimates by case-crossover analyses were roughly similar but less precise than estimates from single-lag Poisson regression model (Table 4). Additional ERs in smokers compared with never-smokers using case-only logistic regression were almost identical to those obtained by single-lag Poisson regression model. The estimated effect modification of smoking on PM\textsubscript{10} effects changed slightly after adjustment for the main effects of each of the 3 copollutants (data now shown) and increased to a greater extent after also adjustment for their interactions with smoking (Supplementary Table 1, available with the online version of the article).

**DISCUSSION**

Health effects of air pollution estimated from daily time-series modeling have been increasingly used in public health decision-making.\textsuperscript{20,21} Health effect estimates based on data from routinely collected whole population mortality, health service utilization, or territory-wide environmental surveillance data\textsuperscript{21,22} are representative of the population. Observable confounding effects from environmental covariates, which vary with time, can be controlled by their inclusion in the core model; unobservable confounding responsible for seasonal and long-time trends can be controlled by inclusion of a smoothing function of the time variable into the Poisson regression model.\textsuperscript{16,23} In all the core models we rigorously checked that the residuals are independent and random, by well-established model diagnostics using partial autocorrelation function plots, to minimize any residual confounding effects.

Although personal risk factors are unlikely to confound time-series studies of air pollution effects because they usually do not change over a short period, these factors may modify the short-term effects of air pollution. The health effect assessment based on the whole population in daily time-series studies is a good measure of effects for all subpopulations provided that health effects are homogenous within the population. If air pollution effects vary among subpopulations, (such as the differing effects we have shown between smokers and never-smokers), the effects for more susceptible subgroups (such as smokers) would be underestimated using overall effect estimates that do not take into account the interaction between this risk factor and air pollution exposure.

We observed substantial effects of PM\textsubscript{10} in smokers, whereas the effects among never-smokers were not conclusive. We cannot determine whether the true effects for never-smokers are close to or equal to the null. One year of data,

<table>
<thead>
<tr>
<th>Table 4. Additional Excess Risk (%) and 95% CI for Mortality From all Natural Causes and Cardiorespiratory Diseases per 10 \mu g/m\textsuperscript{3} Increase in PM\textsubscript{10}, by Age (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Poisson Regression</strong></td>
</tr>
<tr>
<td><strong>Lag 0</strong></td>
</tr>
<tr>
<td>All natural causes</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Cardiorespiratory diseases</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

*Estimates using unconstrained distributed lag model with Poisson regression.
may not provide enough power to address this question; a previous simulation study showed that with one-year data the power to detect an excess risk of 2% is almost 100% and to detect an excess risk of 0.5% is 30%. However, even with this limitation it is apparent that smokers are at higher risk than nonsmokers from air pollution effects.

Although time-series analysis is the most commonly used method for assessing the short-term effects of air pollution, to our knowledge, this is the first time-series study to examine effect modification by an individual lifestyle factor through examining the potential interaction between the fixed individual factor and time-varying concentrations of air pollution. We found positive additional ER of mortality associated with PM$_{10}$ in smokers compared with never-smokers. Sensitivity analyses illustrated that the estimates of additional ER from Poisson regression were robust against 2 alternative statistical models. The estimates by case-crossover analysis were roughly similar but less precise, with larger standard error of effect estimates than those using Poisson regression, consistent with a previous report. In this analysis, the case-only approach provided almost identical estimates to those using Poisson regression, confirming the relationship between these 2 methods demonstrated previously. Time-invariant factors, such as socioeconomic status, are associated with smoking but do not vary between the levels of ambient air pollution, so such factors should not confound the estimation in case-only approach. Further, the interaction effects did not diminish after adjustment for each copollutant nitrogen dioxide, ozone or sulfur dioxide, for the main effects and for the interaction effects between copollutants and smoking status of the individuals.

A number of mechanisms could explain the positive interaction between smoking and ambient particulate pollution on mortality. A possible mechanism induced by smoking may operate through decreased clearance and increased deposition and retention of particles. In a chamber study after exposure to iron oxide particles (2.9 μm aerodynamic diameter), the alveolar long-term clearance kinetics revealed a mean half-time of 124 days in healthy nonsmokers and 208 days in smokers. Mortensen et al. reported faster mucociliary clearance in lifelong nonsmokers than in ex-smokers. Both mainstream and sidestream smoke inhibit ciliary beat frequencies and in some cases completely stop ciliary action. Chronic smoking has been shown to induce ciliary damage, nonreversible even after a long period of smoking cessation. On the other hand, on-site measurement has shown that smokers had a significantly higher total respiratory system deposition of PM$_{2.5}$ than nonsmokers.

It has been established that ultrafine particles are able to penetrate the human lung and enter the systemic circulation after inhalation. Many laboratory and epidemiologic studies have indicated that cigarette smoke induces structural disruption of the airway epithelial barrier and causes vascular endothelial dysfunction, thus increasing particulate entry (and even uptake) into the arterial wall and exacerbating consequent harmful effects. The clearance of particles penetrated and deposited is mainly subject to the macrophage-mediated phagocytosis and digestion. However, smoking has a long-term chronic effect on many important aspects of immune responses, such as neutrophil kinetics (eg, suppression of chemotaxis and phagocytosis), function of lymphocytes and cytokines levels (eg, interleukin 1β and interleukin-6). A recent case-control study found that PM$_{2.5}$ concentrations were associated with absolute neutrophil counts and white blood cell counts in nonsmokers but not in smokers. This impairment of proper immune response in smokers may inhibit the recognition and removal of particulate matter in the body. Furthermore, smoking is associated with oxidation and decreased concentrations of the major endogenous antioxidant, glutathione which would exacerbate oxidative stress induced by particulate air pollution. Lastly, heritability may play a role in etiologic mechanism, although little about this has been studied to date. The prevalence of some specific genotypes (eg, DRD2 Taq1A and GSTP1-105) is higher in smokers than that in nonsmokers. Such genotypes influence smoking behavior, including the initiation and dependence, and effects of smoking on health, and people with such genotypes may be more susceptible to the effects of exposure to air pollution. Taken together, these factors may explain why smoking may exacerbate the adverse effects of inhaled particulate air pollution, and why smokers may suffer more than never-smokers from air pollution.

Our findings could have public health impact. A coherent public health policy aimed at the reduction of avoidable mortality from air pollution should target both environmental air quality and tobacco control. The elimination of either one of these 2 exposures can lead to 2 benefits: one from avoiding the adverse effects of the exposure that has been eliminated, and the other from avoiding the interaction of the 2 exposures. Public health policy in Hong Kong, as elsewhere, needs to be evidence-driven; otherwise the actions of policy makers and law-makers are likely to be inadequate or challenged by vested interests. In addition, the results of this study can be incorporated into health promotion programs to motivate more smokers to quit, especially those who are concerned about air pollution.

**ACKNOWLEDGMENTS**

We thank Deacon Lee, Aberdeen University Medical School, for his contribution to the literature review and assistance in clerical work.

**REFERENCES**

5. Pope CA III, Burnett RT, Thurston GD, et al. Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological


Body Mass Index in Adolescence and Number of Children in Adulthood


**Background:** Body weight is associated with reproduction and related behaviors, but it is unknown whether it has significance for fertility differences in the general population. We examined whether adolescent body mass index (BMI; kg/m²) predicted the number of children in adulthood 21 years later.

**Methods:** The participants were 1298 Finnish women and men (ages 12, 15, and 18 years at baseline) followed in a prospective population-based cohort study (the Cardiovascular Risk in Young Finns) from year 1980 to 2001.

**Results:** There was an inverted J-shaped association between BMI and the number of children, such that underweight adolescents had 10–16% fewer children in adulthood, overweight adolescents 4–8% fewer, and obese adolescents 32–38% fewer than individuals with normal adolescent weight. This association was similar in women and men, and independent of age, education, urbanicity of residence, and timing of menarche (in women). Adolescents with low or high BMI were less likely to have lived with a partner in adulthood, which partly accounted for their decreased number of children. The influence of adolescent BMI was independent of adulthood BMI in women but not in men. Age at menarche also predicted the number of children, such that women with early or late menarche had more children than those with average age at menarche.

**Conclusion:** Underweight and especially obesity may have a negative impact on fertility in the general population. The increasing prevalence of obesity in children and adolescents may represent a concern for future reproductive health.

*Epidemiology 2007;18: 599–606*


From the *Department of Psychology, University of Helsinki, Finland; †International Centre for Health and Society, Department of Epidemiology and Public Health, University College London, UK; ‡National Research and Development Centre for Welfare and Health, Helsinki, Finland; §Department of Medicine, Turku University, Turku, Finland; and ¶Department of Clinical Physiology, Turku University Central Hospital, Turku, Finland.*

Supported by the Academy of Finland (grant nos. 1111056, 77841, and 210283). M. J. was supported by the Finnish Cultural Foundation, M. K. was supported by the Academy of Finland (grants 105195 and 117604), and L. K.-J. was supported by the Yrjö Jahnsson Foundation.

Correspondence: Liisa Keltikangas-Järvinen, Department of Psychology, University of Helsinki, P. O. Box 9, FIN-00014, University of Helsinki, Finland. E-mail: liisa.keltikangas-jarvinen@helsinki.fi

Copyright © 2007 by Lippincott Williams & Wilkins

ISSN: 1044-3983/07/1805-0599

DOI: 10.1097/EDE.0b013e3181257158

Obesity is associated with elevated levels of morbidity and mortality, and is considered to be one of the most important threats to public health. Obesity increases the risk of sexual and reproductive dysfunction in women and men, and obese people are less likely to get married than people of normal weight. Very low body mass index (BMI) has also been associated with dysfunctions of reproductive physiology in both men and women.

Despite the global increase in the prevalence of obesity, very little is known about the potential influence of body weight on fertility (ie, the number of children) in the general population. In a sample of parents of college students, Ellis and Haman found that higher BMI was associated with greater number of children, particularly in women. However, this study was not prospective and, therefore, the direction of causality between weight and number of children could not be determined definitely. The number of births (ie, parity) is associated with increased risk of obesity in women, and parenthood may also increase the risk of obesity in men. Moreover, the sample of Ellis and Haman included only individuals with one or more children, and potential confounding factors, such as age or level of education, were not controlled for.

Prospective epidemiologic studies with representative samples are needed to evaluate the influence of body weight on fertility in the general population. In the present study we examined whether adolescent BMI predicted the number of children in adulthood 21 years later in a population-based sample of Finnish women and men participating in a prospective study. We hypothesized that there is an inverted U-shaped association between BMI and number of children, because both low and high BMI have been associated with risks for reproduction. We also assessed whether the influence of adolescent BMI was mediated by adulthood body weight or the likelihood of attaining a mate. In women, high adolescent BMI has been shown to be related to younger menarche which, in turn, may predict earlier initiation of sexual behavior. Therefore we examined whether age at menarche was involved in the association between BMI and fertility. Skinfold thickness was used as an additional indicator of body fatness.

**METHODS**

**Participants**

The participants were 1298 women (n = 715) and men (n = 583) participating in the ongoing population-based study of Cardiovascular Risk in Young Finns. In this
study, a randomly selected sample of 3596 Finnish healthy children and adolescents from 6 birth cohorts (age 3, 6, 9, 12, 15, and 18 years at baseline) have been followed since 1980, focusing on the development of cardiovascular risk factors. Complete details of the study are provided elsewhere. For the present study, we included 3 of the oldest cohorts, who were adolescents (ie, 12, 15 and 18 years of age) at baseline and thus were adults (ie, 33, 36 and 39 years of age) in the most recent follow-up phase in the year 2001. Of 1790 study participants in these 3 cohorts at baseline, 1298 had data on the present study variables (Table 1). The participants gave written informed consent, and the study was approved by local ethics committees. There was no linear or nonlinear association between adolescent BMI and the likelihood of having participated in the follow-up, indicating that adolescent BMI was not related to study attrition.

**Measures in Adolescence**

Adolescent height and weight were measured at the baseline when the participants were 12, 15, and 18 years of age. Measurements were taken in a medical examination with a Seca weight scale and anthropometer. BMI was calculated as weight in kilograms/(height in meters) squared.

Measures of skinfold thickness were obtained at the baseline examination by Harpenden calipers (Holtain and Bull-British Indicators instruments) to 0.2-mm readings. The combined thickness of 3 skinfold measurements (subscapular, triceps, and biceps) was used in the analysis. The partial correlation (controlling for age and sex) between BMI and skinfold thickness was r = 0.69.

The women participants reported the age at menarche at the baseline and in 2 subsequent follow-up phases 3 and 6 years later. For those who had reported the age at menarche in more than one follow-up, data from the earliest possible follow-up phase were used.

**Measures in Adulthood**

Body mass index was calculated as above based on adulthood height and weight as measured in a medical examination in the follow-up phase in 2001. Waist circumference was measured midway between iliac crest and lowest rib as the average of 2 measurements with an accuracy of 0.1 cm. The partial correlation (controlling for age and gender) between adulthood BMI and waist circumference was r = 0.91.

The number of the participant’s children and their years of birth were reported by the participants in the follow-up phase when the participants were 33, 36, and 39 years of age. For partnership history we created a dichotomous variable that indicated whether or not the participant had ever lived together with a partner (ie, had either been married or cohabiting). Data for this variable were taken from all the 6 follow-up phases, in which the participants have reported their current marital status and changes in marital status between the follow-ups.

Education level was measured by the completed years of education in adulthood reported by the participants. Place of residence in adulthood was reported by the participants on a four-point scale (1 = city; 2 = suburban area; 3 = rural area; 4 = remote rural area), and was used as a continuous covariate in the analyses.

**Statistical Analysis**

**Poisson Regression**

The association between BMI and the number of children was assessed with Poisson regression, for which the equation with n independent variables is

\[
\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in})
\]

The Poisson model provided an adequate fit for the present data (Pearson goodness-of-fit \( \chi^2 = 1310.65, df = 1291 \)), and there was no significant overdispersion in the dependent variable (mean = 1.6, variance = 1.7). The Poisson regression was fitted with adolescent BMI and its square as the independent variables; adulthood number of children was the dependent variable; and covariates were age, sex, years of education and place of residence. The associations of skinfold thickness and of age at menarche with the number of children were also assessed with Poisson regression, with the same covariates as mentioned above. The parameter estimates were expressed as incidence-risk ratios (IRRs).

**Survival Analysis**

In order to determine the temporal patterning of the fertility differences, discrete-time survival analysis was used to examine whether and when the participants had their first, second, and third child, and whether adolescent BMI predicted these events. The birth of the fourth or any subsequent child was not considered since only 6% of the participants had a fourth child and only 2% had 5 or more children (Table 1). Survival analysis takes into account the phenom-

---

**TABLE 1. Descriptive Statistics** of the Sample (n = 1298)

<table>
<thead>
<tr>
<th>Sex; no. (%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>715 (55.1)</td>
</tr>
<tr>
<td>Men</td>
<td>583 (44.9)</td>
</tr>
<tr>
<td>Age in 2001</td>
<td>35.8 ± 2.4</td>
</tr>
<tr>
<td>Education</td>
<td>14.3 ± 3.3</td>
</tr>
<tr>
<td>Residence</td>
<td>2.5 ± 1.0</td>
</tr>
<tr>
<td>Has lived with a partner; no. (%)</td>
<td>1130 (87.0)</td>
</tr>
<tr>
<td>Adolescent body mass index (kg/m²)</td>
<td>19.7 ± 2.9</td>
</tr>
<tr>
<td>Adulthood body mass index (kg/m²)</td>
<td>25.4 ± 4.2</td>
</tr>
<tr>
<td>Adolescent skinfold thickness (mm)</td>
<td>29.7 ± 13.4</td>
</tr>
<tr>
<td>Adulthood waist circumference (cm)</td>
<td>85.9 ± 12.3</td>
</tr>
<tr>
<td>Age at menarche in women</td>
<td>13.2 ± 1.2</td>
</tr>
<tr>
<td>Number of children</td>
<td></td>
</tr>
<tr>
<td>Mean ± SD</td>
<td>1.6 ± 1.3</td>
</tr>
<tr>
<td>No. (%)</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>335 (25.8)</td>
</tr>
<tr>
<td>1</td>
<td>244 (18.8)</td>
</tr>
<tr>
<td>2</td>
<td>439 (33.8)</td>
</tr>
<tr>
<td>3</td>
<td>205 (15.8)</td>
</tr>
<tr>
<td>4</td>
<td>52 (4.0)</td>
</tr>
<tr>
<td>5+</td>
<td>23 (1.8)</td>
</tr>
</tbody>
</table>

*Mean ± SD, unless otherwise indicated.*
enon of censoring, ie, the fact that not all participants have their children within the study period, and some participants do not have children at all. In survival analysis a participant is censored either when the event of interest occurs (ie, a child is born) or when the study period ends (ie, at the age of 33, 36, or 39, depending on whether the participant belonged to the youngest, middle or oldest birth cohort). The survival analyses thus allowed us to determine the population estimates for the probabilities of having the first, second, and third child by the age 39. Age, sex, education and place of residence were entered as covariates. The age-specific fertility is known to follow a bell-shaped curve, so the effect of time was modeled as a nonlinear function. The parameter estimates were expressed as odds ratios of logit hazard function for one unit increase in the independent variable.

Adolescent BMI Categorization for Illustration

The results were illustrated by categorizing participants into groups of underweight, normal weight, overweight, and obese on the basis of their adolescent BMI. For the overweight and obese groups we used the international cut-off values provided by Cole et al. Based on these cut-offs, 7% of adolescent females and 9% of males were categorized as overweight and 0.9% of females and 1.4% of males were categorized as obese. In the absence of established cut-off values for adolescent underweight, we categorized BMI in the lowest 5% as underweight. This categorizing was carried out within sex and birth cohort groups, and provided the respective BMI cut-off values of 14.8, 16.6, and 17.6 for females, 12, 15, and 18 years of age, and the corresponding values of 15.2, 17.0, and 18.3 for males. Next we determined the median adolescent BMI values within the underweight, normal weight, overweight and obese groups, (16.2, 19.4, 25.1, and 30.4, respectively). The corresponding values were 15.7, 19.5, 25.1, and 29.9 for females, and 16.4, 19.2, 25.1, and 30.5 for males. These values were used as prototypical or average cases of the 4 adolescent BMI groups in illustrations of the model-predicted estimates. In these calculations other covariates were assigned their mean values.

Logistic Regression

The association between adolescent BMI and partnership history was assessed with logistic regression analysis. Adolescent BMI and its square were the independent variables; adulthood number of children was the dependent variable; and the covariates were age, sex, years of education and place of residence.

Regression Diagnostics

BMI was positively skewed, so we examined whether a transformation of BMI affected the results using an inverse transformation (ie, X=1/BMI) which corrected the skewness. The results were similar and, in terms of confidence intervals of the estimates, even slightly more precise when the transformed BMI was used (data not shown). However, the original scale provided more convenient regression coefficients, so the analyses were fitted with the original BMI scale. Outlier diagnostics indicated that extreme observations did not substantially influence the results.

With one exception (see below), the findings were similar for women and men, with no statistical evidence of differences by sex. However, because of the sex-specific nature of fertility, we also present the results separately for women and men. In some analyses the confidence intervals of the relevant parameter estimates included 1.00 in separate analyses for women and men, but not when sexes were combined. In the absence of statistical interaction by sex, we will interpret these results based on the estimates for women and men combined.

RESULTS

There was a nonlinear association between adolescent BMI and the number of children in adulthood (Table 2, Model 1), indicating that underweight and obese individuals had fewer children than those with normal weight (Fig. 1A). The predicted number of children of average members (ie, median values) of adolescent underweight, normal weight, overweight, and obese groups were, respectively, 1.49, 1.78, 1.64, and 1.10 in women and 1.32, 1.46, 1.40, and 0.99 in men. In other words, average underweight, overweight, and obese women had, respectively, 16%, 8%, and 38% fewer children than women with normal weight. The corresponding percentages in men were 10%, 4%, and 32%. Individuals with BMI about 2 units above the mean (ie, around the cut-off point of the 75th percentile) were estimated to have 2–3% more children than individuals with mean BMI. Note that women had more children than men because women begin to have children at a younger age."

The survival analyses indicated that there was a non-linear association between adolescent BMI and the probability of having the first, second or third child at a given age (Table 3; Fig. 2). There were no linear or nonlinear interaction effects between BMI and time, indicating that the strength of the association between BMI and the probability of having children did not increase or decrease with time. Figure 2 shows the cumulative probability (ie, P = 1 – value of survival function) of having children plotted against age in women and men. Table 4 shows in numerical form the likelihoods of having the first, second, and third child by the age of 39 in women and men.

The partial correlation (controlling for age and sex) between Year-0 adolescent and Year-21 adulthood BMI was r=0.54. We tested whether adulthood BMI mediated the association between adolescent BMI and the number of children by entering both adolescent and adulthood BMI into the same model. The results differed by sex (sex × adulthood BMI interaction effect: linear IRR = 1.28, 95% confidence interval [95% CI] = 1.02–1.60; quadratic IRR = 0.996, 95% CI = 0.992–0.999). When both BMIs were in the model, only adolescent BMI was important in women, while only adulthood BMI was important in men (Table 2, Model 3). In other words, the influence of adolescent BMI was mediated by adulthood BMI in men but not in women. This result was the same when waist circumference rather than BMI was used as an indicator of adulthood body fatness (data not shown).

Next we assessed whether adolescent BMI predicted the likelihood of having ever lived with a partner (ie,
having been married or cohabiting), and the degree to which this accounted for the association between BMI and the number of children. Adolescent BMI predicted in a nonlinear fashion the likelihood of having ever lived with a partner in women (linear OR \(1.83, 95\% \text{ CI} 1.03–3.26\); quadratic OR \(0.986, 95\% \text{ CI} 0.973–0.999\)) and men (linear OR \(1.62, 95\% \text{ CI} 0.996–2.64\); quadratic OR \(0.989, 95\% \text{ CI} 0.978–0.999\); Fig. 1B) with the respective probability estimates of 0.84, 0.90, 0.90, and 0.81 in women and 0.79, 0.84, 0.84, and 0.74 in men for average members of the 4 BMI groups. Adolescent BMI was still associated with number of children even after controlling for partnership history (Table 2, Model 2).

When the Poisson regression models were fit including only participants who had ever lived together with a partner, the evidence of the BMI-fertility association was weaker, all \((n = 1130)\): linear IRR \(1.14, 95\% \text{ CI} 0.99–1.33\); quadratic IRR \(0.997, 95\% \text{ CI} 0.993–1.000\); Women \((n = 640)\): linear IRR \(1.15, 95\% \text{ CI} 0.94–1.42\); quadratic IRR \(0.997, 95\% \text{ CI} 0.992–1.001\); Men \((n = 490)\): linear IRR \(1.13, 95\% \text{ CI} 0.91–1.40\); quadratic IRR \(0.997, 95\% \text{ CI} 0.992–1.002\), although not absent. In this subsample, the predicted number of children in the 4 adolescent BMI groups was 1.77, 1.93, 1.82 and 1.46 in women, and 1.63, 1.69, 1.56, and 1.21 in men. In other words, underweight, overweight and obese women had approximately 8%, 5%, and 24% fewer children, respectively, than women with normal adolescent weight. In men the corresponding percentages were 4%, 8%, and 29%, respectively.

### TABLE 2. Association of Adolescent BMI With Number of Children in Adulthood: Three Adjusted* Poisson Regression Models \((n = 1298)\)

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRR (95% CI)</td>
<td>IRR (95% CI)</td>
<td>IRR (95% CI)</td>
</tr>
<tr>
<td><strong>All (n = 1298)</strong></td>
<td><strong>All (n = 1298)</strong></td>
<td><strong>All (n = 1298)</strong></td>
</tr>
<tr>
<td>BMI</td>
<td>1.28 (1.10–1.48)</td>
<td>1.18 (1.02–1.36)</td>
</tr>
<tr>
<td>BMI(^2)</td>
<td>0.994 (0.990–0.998)</td>
<td>0.996 (0.993–0.999)</td>
</tr>
<tr>
<td>Partnership</td>
<td>5.01 (3.92–6.40)</td>
<td>1.04 (0.95–1.14)</td>
</tr>
<tr>
<td>Adult BMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult BMI(^2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Women (n = 715)</strong></td>
<td><strong>Women (n = 715)</strong></td>
<td><strong>Women (n = 715)</strong></td>
</tr>
<tr>
<td>BMI</td>
<td>1.32 (1.07–1.63)</td>
<td>1.19 (0.97–1.46)</td>
</tr>
<tr>
<td>BMI(^2)</td>
<td>0.993 (0.988–0.998)</td>
<td>0.996 (0.991–1.001)</td>
</tr>
<tr>
<td>Partnership</td>
<td>4.72 (3.35–6.64)</td>
<td>0.92 (0.87–1.08)</td>
</tr>
<tr>
<td>Adult BMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult BMI(^2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Men (n = 583)</strong></td>
<td><strong>Men (n = 583)</strong></td>
<td><strong>Men (n = 583)</strong></td>
</tr>
<tr>
<td>BMI</td>
<td>1.24 (1.00–1.54)</td>
<td>1.16 (0.93–1.43)</td>
</tr>
<tr>
<td>BMI(^2)</td>
<td>0.995 (0.990–0.999)</td>
<td>0.996 (0.992–1.001)</td>
</tr>
<tr>
<td>Partnership</td>
<td>5.30 (3.74–7.52)</td>
<td>1.28 (1.05–1.55)</td>
</tr>
<tr>
<td>Adult BMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult BMI(^2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*All models were adjusted for age, education, residence, and sex.

**FIGURE 1.** (A) and (B) show the predicted number of children and the likelihood of having ever lived with a partner, respectively, plotted against adolescent BMI. In (C), the predicted number of children is plotted against the age at menarche (in women). In (A) and (B), the range of the x-axis is approximately from the median value of the underweight group to the median value of the obese group. The range of the x-axis in (C) is approximately ± 2 standard deviations.
Among women the partial correlation (controlling for age) between BMI and age at menarche was r = –0.27, indicating that heavier women had earlier menarche. There was a nonlinear association between age at menarche and the number of children, such that women with early or late menarche had more children than those in between (linear \( IRR = 0.98 \), 95% CI = 0.97–0.98; quadratic \( IRR = 1.010 \), 95% CI = 1.001–1.019; Fig. 1C). Controlling for age at menarche did not change the association between adolescent BMI and fertility (data not shown).

The association between adolescent skinfold thickness and adulthood number of children was in the same direction (all: linear \( IRR = 1.13 \), 95% CI = 0.98–1.31; quadratic \( IRR = 0.988 \), 95% CI = 0.973–1.004; women: linear \( IRR = 1.13 \), 95% CI = 0.92–1.39; quadratic \( IRR = 0.986 \), 95% CI = 0.960–1.013; men: linear \( IRR = 1.16 \), 95% CI = 0.92–1.48; quadratic \( IRR = 0.987 \), 95% CI = 0.963–1.012) as with BMI, although the confidence intervals of the parameter estimates included 1.00.

### DISCUSSION

Our findings suggest that underweight and obesity are associated with reduced reproduction in the general population. The participants of the present study were still in their reproductive years, so their current fertility may not have accurately represented their completed fertility. However, the survival analyses suggested that this may not have been a major limitation, because the reproductive differentials between BMI groups became stronger over time. Moreover, less than 5% of babies in Finland are born to women over 39 years of age,\(^21\) the corresponding percentage being somewhat higher for men.\(^22\) It is therefore unlikely that the observed fertility differences related to body weight would represent only transitory effects, although this should be confirmed with further follow-up of this cohort.

In women, underweight is known to cause menstrual dysfunction,\(^23\) and obesity has been associated with decreased fecundability and increased risk of pregnancy complications.\(^4\)–\(^6\) Obesity is also a risk factor for polycystic ovary syndrome, which causes infertility.\(^24\) In men, low (<20 kg/m\(^2\)) and high (>25 kg/m\(^2\)) adulthood BMI have been associated with decreased semen quality,\(^8\) and obesity has been shown to increase the risk of erectile dysfunction.\(^9\) These physiological correlates of body weight may contribute to the lower fertility in underweight and obese people.

The lowered fertility of underweight and obese individuals was partly accounted by their decreased likelihood of attaining a partner. We were not able to examine whether other factors related to mating (eg, the number of partners, the stability of relationships) might also be involved. The role of other psychologic and social variables should also be considered. Taller young women have been found to have lower maternal tendencies and decreased preference for having children.\(^25\) Whether a similar phenomenon is related to BMI is unclear\(^26\); the possible correlation between childbearing motivation and body weight merits further research. The extremes of body weight may also be related to psychiatric disorders.\(^27\)–\(^28\)

Research suggests that there is assortative mating (ie, nonrandom mating that results in similarity between spouses) for BMI in general,\(^29\) and obesity in particular.\(^30\) This may act to cumulate the influence of BMI on fertility in underweight and obese couples. Assortative mating also implies that the BMI-fertility association observed in one sex might be partly accounted by their spouse’s BMI, for which we did not have data. This possibility has received little attention in fertility...
TABLE 4. The Predicted Probability of Having the First, Second, and Third Child by the Age of 39 (ie, $P = 1 - \text{value of survivor function at 39}$) by BMI Group

<table>
<thead>
<tr>
<th></th>
<th>Underweight</th>
<th>Normal Weight</th>
<th>Overweight</th>
<th>Obese</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.50</td>
<td>0.57</td>
<td>0.55</td>
<td>0.39</td>
<td>0.50</td>
</tr>
<tr>
<td>First child</td>
<td>0.75</td>
<td>0.81</td>
<td>0.80</td>
<td>0.65</td>
<td>0.75</td>
</tr>
<tr>
<td>Second child</td>
<td>0.57</td>
<td>0.63</td>
<td>0.60</td>
<td>0.42</td>
<td>0.55</td>
</tr>
<tr>
<td>Third child</td>
<td>0.19</td>
<td>0.28</td>
<td>0.25</td>
<td>0.09</td>
<td>0.20</td>
</tr>
<tr>
<td>Women</td>
<td>0.54</td>
<td>0.63</td>
<td>0.58</td>
<td>0.38</td>
<td>0.53</td>
</tr>
<tr>
<td>First child</td>
<td>0.81</td>
<td>0.88</td>
<td>0.84</td>
<td>0.64</td>
<td>0.79</td>
</tr>
<tr>
<td>Second child</td>
<td>0.61</td>
<td>0.68</td>
<td>0.63</td>
<td>0.43</td>
<td>0.59</td>
</tr>
<tr>
<td>Third child</td>
<td>0.20</td>
<td>0.34</td>
<td>0.27</td>
<td>0.07</td>
<td>0.22</td>
</tr>
<tr>
<td>Men</td>
<td>0.45</td>
<td>0.51</td>
<td>0.51</td>
<td>0.40</td>
<td>0.47</td>
</tr>
<tr>
<td>First child</td>
<td>0.67</td>
<td>0.74</td>
<td>0.77</td>
<td>0.66</td>
<td>0.71</td>
</tr>
<tr>
<td>Second child</td>
<td>0.52</td>
<td>0.57</td>
<td>0.57</td>
<td>0.43</td>
<td>0.52</td>
</tr>
<tr>
<td>Third child</td>
<td>0.16</td>
<td>0.21</td>
<td>0.20</td>
<td>0.10</td>
<td>0.17</td>
</tr>
</tbody>
</table>

FIGURE 2. The cumulative probability ($P = 1 - \text{value of survivor function}$) of having the first, second, and third child plotted against age in women and men by body weight groups. Note the different scales of y-axis in each of the panels.
research. Sallmen et al\(^7\) found that men’s overweight decreased the likelihood of conception independently of their wives’ BMI. Further data on couples are needed to investigate the combined influence of partners’ body weight on the number of their children.

Adolescent BMI predicted fertility independently of adulthood BMI in women but not in men. This difference may be because in women the association between adulthood BMI and fertility is probably confounded by the influence of pregnancies on adulthood BMI.\(^12\) It may also be that women’s adolescent body weight has independent long-term implications for reproductive potential. This would parallel the finding that adolescent overweight may increase adulthood mortality independently of adulthood body weight.\(^2\) Indeed, adolescent overweight has been implicated as a risk factor for adulthood polycystic ovary syndrome\(^31\) and menstrual disturbances.\(^31–33\) Women with a history of eating disorders and severe underweight appear to suffer no major long-term consequences of infertility,\(^34–36\) although they are more likely to have obstetric complications.\(^27,34–36\) More detailed information on possible biologic and social mechanisms would help to evaluate the relative importance of adolescent and adulthood body weight on fertility.

The ongoing increases in obesity\(^3\) may represent a concern for future reproductive health.\(^57\) In Finnish adolescents the prevalence of obesity increased 2- to 3-fold from 1977 to 1999.\(^38\) Hence, the lowered fertility associated with high BMI may affect an increasing proportion of the population in generations younger than the present study participants, who were adolescents in the early 1980s.

Adolescent BMI is a highly heritable trait with an estimated heritability of about 85% in Finnish adolescent twins.\(^39\) Heritable traits associated with reproductive differentials have the potential to evolve via the mechanism of natural selection.\(^40\) The present results suggest that stabilizing selection may be acting on BMI, since the low and high ends of BMI are selected against in terms of reproductive success. On the other hand, individuals with a BMI about one-half standard deviations above the mean were estimated to have 2–3% more children than those with mean BMI. If this pattern were to hold over generations, genetic factors might be shifting the mean BMI of the population about 2 units upwards in future generations, albeit only at a modest rate. Reproductive differentials are unlikely to account for any substantial portion of the recent obesity epidemic.

The higher fertility of women with early or late menarche was an unexpected finding. Early-maturing girls tend to enter sexual and marital relations earlier,\(^17\) which might contribute to their greater number of children in adulthood. On the other hand, a study conducted in China\(^41\) found that late rather than early menarche was associated with increased fertility. The pattern of the present finding suggests that timing of menarche may be under disruptive natural selection where individuals in the extreme ends of a trait have a reproductive advantage over the others. It would be of interest to investigate further the developmental pathways linking early and late menarche to increased adulthood fertility.

### REFERENCES


Body Mass Index and Lung Cancer Risk in Women

Geoffrey C. Kabat,* Anthony B. Miller,† and Thomas E. Rohan*

Background: Studies have suggested that leanness in adulthood may be a risk factor for lung cancer; however, there is justifiable concern that the observed association may be due to residual confounding by smoking, preclinical weight loss, competing causes of death, or some combination of these.

Methods: To examine this association we used data from the Canadian National Breast Screening Study, which included 89,835 women ages 40–59 years at recruitment between 1980 and 1985. During a mean of 16 years of follow-up, we observed 750 incident lung cancer cases. We used Cox proportional hazards models to estimate hazard ratios and 95% confidence intervals for the association between body mass index (BMI) and lung cancer.

Results: After adjustment for pack-years of smoking and other covariates, there was some evidence for inverse associations in current smokers (hazard ratio for highest BMI quintile relative to the lowest = 0.63; 95% confidence interval = 0.48–0.83) and in former smokers (0.69; 0.39–1.23), whereas in never-smokers, BMI was positively associated with lung cancer (2.19; 1.00–4.80). The results for current and former smokers were not altered by exclusion of cases diagnosed within the first 5 years of follow-up; however, in never-smokers the strength of the association was reduced.

Conclusions: The present study contributes to the aggregate evidence suggesting that there may be an inverse association between BMI and lung cancer among smokers. However, the contrasting pattern of associations between BMI and lung cancer seen in ever-smokers and never-smokers in this study requires explanation.

(Epidemiology 2007;18: 607–612)

Excess body weight, as measured by body mass index (BMI; weight [kg] divided by height squared [m²]) is an important predictor of total mortality, cardiovascular disease mortality, and mortality from certain cancers.¹ A recent analysis of a large cohort of retirees indicated that even moderate excess body weight in middle age is associated with increased total mortality.² In apparent contrast to these deleterious effects of overweight and obesity, a number of epidemiologic studies, (both case-control and cohort designs) have reported an inverse association of body mass index with lung cancer risk, suggesting that leanness may be a risk factor for the development of lung cancer.³–¹³ However, other studies have either failed to find an association,¹⁴–¹⁶ or the association disappeared when smoking habits and health status were taken into account.¹⁷,¹⁸ One study in never-smokers and long-term ex-smokers reported a positive association between body mass index and lung cancer risk.¹⁹

The observed association between leanness and lung cancer risk may be explained by the fact that cigarette smokers tend to be leaner than nonsmokers; by weight loss due to pre-existing disease; by competing causes of death; or by some combination of these.¹⁸,²⁰ However, several studies have at least partially addressed these concerns.³,⁴,⁷–⁹,¹¹ Furthermore, although current smokers tend to weigh less than former smokers and never-smokers, BMI among current smokers tends to be higher in heavy smokers than in those smoking less.²¹–²⁸ Thus, it is unclear that the inverse association between BMI and lung cancer risk can be explained by smoking. In view of the conflicting evidence, studies capable of addressing sources of bias or confounding can help determine whether there is a reproducible association between BMI and lung cancer. If BMI is truly associated with lung cancer risk, a second step would be to explain the biologic basis of the association.

We report here on the association between BMI and risk of lung cancer in a large prospective cohort of women followed for an average of 16 years.

METHODS

Study Population

The study, which has been described in detail elsewhere,²⁹,³⁰ was conducted among participants in the Canadian National Breast Screening Study, a randomized controlled trial of screening for breast cancer. A total of 89,835 women ages 40–59 years with no history of breast cancer were recruited into the trial between 1980 and 1985. The study was approved by the appropriate Institutional Review Boards, and informed consent was obtained from all study participants.

Questionnaires

At recruitment into the study, participants completed a self-administered questionnaire that sought information on demographic characteristics, smoking history, menstrual and reproductive history, and use of oral contraceptives and replacement estrogens. Information about smoking habits included smoking status, number of cigarettes smoked per day, number of years of smoking, and, for former smokers, the year they had stopped smoking. Height and weight were

From the *Department of Epidemiology and Population Health, Albert Einstein College of Medicine, New York; and †Department of Public Health Sciences, University of Toronto, Toronto, Canada.
Correspondence: Geoffrey C. Kabat, Department of Epidemiology and Population Health, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Room 1301, NY 10461. E-mail: gkabat@acem.yu.edu.
Copyright © 2007 by Lippincott Williams & Wilkins
ISSN: 1044-3983/07/1805-0607
DOI: 10.1097/EDE.0b013e31812713d1

Epidemiology • Volume 18, Number 5, September 2007

607
measured at the initial physical examination. This information was available on the full cohort of 89,835 women. Starting in 1982 (that is, after some participants had completed their scheduled visits to the screening centers), a self-administered food-frequency questionnaire containing questions on 86 food items was distributed to all new attendees at all screening centers and to women returning to the screening centers for rescreening. This questionnaire also included questions on current weight, weight 1 year prior to enrollment, weight at age 20, height, alcohol consumption, and physical activity. A total of 49,654 women completed the food frequency questionnaire.

**Computed Variables**

BMI was defined as weight (kg)/height (m²), (either as measured or self-reported). Pack-years of smoking were computed by multiplying the number of cigarettes smoked per day by the total years of smoking, and dividing by 20. Change in weight was computed by subtracting weight at age 20 (ie, self-reported weight in the subgroup with dietary information) from weight measured at enrollment. Data from the food frequency questionnaire were also used to calculate daily total energy intake, intake of alcohol, and self-reported vigorous physical activity.

**Ascertainment of Incident Lung Cancer Cases and Deaths**

Cases were women who were diagnosed during follow-up with incident lung cancer (ICD-9 codes 162.0–162.9), ascertained by means of computerized record linkage to the Canadian Cancer Database. Deaths from all causes were ascertained by means of record linkage to the National Mortality Database. Both databases are maintained by Statistics Canada. The linkages to the databases yielded data on cancer incidence and mortality to 31 December 2000 for women in Ontario, to 31 December 1998 for women in Quebec, and to 31 December 1999 for women in other provinces in Canada. Of the 89,835 women recruited into the study, we excluded 23 lung cancer cases who lacked a date of diagnosis and date of enrollment (7 women) or whose date of diagnosis preceded their date of enrollment (16 women), as well as 24 cases of nonepithelial cancer of the lung. Among the remaining 89,788 women, we identified 750 incident lung cancers. In the subgroup of 49,654 women with dietary information, 42,444 noncases and 342 lung cancer cases had information on weight at age 20.

**Statistical Analysis**

We used Cox proportional hazards models (using age as the time scale) to estimate hazard ratios (HRs) and 95% confidence intervals (CIs) for the association between BMI and lung cancer risk. Study participants were considered at risk from their date of enrollment until the date of diagnosis of their lung cancer, termination of follow-up (the date to which cancer incidence data were available for women in the corresponding province) or death, whichever occurred earliest. Analyses were carried out both on the full cohort (with 750 cases) and on the subgroup for which data on weight at age 20 was available. BMI was categorized into quintiles. Additional models were fitted in the full cohort with BMI categorized into deciles and also as a continuous variable.

Because smoking is the dominant risk factor for lung cancer, all analyses were stratified by smoking status (current, former, and never-smoker). In current smokers, adjustment for number of cigarettes smoked per day, plus years of smoking and pack-years, yielded similar results; we present the data adjusted for pack-years. In former smokers, we adjusted for pack-years plus years since quitting smoking. Additional covariates in the multivariate models were education (3 levels), menopausal status (pre, peri-, postmenopausal), randomization group (intervention versus control), and study center. These models were repeated for subgroups defined by age at diagnosis (dichotomized at the median: <65.5/65.5+), menopausal status (premenopausal/postmenopausal), and for adenocarcinoma of the lung. Adenocarcinoma was the most common histologic subtype (accounting for 47% of cases with defined histology) followed by small cell carcinoma (16%), squamous cell carcinoma (13%), large cell carcinoma (7%), and mixed or unknown cell types (17%).

Analyses in the subgroup with dietary information additionally included terms for total caloric intake (continuous), vigorous physical activity (any/none), and alcohol consumption (4 levels). Change in weight from age 20 to enrollment was categorized as follows: minimal change (+5 pounds) as the reference group; −6 to +6 pounds, +6 to 14 pounds, +15 to 24 pounds, +25 to 49 pounds, and +50 pounds or more. To test for trends in the categorical variables of interest, study participants were assigned the median value of their category, and the resulting variable was fitted as a continuous variable in the regression models; the statistical significance of the coefficients was evaluated using the Wald test. Use of the Lifetest procedure in SAS (SAS Institute, Cary, NC) showed that the proportional hazards assumption was met. All significance tests were two-sided.

**RESULTS**

Table 1 presents the baseline characteristics of lung cancer cases and noncases. Cases were older than noncases, had a higher proportion of postmenopausal women, and had higher indices of cigarette smoking. Cases also had lower mean years of education and lower mean body mass index at enrollment. Body mass index at age 20 did not differ between the 2 groups.

After adjusting for age at enrollment, BMI at enrollment differed by smoking status and by case status (Table 2). Among noncases, current smokers as a group had a lower mean BMI compared with both never and former smokers (24.6 current smokers; 25.1 never-smokers; 25.1 former smokers). Among noncases who were current smokers, except for the lightest smoking category, there was a small increase in BMI with increasing number of cigarettes smoked per day. Among lung cancer cases, all analyses were stratified by smoking status (current, former, and never-smoker). In current smokers, adjustment for number of cigarettes smoked per day, plus years of smoking and pack-years, yielded similar results; we present the data adjusted for pack-years. In former smokers, we adjusted for pack-years plus years since quitting smoking. Additional covariates in the multivariate models were education (3 levels), menopausal status (pre, peri-, postmenopausal), randomization group (intervention versus control), and study center. These models were repeated for subgroups defined by age at diagnosis (dichotomized at the median: <65.5/65.5+), menopausal status (premenopausal/postmenopausal), and for adenocarcinoma of the lung. Adenocarcinoma was the most common histologic subtype (accounting for 47% of cases with defined histology) followed by small cell carcinoma (16%), squamous cell carcinoma (13%), large cell carcinoma (7%), and mixed or unknown cell types (17%).

Analyses in the subgroup with dietary information additionally included terms for total caloric intake (continuous), vigorous physical activity (any/none), and alcohol consumption (4 levels). Change in weight from age 20 to enrollment was categorized as follows: minimal change (+5 pounds) as the reference group; −6 to +6 pounds, +6 to 14 pounds, +15 to 24 pounds, +25 to 49 pounds, and +50 pounds or more. To test for trends in the categorical variables of interest, study participants were assigned the median value of their category, and the resulting variable was fitted as a continuous variable in the regression models; the statistical significance of the coefficients was evaluated using the Wald test. Use of the Lifetest procedure in SAS (SAS Institute, Cary, NC) showed that the proportional hazards assumption was met. All significance tests were two-sided.

**RESULTS**

Table 1 presents the baseline characteristics of lung cancer cases and noncases. Cases were older than noncases, had a higher proportion of postmenopausal women, and had higher indices of cigarette smoking. Cases also had lower mean years of education and lower mean body mass index at enrollment. Body mass index at age 20 did not differ between the 2 groups.

After adjusting for age at enrollment, BMI at enrollment differed by smoking status and by case status (Table 2). Among noncases, current smokers as a group had a lower mean BMI compared with both never and former smokers (24.6 current smokers; 25.1 never-smokers; 25.1 former smokers). Among noncases who were current smokers, except for the lightest smoking category, there was a small increase in BMI with increasing number of cigarettes smoked per day. Among lung cancer cases, all analyses were stratified by smoking status (current, former, and never-smoker). In current smokers, adjustment for number of cigarettes smoked per day, plus years of smoking and pack-years, yielded similar results; we present the data adjusted for pack-years. In former smokers, we adjusted for pack-years plus years since quitting smoking. Additional covariates in the multivariate models were education (3 levels), menopausal status (pre, peri-, postmenopausal), randomization group (intervention versus control), and study center. These models were repeated for subgroups defined by age at diagnosis (dichotomized at the median: <65.5/65.5+), menopausal status (premenopausal/postmenopausal), and for adenocarcinoma of the lung. Adenocarcinoma was the most common histologic subtype (accounting for 47% of cases with defined histology) followed by small cell carcinoma (16%), squamous cell carcinoma (13%), large cell carcinoma (7%), and mixed or unknown cell types (17%).

Analyses in the subgroup with dietary information additionally included terms for total caloric intake (continuous), vigorous physical activity (any/none), and alcohol consumption (4 levels). Change in weight from age 20 to enrollment was categorized as follows: minimal change (+5 pounds) as the reference group; −6 to +6 pounds, +6 to 14 pounds, +15 to 24 pounds, +25 to 49 pounds, and +50 pounds or more. To test for trends in the categorical variables of interest, study participants were assigned the median value of their category, and the resulting variable was fitted as a continuous variable in the regression models; the statistical significance of the coefficients was evaluated using the Wald test. Use of the Lifetest procedure in SAS (SAS Institute, Cary, NC) showed that the proportional hazards assumption was met. All significance tests were two-sided.
Table 1 presents multivariate-adjusted HRs and 95% CIs for the association of BMI with lung cancer stratified by smoking status (premenopausal/postmenopausal). When cases were restricted to those with adenocarcinoma, results were similar. We repeated the analyses after excluding the 132 lung cancers ascertained within the first 5 years of follow-up. Results were unchanged in all subjects, and in current and former smokers. In never-smokers, however, the HRs were reduced (HR for extreme quartiles = 1.45; 95% CI = 0.64–3.30, and the gradient with higher BMI was weakened (P for trend = 0.11). (The latter analysis was based on only 75 cases). Furthermore, when the follow-up period was partitioned into 3 segments (<8 years, 8–12 years, 13+ years), the inverse association in current smokers was apparent in all 3 time periods. In ex-smokers, the inverse association was apparent particularly in the second and third time periods. In never-smokers, the positive association was seen mainly in the first and third time periods (data not shown).

Among both cases and noncases, weight increased between age 20 and age at enrollment, with a larger increase in noncases (mean 15.9 pounds compared with 14.6 pounds in cases; P < 0.0001). Among current smokers, women who lost more than 5 pounds were not at increased risk compared with women who experienced minimal change in weight (HR = 1.15; 95% CI = 0.72–1.82; Table 4). Among women who gained weight there was a trend toward decreasing risk with increasing level of weight gain (P = 0.004). For women who gained 50+ pounds the HR was 0.50 (95% CI = 0.22–1.12). The effect of change in weight could not be assessed in former and never-smokers due to the small numbers of cases (62 and 45, respectively).

**DISCUSSION**

In the present study BMI was inversely associated with lung cancer risk among current and former smokers, whereas in never-smokers it was positively associated with risk. Adjustment for pack-years of smoking (or for duration and intensity of smoking) did not weaken the observed associations in current and former smokers. The inverse association of BMI with lung cancer in current and former smokers was unaffected after excluding lung cancer cases diagnosed in the first 5 years of follow-up, thereby minimizing the likelihood that the observed findings are due to weight loss caused by preclinical disease at enrollment. In addition, among current smokers weight gain between age 20 and age at enrollment was associated with reduced risk.

Our results in smokers are consistent with the findings from several cohort and case-control studies. Studies that analyzed the association of BMI with lung cancer in never-smokers have shown conflicting results. Three case-control studies and 3 cohort studies provide evidence...
TABLE 3. Multivariate-Adjusted Associations of BMI (kg/m²) and Risk of Lung Cancer, by Smoking Status

<table>
<thead>
<tr>
<th>BMI</th>
<th>Current Smokers (n = 520)</th>
<th>Ex-Smokers (n = 123)</th>
<th>Never-Smokers (n = 98)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR (95% CI)</td>
<td>HR (95% CI)</td>
<td>HR (95% CI)</td>
</tr>
<tr>
<td>No adjustment for smoking*</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>≤21.6†</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>21.6–23.3</td>
<td>0.75 (0.59–0.97)</td>
<td>1.12 (0.66–1.91)</td>
<td>0.91 (0.36–2.31)</td>
</tr>
<tr>
<td>23.3–25.1</td>
<td>0.77 (0.59–0.99)</td>
<td>0.75 (0.42–1.33)</td>
<td>1.97 (0.88–4.41)</td>
</tr>
<tr>
<td>25.1–27.9</td>
<td>0.62 (0.47–0.82)</td>
<td>0.56 (0.30–1.05)</td>
<td>1.81 (0.80–4.06)</td>
</tr>
<tr>
<td>27.9+</td>
<td>0.68 (0.52–0.89)</td>
<td>0.87 (0.49–1.53)</td>
<td>2.19 (1.00–4.80)</td>
</tr>
<tr>
<td>P for trend</td>
<td></td>
<td>0.001</td>
<td>0.14</td>
</tr>
<tr>
<td>With additional adjustment for smoking†</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>≤21.6†</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>21.6–23.3</td>
<td>0.77 (0.60–0.99)</td>
<td>1.06 (0.63–1.80)</td>
<td></td>
</tr>
<tr>
<td>23.3–25.1</td>
<td>0.77 (0.60–1.00)</td>
<td>0.65 (0.36–1.16)</td>
<td></td>
</tr>
<tr>
<td>25.1–27.9</td>
<td>0.64 (0.49–0.85)</td>
<td>0.47 (0.25–0.88)</td>
<td></td>
</tr>
<tr>
<td>27.9+</td>
<td>0.63 (0.48–0.83)</td>
<td>0.69 (0.39–1.23)</td>
<td></td>
</tr>
<tr>
<td>P for trend</td>
<td></td>
<td>0.0003</td>
<td>0.02</td>
</tr>
</tbody>
</table>

*Adjusted for age (as time-scale), education (3 levels), and menopausal status (pre-, peri-, postmenopausal).
†Reference category.
‡HRs for current smokers are additionally adjusted for pack-years of smoking as a continuous variable. HRs for former smokers are additionally adjusted for pack-years and years since quitting.

In smokers, the inverse association of BMI with lung cancer risk could be explained by a number of potential biases. First, if heavier smokers tended to have lower BMI, this could lead to a spurious association. Second, the observed association could be due to preclinical weight loss prior to a diagnosis of lung cancer. Finally, the association could be due to competing mortality from diseases for which obesity is a risk factor. We address each of these points below.

Although current smokers tend to be leaner than former and never-smokers, when attention is restricted to current smokers, those who smoke more tend to have higher BMIs than moderate smokers.21–28 This was true for noncases in our study, although not necessarily for cases. Among noncases who were smokers at baseline, BMI increased modestly over increasing strata of cigarettes smoked per day (Table 2). Among cases, BMI was essentially unchanged over categories of cigarettes per day above the category of 1–10 cigarettes. After adjustment for pack-years of smoking, the association of BMI with lung cancer among current smokers was slightly strengthened rather than attenuated.

Aside from a previous history of breast cancer, no information was available regarding pre-existing illness. However, when lung cancer cases diagnosed during the first 5 years of follow-up were excluded, the association of BMI with lung cancer in current and former smokers was unchanged. In addition, the association of BMI with lung cancer in current smokers was similar when the follow-up period was partitioned into 3 segments. This suggests that the association cannot be explained by preclinical disease prior to baseline. Several other studies have presented similar findings.7,11

Regarding the possibility that competing causes of death could be responsible for the observed association, it

dence of an inverse association. In contrast, a large cohort study16 found no association of BMI with lung cancer, and a fourth case-control study17 found a 2.6-fold increased odds ratio for the highest octile of BMI versus the lowest octile in never and long-term former smokers. Our findings in never-smokers are consistent with those of Rauscher et al.19; however, our results were attenuated when cases diagnosed within the first 5 years of follow-up were excluded, suggesting that weight gain close to the time of diagnosis in never-smokers may be responsible for the observed positive association.

TABLE 4. Multivariate-Adjusted Association of Change in Weight since Age 20 and Risk of Lung Cancer Among Current Smokers

<table>
<thead>
<tr>
<th>Weight Change</th>
<th>No. Cases/Person-Years</th>
<th>HR† (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight loss &gt;5 lbs</td>
<td>38/23,557</td>
<td>1.15 (0.72–1.82)</td>
</tr>
<tr>
<td>No/minimal change: ±5 lbs†</td>
<td>34/16,687</td>
<td>1.00</td>
</tr>
<tr>
<td>Weight gain: &gt;5–&lt;15 lbs</td>
<td>39/26,774</td>
<td>0.80 (0.51–1.26)</td>
</tr>
<tr>
<td>15–&lt;25 lbs</td>
<td>39/28,406</td>
<td>0.73 (0.46–1.14)</td>
</tr>
<tr>
<td>≥25–&lt;50 lbs</td>
<td>47/30,007</td>
<td>0.74 (0.48–1.14)</td>
</tr>
<tr>
<td>≥50 lbs</td>
<td>8/6,708</td>
<td>0.50 (0.22–1.12)</td>
</tr>
<tr>
<td>P for trend</td>
<td></td>
<td>0.004</td>
</tr>
</tbody>
</table>

*No. cases among current smokers in the dietary subcohort = 209.
†Adjusted for parity (parous/nulliparous), menopausal status (pre-, peri-, postmenopausal), education (3 categories), pack-years of smoking (continuous), total caloric intake (continuous), vigorous physical activity (any, none), alcohol consumption (4 levels), study center, and randomization group (intervention vs. control).
‡Reference category.
should be noted that survival analysis partially takes into account competing causes of death. In the present study the inverse association between BMI and lung cancer was actually stronger in women ages 40–49 at enrollment than in women ages 50–59, suggesting that the association is not an artifact due to early loss of women due to obesity-related conditions.

The paper by Henley et al\textsuperscript{18} is of particular interest because of the large size of its study population. The authors examined the association of BMI with subsequent lung cancer mortality among 941,105 men and women enrolled in the American Cancer Society’s Cancer Prevention Study II and followed for 14 years. To assess bias, they successively eliminated groups in which the association might be confounded. Henley and colleagues concluded that in lifetime nonsmokers who did not report pre-existing disease, leanness was not associated with lung cancer mortality. However, they did not discuss their results in smokers, who accounted for 98\% of the lung cancers cases in men and 93\% in women. Their data on all cases after exclusion of persons with pre-existing disease and after exclusion of the first 5 years of follow-up tend to support the existence of a modest increased risk associated with low BMI. In men, for those with BMI < 20 kg/m\textsuperscript{2} the hazard ratio was 1.24 (95\% CI = 1.05–1.46) after exclusion of those with pre-existing disease and 1.23 (1.01–1.50) after the additional exclusion of the first 5 years of follow-up. The corresponding hazard ratios in women with BMI < 19 were 1.32 (1.13–1.55) and 1.28 (1.07–1.53). In women there was suggestive evidence a trend toward increasing risk with decreasing BMI. It is noteworthy that, despite their large study population, Henley et al only categorized number of cigarettes smoked per day as < 20, 20, and > 20. It is possible that finer adjustment of intensity of smoking would have strengthened the inverse association with BMI.

Few studies have examined weight gain/loss since early adulthood relative to lung cancer risk. In a cohort study from Hawaii, Nomura et al\textsuperscript{9} found that weight loss since age 25 was associated with increased risk of lung cancer, but found no association between BMI at entry and lung cancer risk. Goodman and Wilkens\textsuperscript{9} reported in a case-control study that weight gain during adulthood was associated with reduced risk. Both cases and controls tended to gain weight during adulthood, but cases tended to gain less weight, which is very similar to our findings.

Among the strengths of the present study are its prospective nature, the high level of follow-up, and the availability of weight and height measured at enrollment. However, a number of limitations also need to be borne in mind. First, as already mentioned, we did not have information on general health status or history of chronic diseases at enrollment, with the exception of breast cancer and benign breast disease. Second, information on smoking habits was limited to that obtained at baseline. Thus, even though all smoking variables in this study were strong predictors of subsequent lung cancer risk, some misclassification is likely due to changes in smoking habits after enrollment. Furthermore, we did not have information on more detailed aspects of smoking (depth of inhalation, puff frequency, interpuff interval, tar and nicotine content of the cigarette smoked, and mentholation), which may be associated with body mass index and lung cancer risk. Thus, in spite of our adjustment for smoking parameters, there could still be residual confounding of the association of BMI with lung cancer due to smoking.

Two recent studies suggest a possible biologic mechanism underlying the observed inverse association of BMI with lung cancer. In one small study,\textsuperscript{32} the level of DNA adducts in peripheral lymphocytes was inversely related to BMI level after accounting for the number of cigarettes smoked per day. In another study,\textsuperscript{33} there was a strong inverse association between BMI and levels of 8-hydroxydeoxyguanosine, a marker of oxidative DNA damage, whereas number of cigarettes smoked per day was not associated with BMI. However, it is unclear from these studies whether leanness itself or some factor correlated with leanness is an independent risk factor for lung cancer, or perhaps modifies the carcinogenic effect of smoking.

In conclusion, the present study contributes to the aggregate evidence suggesting an inverse association between body mass index and lung cancer among smokers. However, the contrasting patterns in ever-smokers and never-smokers, as well as the conflicting results in previous studies of never-smokers, requires explanation. In view of the potential for uncontrolled confounding in most studies to date, the association of body mass index with lung cancer merits further study, particularly in large cohorts with serial measurements of weight and height at regular intervals over the follow-up period, and with updated information on smoking habits and health status.

ACKNOWLEDGMENTS

We thank Statistics Canada, the provincial and territorial Registrars of Vital Statistics, and the Cancer Registry directors for their assistance in making the cancer incidence and mortality data available.

REFERENCES


Pet Ownership and Blood Pressure in Old Age

Joel David Wright,*† Donna Kritz-Silverstein,* Deborah J. Morton,* Deborah L. Wingard,* and Elizabeth Barrett-Connor*

Background: It has been proposed that pet ownership improves cardiovascular health. This study examines the relation of pet ownership with systolic and diastolic blood pressure, pulse pressure, mean arterial pressure, and hypertension in a large sample of older men and women.

Methods: Participants were 1179 community-dwelling men (n = 498) and women (n = 681) age 50–95 years. Participants responded to a 1991–1992 mailed questionnaire ascertaining pet ownership, and they attended a 1992–1996 clinic visit at which systolic (SBP) and diastolic (DBP) blood pressures were measured and use of antihypertensive medication was validated. Pulse pressure was calculated as SBP minus DBP. Mean arterial pressure was calculated as (SBP + DBP)/2. Body mass index, waist-hip ratio, and information on other potential confounders were obtained.

Results: Average age of participants was 70.4 ± 10.8 years; 30.0% reported current pet ownership. Mean SBP was 137.5 ± 21.4 mm Hg, and DBP was 76.1 ± 9.3 mm Hg; 55.6% were hypertensive (SBP ≥ 140, DBP ≥ 90 or taking hypertension medication). Pet owners were younger and slightly more overweight and they exercised less than nonowners; owners were somewhat more likely to have diabetes and to use beta-blockers. In unadjusted analyses, pet owners had lower SBP, pulse pressure, and mean arterial pressure, and a reduced risk of hypertension (odds ratio = 0.62; 95% confidence interval = 0.49–0.80). However, after adjustment for age and other confounders, pet ownership was not associated with systolic or diastolic blood pressure, pulse pressure, mean arterial pressure or risk of hypertension.

Conclusions: Results suggest that pet ownership is not independently associated with blood pressure, vascular reactivity, or hypertension.

(ClinicalTrials.gov number: NCT01652765.)

Cardiovascular disease is the leading cause of death in the United States.† Results of research concerning the effect of pet ownership on cardiovascular health have been inconsistent.‡,§ A clinical trial of 48 hypertensive stockbrokers from New York found that pet ownership was more effective than angiotensin-converting enzyme (ACE) inhibitors in blocking the increase of blood pressure (BP) with mental stress.¶ Similarly, a clinical trial in 240 married couples found that those with pets had significantly lower BP levels, less BP reactivity, and a quicker recovery from stressors.‖ Another clinical trial in 23 healthy children, aged 3–6 years, reported that BP was reduced in the presence of a dog during a doctor’s office visit.¶ Two other small intervention trials of pet snakes showed mixed results.⁰ Although 3 of these studies reported significant BP benefits, the other 2 did not, and they were relatively small studies that did not include older adults, and most were of dog and cat owners only.

In contrast, a cross-sectional study of 5079 Australian men and women aged 40–64 found no evidence that pet (dog/cat/other) ownership (of any type of animal) was associated with cardiac risk factors; in fact, participants with pets had higher diastolic BP than those without pets, even after adjusting for confounders. Editorials and letters have since debated these results.⁵,⁶ Although some argue that pet ownership benefits cardiovascular health, others state that there is no clear evidence that this is the case.⁷ There have been no other large population-based studies examining the association of pet ownership and BP in older men and women. We examined the association between pet ownership and BP before and after adjustment for confounders in a large population-based study of older men and women.

METHODS

Participants

Between 1972 and 1974, 82% of all adults age 30 to 79 residing in the southern California community of Rancho Bernardo were enrolled in a study of heart disease risk factors. These 6339 individuals have been followed with periodic clinic visits and yearly mailed questionnaires; vital status is also obtained annually. Participants are white, middle class, and relatively well-educated.

Between 1992 and 1996, members of this cohort age 50 years and older were invited to attend a follow-up clinic visit that included BP measurements; 1782 participated. Of these individuals, 1322 had responded to a mailed survey in 1991–1992 on pet ownership. After eliminating 29 individuals for lack of data on pet ownership and 7 whose BP was not measured, there were 1179 (498 men and 681 women) age 50

Submitted 26 July 2006; accepted 25 May 2007.

From the *University of California, San Diego, Department of Family and Preventive Medicine, La Jolla, California and †San Diego State University, Graduate School of Public Health, Department of Epidemiology and Biostatistics, San Diego, California. Supported by NIH grant no. DK31801 from the National Institute of Diabetes and Digestive and Kidney Disorders and grant no. AG02181 from the National Institute of Aging.

Correspondence: Donna Kritz-Silverstein, University of California, San Diego, Department of Family and Preventive Medicine, 9500 Gilman Drive, 0631-C, La Jolla, CA 92093-0631. E-mail: dsilverstein@ucsd.edu

Copyright © 2007 by Lippincott Williams & Wilkins

ISSN: 1044-3983/07/1805-0613

DOI: 10.1097/EDE.0b013e3181271398

Epidemiology • Volume 18, Number 5, September 2007

613
and older who provided data for this analysis. This study was approved by the University of California, San Diego (UCSD) Human Subjects Protections Program; all participants were ambulatory, and written informed consent was obtained prior to participation.

**Measurements**

During the 1992–1996 clinic visit, 2 BP measurements were obtained using a standard mercury sphygmomanometer, after the participant had been seated quietly for at least 5 minutes, by a specially trained nurse using the Hypertension Detection and Follow-up Program protocol. A 30-second pulse rate was measured at the wrist.

Weight and height were assessed with the participants wearing light clothing and no shoes; body-mass index (BMI; kg/m²) was calculated as an estimate of obesity. Percent body fat was estimated using bioelectrical impedance (model 1990B, Valhalla Scientific Inc., San Diego, CA). Waist and hip girths were measured, and waist-hip ratio was calculated [(waist (cm)/hip (cm)) × 100]. A blood sample was obtained in the morning after a requested 12- to 16-hour overnight fast; total cholesterol was measured in a Centers for Disease Control-certified Lipid Research Clinic laboratory using enzymatic techniques and an ABA-200 biochromatic analyzer (Abbott Laboratories, Irving, TX).

Participants answered standardized questions on current exercise 3 or more times per week (no/yes), the frequency of alcohol consumption 3 or more times per week (no/yes), and cigarette smoking history (never/past/current). Participants also reported whether they had ever been diagnosed by a physician as having diabetes or hypertension. Current medication use (including use of betablockers, thiazides, and other antihypertensives) was queried and validated by a nurse who examined pills and prescriptions brought to the clinic for that purpose.

On the mailed questionnaire in 1991–1992, participants were asked about lifetime and current pet ownership, including the kinds of indoor pets they had ever owned (dogs, cats, birds, hamsters/gerbils, or others). Pet ownership was defined as current ownership of a pet. Lifetime pet ownership was defined as ever owning an indoor pet.

**Statistical Analysis**

Means of the 2 systolic BP measurements (SBP) and the 2 diastolic BP (DBP) measurements were used to analyze BP as a continuous variable. In categorical analyses, individuals were categorized as hypertensive if their mean SBP was ≥140 mm Hg, their mean DBP was ≥90 mm Hg, or they reported current use of antihypertensive medication. Because of changes in recent guidelines for hypertension, a second analysis was performed in which individuals were categorized as hypertensive if they had mean SBP ≥130 mm Hg or mean DBP ≥85 mm Hg, or they reported current use of antihypertensive medication. Pulse pressure was calculated as SBP minus DBP. Mean arterial pressure was calculated as (SBP + DBP)/2.

Participants were dichotomized on pet ownership based on whether they currently owned a pet. Because it was thought that current pet ownership would have a greater effect on current BP than lifetime pet ownership, current ownership was used in most analyses. However, analyses were also performed after categorizing participants based on lifetime pet ownership and type of pet owned. Participants were categorized as lifetime dog owners if they owned only a dog or if they owned a dog and any other pet; as lifetime cat owners if they owned only a cat or a cat and any pet other than a dog; and as lifetime other pet owners if they owned only pets other than a dog or cat.

Comparisons of descriptive statistics by pet ownership controlling for age were performed using analysis of covariance (ANCOVA). Multiple linear regression analysis was used to examine the associations of pet ownership with SBP, DBP, pulse pressure and mean arterial pressure after adjustment for potential confounders. Multiple logistic regression analysis was used to determine whether pet ownership was associated with risk of categorically defined hypertension after adjustment for potential confounders. Analyses were also performed stratified by age, sex, pet type (dog owners/other pet owners/nonowners) and time when pets were owned (current/past/never). Multiple logistic regression analysis was used to determine survival bias by examining associations between pet ownership and vital status adjusted for potential confounders. Statistical analyses were performed with SAS statistical software (version 8.1, SAS Institute, Cary, NC).

**RESULTS**

Of these 1179 participants, 42% (n = 498) were men and 57% (n = 681) were women. Age ranged from 50.2 to 95.4 years with mean ± SD of 70.4 ± 10.8 years. Overall, 30% (n = 354) reported current pet ownership and 80% (n = 947) reported ever owning a pet. Mean (±SD) SBP was 137.5 (±21.4) mm Hg and DBP was 76.1 (±9.3) mm Hg. Using a categorical definition based on SBP ≥140, DBP ≥90, or use of antihypertensive medication, 55% were hypertensive; based on SBP ≥130, DBP ≥85, or use of antihypertensive medication, 71% were defined as hypertensive. Among the 947 participants who ever owned a pet, 80% reported ever owning a dog, 38% a cat, 20% a bird, 8% a hamster or gerbil, and 5% another type of pet.

As shown in Table 1, current pet owners were younger (mean = 64 years) than those who did not own pets (mean = 73.0 years). Age-adjusted comparisons of other characteristics by current pet ownership are also shown in Table 1. Current pet owners were somewhat more likely to be overweight (defined as BMI ≥25.0) than those who did not own pets (58% vs. 46% although mean BMI was similar between groups (mean = 25.4 and 25.7, respectively). Current pet owners were also less likely than nonowners to exercise 3 or more times per week (66% compared with 73%), slightly more likely to report diabetes (7% compared with 6%), and slightly more likely to take beta-blockers (23% compared with 22%). Those who currently owned pets did not differ from nonowners by sex, percent body fat, smoking status, alcohol consumption, physical health status, or other BP medication use. Results were similar in comparisons of those who were lifetime ever versus never pet owners (data not shown).
Table 2 shows unadjusted, age-adjusted, and multivariate-adjusted comparisons of mean BP by current pet ownership. Based on the unadjusted analyses, current pet owners had lower SBP than nonowners (mean = 132.8 mm Hg [130.5–135.0] vs. 139.5 mm Hg [138.1–141.0]), lower pulse pressure (mean = 55.5 mm Hg [53.6–57.5] vs. 63.9 mm Hg [62.6–65.1]), lower mean arterial pressure (mean = 105.0 mm Hg [103.6–106.4] vs. 107.6 mm Hg [106.7–108.5]), but somewhat higher DBP (mean = 77.2 mm Hg [76.3–78.2] vs. 75.7 mm Hg [75.0–76.3]), compared with those who did not own pets. However, these differences were attenuated after adjustment for age, or for age, sex, BMI, antihypertensive medication use, exercise, and diabetes diagnosis. Stratification by sex, by pet type (dog owners/other pet owners/nonowners), or by time when pets were owned (current/past/never) yielded similarly minor differences in SBP, DBP, PP, and MAP (data not shown). In addition, similar results were obtained when comparisons were made between those who had ever versus never owned a pet and when age and BMI interactions were considered (data not shown).

Table 3 shows the risk of hypertension in current pet owners as compared with nonowners overall and in sex-specific models after adjustment for age, or for age and other covariates. Before adjustment for age, the risk of hypertension among all pet owners (men and women combined) was lower than for nonowners, whether using the more stringent criteria (OR = 0.62; 95% CI = 0.49–0.80) or the less stringent criteria (0.55; 0.43–0.72). However, after adjustment for age, or for age and the other covariates, there were no substantial differences by pet ownership in risk of hypertension. Similar results were obtained when stratified by sex, time when pets were owned (current/past/never) or when lifetime pet ownership (ever/never) was used instead of current pet ownership (data not shown). Similar results were obtained when BP data collected at a 1988–1991 clinic visit was used instead of data from 1992–1996.

Results of analyses examining the association of type of pet ever owned with BP, pulse pressure, and mean arterial pressure are presented in Table 4. In unadjusted analyses, participants who had ever owned a dog had a lower SBP (mean = 136.2 mm Hg [134.7–137.7] vs. 141.9 mm Hg [139.3–144.6]) and pulse pressure (mean = 59.9 [58.6–61.2] vs. 66.3 [63.9–68.6]) than those who never owned pets. There were no other pairwise differences in SBP, DBP, pulse pressure, or mean arterial pressure by type of pet ever owned (dogs vs. cats vs. other vs. none). There similarly were no important pairwise differences after adjustment for age alone or for age, sex, exercise, self-reported physical health, beta-blocker use, other antihypertensive use, exercise, and diabetes diagnosis.

Overall, 415 (35.2%) of this sample died between the 1991–1992 questionnaire and 2004. In an age-adjusted analysis of pet ownership from 1991–1992 with vital status through 2004 (data not shown), those who did not own pets were more likely to have died than those who did own pets (OR = 2.09; 95% CI = 1.61–2.72).
DISCUSSION

Millions of Americans own pets. Results from some studies have suggested that pets can reduce BP.2,3,5 Plausible mechanisms have been suggested, including reduced stress, companion support, and walking a pet (usually a dog) for exercise. In the present study, current pet ownership was associated with both lower BP and hypertension in unadjusted analyses, but this association was explained by age, and not modified by further adjustment for covariates. Similarly, unadjusted analyses suggested that dog ownership conferred some benefit with regard to systolic BP and pulse pressure. However, after adjustment for age, there were no important differences between dog owners and nonowners on these variables. These results show the importance of adjustment for confounders, especially age, when examining BP and related outcomes. It is of interest that some, though not all, previous studies that suggested a reduction in BP with pets ownership2,5 did not adjust for confounders.

The results of this study are in accord with those of a randomized study in 72 healthy, middle-aged Australian men.

### TABLE 3. Unadjusted, Age-Adjusted and Multivariate-Adjusted* Risk of Hypertension in Current Pet Owners as Compared With Nonowners

<table>
<thead>
<tr>
<th>Variable</th>
<th>All (n = 354)</th>
<th>Men (n = 160)</th>
<th>Women (n = 194)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension (≥140/90)†</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.62 (0.49–0.80)</td>
<td>0.66 (0.45–0.96)</td>
<td>0.60 (0.43–0.84)</td>
</tr>
<tr>
<td>Age-adjusted</td>
<td>1.10 (0.83–1.46)</td>
<td>1.05 (0.69–1.59)</td>
<td>1.09 (0.74–1.60)</td>
</tr>
<tr>
<td>Multivariate-adjusted</td>
<td>0.91 (0.61–1.34)</td>
<td>0.74 (0.39–1.40)</td>
<td>1.00 (0.61–1.63)</td>
</tr>
<tr>
<td>Hypertension (≥130/85)‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.55 (0.43–0.72)</td>
<td>0.54 (0.36–0.81)</td>
<td>0.57 (0.40–0.81)</td>
</tr>
<tr>
<td>Age-adjusted</td>
<td>0.94 (0.70–1.27)</td>
<td>0.79 (0.51–1.22)</td>
<td>1.13 (0.75–1.70)</td>
</tr>
<tr>
<td>Multivariate-adjusted</td>
<td>0.81 (0.57–1.14)</td>
<td>0.69 (0.42–1.12)</td>
<td>1.09 (0.69–1.73)</td>
</tr>
</tbody>
</table>

For all comparisons, reference group is nonowners. Adjusted comparisons performed with multiple logistic regression.

*Adjusted for age, sex, BMI, beta blocker use, exercise and other antihypertensive use.
†SBP ≥ 140 mm Hg, DBP ≥ 90 mm Hg or taking hypertension medication.
‡SBP ≥ 130 mm Hg, DBP ≥ 85 mm Hg or taking hypertension medication.

### TABLE 4. Unadjusted, Age-Adjusted, and Multivariate-Adjusted* Comparisons of Blood Pressure, Pulse Pressure, and Mean Arterial Pressure by Lifetime Pet Ownership

<table>
<thead>
<tr>
<th>Variable</th>
<th>Dogs* (n = 780)</th>
<th>Cats* (n = 133)</th>
<th>Other (n = 19)</th>
<th>No Pets (n = 247)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBP (mm Hg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>136.2 (134.7–137.7)</td>
<td>137.1 (133.5–140.8)</td>
<td>136.6 (127.0–146.2)</td>
<td>141.9 (139.3–144.6)</td>
</tr>
<tr>
<td>Age-adjusted</td>
<td>137.0 (135.6–138.4)</td>
<td>138.7 (135.3–142.1)</td>
<td>138.0 (129.1–146.9)</td>
<td>138.5 (136.0–141.0)</td>
</tr>
<tr>
<td>Multivariate-adjusted†</td>
<td>142.5 (140.6–144.4)</td>
<td>143.8 (140.3–147.3)</td>
<td>142.6 (134.0–151.2)</td>
<td>144.3 (141.4–147.1)</td>
</tr>
<tr>
<td>DBP (mm Hg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>76.3 (75.6–77.0)</td>
<td>76.2 (74.6–77.8)</td>
<td>75.6 (71.4–79.8)</td>
<td>75.7 (74.5–76.8)</td>
</tr>
<tr>
<td>Age-adjusted</td>
<td>76.1 (75.4–76.7)</td>
<td>75.9 (74.3–77.4)</td>
<td>75.3 (71.2–79.4)</td>
<td>76.5 (75.3–77.7)</td>
</tr>
<tr>
<td>Multivariate-adjusted†</td>
<td>77.4 (76.7–78.3)</td>
<td>77.4 (75.7–79.0)</td>
<td>76.7 (72.7–80.8)</td>
<td>78.1 (76.8–79.4)</td>
</tr>
<tr>
<td>Pulse pressure (mm Hg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>59.9 (58.6–61.2)</td>
<td>60.9 (57.7–64.2)</td>
<td>60.9 (52.4–69.5)</td>
<td>66.3 (63.9–68.6)</td>
</tr>
<tr>
<td>Age-adjusted</td>
<td>60.9 (59.7–62.1)</td>
<td>62.8 (60.0–65.6)</td>
<td>62.7 (55.3–70.1)</td>
<td>61.9 (59.8–64.0)</td>
</tr>
<tr>
<td>Multivariate-adjusted†</td>
<td>66.9 (64.8–69.1)</td>
<td>68.3 (65.0–71.5)</td>
<td>67.9 (60.7–75.2)</td>
<td>68.0 (65.2–70.7)</td>
</tr>
<tr>
<td>Mean arterial pressure (mm Hg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>106.2 (105.3–107.2)</td>
<td>106.7 (104.4–108.9)</td>
<td>106.1 (100.1–112.1)</td>
<td>108.8 (107.1–110.5)</td>
</tr>
<tr>
<td>Age-adjusted</td>
<td>106.5 (105.6–107.4)</td>
<td>107.3 (105.1–109.5)</td>
<td>106.6 (100.8–112.5)</td>
<td>107.5 (105.9–109.2)</td>
</tr>
<tr>
<td>Multivariate-adjusted†</td>
<td>110.2 (109.0–111.4)</td>
<td>110.8 (108.5–113.1)</td>
<td>109.9 (104.2–115.6)</td>
<td>111.4 (109.6–113.3)</td>
</tr>
</tbody>
</table>

For all comparisons, the reference group is non-owners. Adjusted comparisons performed with analysis of covariance.

* “Dogs” includes ownership of dogs and other pets; “Cats” includes cats and pets other than dogs.
†Adjusted for age, sex, BMI, beta blocker or other antihypertensive use, exercise, and diabetes.
and women, which showed that the presence of a dog had no effect on BP during a task inducing mental stress.\textsuperscript{11} Mental stress increased BP in the absence of a dog, but BP was not affected by the presence of a dog. A large cross-sectional study performed with more than 5000 middle-aged men and women drawn from the electoral roll in Australia\textsuperscript{4} also has reported that pet ownership did not improve BP. In that study, current pet owners had higher DBP than those who did not currently own pets after adjustment for age, BMI, alcohol consumption, smoking status, physical activity, diabetes diagnosis, education, and socioeconomic status.

The results of the present study are in contrast to those of a number of small studies.\textsuperscript{2,3,5} For example, Allen and colleagues\textsuperscript{2} found that pet ownership reduced BP increases in response to mental stress in 48 hypertensive stockbrokers, but only persons who agreed to acquire a pet were included in the study, possibly biasing the results. The same research group found that among 240 healthy, middle-aged married couples, those who owned pets had lower BP levels, greater reactivity, and a quicker recovery from stressors.\textsuperscript{3} These results persisted even after adjustment for age, sex, BMI, education, income, presence of children in the home, mental health, or interpersonal support, although the data were not adjusted for physical activity.

Several sources of confounding in the present study were considered. Obesity is associated with high BP\textsuperscript{12} but is unlikely to explain the similarity of pet owners and nonowners observed here. Pet owners actually had a slightly higher BMI and exercised less than those who did not own pets, and adjusting for BMI or exercise 3 or more times per week did not change the associations. Furthermore, pet owners and those who did not own pets did not differ on either waist-hip ratio or percent body fat, and adjusting for these variables did not materially alter the results. In western cultures, BP, especially SBP, increases with age.\textsuperscript{1,3} The favorable associations of pets with BP in the present study were not independent of age and thus do not support the thesis that current pet ownership has an independent favorable effect on SBP or any other BP variable in older adults.

It is plausible that the type of pet owned or the owner’s affinity for the pet could affect BP. For example, a dog owner might be more physically fit than a cat owner, because dog owners might walk more than cat owners. However, no differences by type of pet owned were observed in this study. Those who have a greater affinity for their pet and perceive their pets as having an affinity for them may have different brain levels of serotonin or cortisol\textsuperscript{4,6}; these variables were not assessed in the present study. Likewise, duration of pet ownership could have an effect but was not assessed.

Several potential limitations of this study should be considered. The outcome measures of BP, pulse pressure were all objectively assessed, thereby reducing the risk of response bias. However, we cannot exclude the possibility of survival bias, whereby those without pets had the highest BPs but died prior to the study. An age-adjusted analysis of pet ownership (1991–1992) with current vital status through 2004 showed that those who did not own pets were more likely to have died than those who did own pets. Selection bias may also have occurred; healthier participants may have been more likely to answer questionnaires and attend clinic visits. However, if this had occurred, we would have been more likely to find differences between pet and nonpet groups after adjustment for covariates. There is also the possibility that participants may have given up their pets as illness limited their ability to care for pets, but participants were not queried about reasons for not owning pets. Finally, results from this study of well-educated, middle-class men and women with good access to medical care may not generalize to other samples of men and women. Previously published papers have shown that, as compared with national, representative samples of persons their age, participants in the Rancho Bernardo cohort are on average somewhat leaner,\textsuperscript{15} but are no more or less likely to have been cigarette smokers,\textsuperscript{16} and have similar levels of alcohol consumption, SBP, diabetes, impaired glucose tolerance, and plasma total cholesterol.\textsuperscript{17–22}

In conclusion, the overall results of the present cross-sectional study suggest that BP, hypertension, and vascular reactivity do not vary by pet ownership. The similarity of BP levels in those who did and did not own pets does not support the thesis that pet ownership lowers BP and risk of hypertension. The overall results of this study suggest that there is no beneficial effect of pet ownership on BP in old age.

REFERENCES


Does Low Maternal Blood Pressure During Pregnancy Increase the Risk of Perinatal Death?

Aimin Chen* and Olga Basso†

Background: A recent report described an association between low maximum diastolic blood pressure (DBP) during pregnancy and perinatal death (stillbirth and death in the first week combined). The authors did not account for gestational length, a strong predictor of perinatal death.

Methods: We studied 41,089 singleton pregnancies from the U.S. Collaborative Perinatal Project (1959–1966).

Results: We observed an association between low maximum DBP and elevated risk of perinatal death. However, this association disappeared after accounting for reverse causation related to gestational length. At any given gestational week, women whose offspring ultimately experienced perinatal death did not have significantly lower maximum DBP than women whose offspring survived the perinatal period. When accounting for the trend of increasing DBP during late pregnancy through gestational-age-specific DBP standardized score, we saw no association between low diastolic blood pressure and perinatal death.

Conclusions: Low maximum maternal DBP during pregnancy is a post hoc correlate of perinatal death, not a true risk factor.

A recent U.K. study suggested that babies whose mothers had low diastolic blood pressure (DBP) during pregnancy had an increased risk of perinatal death.1 A similar finding had been reported in an earlier study among a segment of women enrolled in the U.S. Collaborative Perinatal Project.2 Both studies used as a marker the highest recorded value of DBP during pregnancy. However, because blood pressure normally decreases in the early second trimester and then increases in late pregnancy,3 the maximum recorded DBP will depend in part on gestational length, a strong predictor of perinatal death. Furthermore, the maximum value within any given set of measurements becomes higher as the number of measurements increases.

This may result in a tendency for shorter pregnancies to have a lower maximum blood pressure.

To assess the association between low DBP and perinatal death, we reanalyzed the full data from the Collaborative Perinatal Project, with particular attention to the problems involved in using as a predictor the maximum DBP over the whole pregnancy.

Methods

The Collaborative Perinatal Project was a prospective pregnancy cohort of 58,760 pregnancies from 48,197 women recruited in 12 academic medical centers in the U.S. between 1959 and 1966.4 Women were enrolled during pregnancy (median gestation at enrollment: 21 weeks; interquartile range: 15–28 weeks).

We excluded 623 multifetal pregnancies and 3432 pregnancies lost before 20 weeks. For 9759 women contributing more than one singleton birth, we retained only the earliest. We further excluded 302 pregnancies with gestational length (based on the date of last menstrual period) that was missing or recorded as being below 20 or above 51 weeks (allowing for some degree of error in dating). We additionally excluded 1623 pregnancies with no valid maternal DBP measurement and 22 women with no DBP measurement higher than 40 mm Hg or lower than 110 mm Hg. Among women with at least one measurement within the above range, we excluded any value <40 or >110 mm Hg. Although we included women with hypertension during pregnancy in most analyses, we excluded those with reported history of hypertension before pregnancy and those who developed proteinuria during pregnancy (n = 1910). These restrictions left 41,089 singleton births for analysis.

Blood pressure was measured by sphygmomanometer at registration, at each subsequent prenatal visit, and at the time of labor. For diastolic blood pressure, either Korotkoff phase 4 (muffling of the sound) or phase 5 (disappearance of the sound) was used.5 If a woman had more than one blood pressure measurement during a specific gestational week, we used only the earliest valid measurement within that gestational week. Perinatal death was defined as the combination of fetal death after 20 weeks of gestation and death within the first 7 days after birth.

We started by using the highest DBP value during pregnancy as a predictor of perinatal death, as in previous studies.1,2 We then checked whether maximum DBP depended on the length of gestation and the number of measurements. By comparing the maximum DBP up to a given week between women who ultimately experienced perinatal death.
death and women who did not, we assessed whether a low maximum DBP predicted perinatal death. Finally, to account for the trend of increasing DBP in the third trimester of pregnancy, we calculated a gestational-age-specific DBP z-score, and used percentiles of mean z-score to predict perinatal death. In the logistic regression models, we adjusted for race, maternal age, parity, socioeconomic status, smoking during pregnancy, and maternal prepregnancy body mass index (BMI) (covariates categorized as in Table 1). We additionally adjusted for study center (Boston, Buffalo, New Orleans, New York Columbia, Baltimore, Richmond, Minneapolis, New York Metropolitan, Portland, Philadelphia, Providence, and Memphis). We used R 2.3.16 for graphs and splines, and SAS 9.17 for statistical analysis.

**RESULTS**

Among the 41,089 babies in the analysis sample, 671 were stillborn (1.6%) and 563 died within the first 7 days of life (1.4%), yielding a total of 1234 perinatal deaths (3.0%). Black race, older age, higher parity, lower socioeconomic status, smoking, and obesity were all associated with higher perinatal mortality, as was a shorter gestation (Table 1).

The mean ± SD gestational age at delivery was 39 ± 3 weeks. The mean number of blood pressure measurements per woman was 7 ± 4 (range, 1–29). Women whose baby died perinatally had shorter gestation (32 ± 7 weeks; range, 20–47), and fewer blood pressure measurements (5 ± 4; range, 1–26).

As expected, DBP decreased in the second trimester and increased during the third trimester (not shown). The mean maximum DBP increased with gestational length and also with increasing number of blood pressure measurements (from 75 mm Hg with 2 measurements to 83.5 mm Hg with 13 or more measurements). The effect of gestational length and number of measurements on mean DBP was, on the other hand, limited.

Similar to the findings of Steer et al,1 the relationship between maximum DBP and perinatal mortality was U-shaped, with highest mortality at the lowest maximum DBP (Fig. 1).

Women with lower maximum DBP had, on average, a shorter gestation and fewer BP measurements. Both of these factors could contribute to biasing any inference based on Figure 1, as suggested by the fact that, at any gestational age, a low maximum DBP up to that point was not predictive of perinatal death (Fig. 2).

When using gestational-age-specific mean DBP z-score percentiles, we saw little evidence of elevated risk of perinatal death among women with low z scores (Fig. 3). To obtain more precise estimates, we defined the lowest and highest 10% of z-score as “low” and “high” DBP. The perinatal death rate was 2.4% for the lowest and 5.5% for the highest decile. The corresponding adjusted ORs were 0.88 (95% confidence interval = 0.67–1.16) and 2.08 (1.68–2.59), respectively, compared with a DBP z-score between the 40th and 59th percentiles.

The results persisted when including DBP measurements with values <40 or >110 mm Hg, using mean DBP, restricting to women registered early in pregnancy or with DBP ≤90 mm Hg, or using different categorizations of mean DBP z-score (data not shown).

**DISCUSSION**

In this analysis, low maximum maternal DBP during pregnancy appeared to be a post hoc correlate of perinatal death. At no point in pregnancy was a low maximum DBP predictive of perinatal death, suggesting that such a marker would not be a useful clinical predictor of risk. The

---

** TABLE 1.** Perinatal Mortality by Sociodemographic Characteristics in 41,089 Pregnant Women, Collaborative Perinatal Project, 1959–1966

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>No.*</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>18950</td>
<td>2.75</td>
</tr>
<tr>
<td>Black</td>
<td>18653</td>
<td>3.23</td>
</tr>
<tr>
<td>Others</td>
<td>3486</td>
<td>3.13</td>
</tr>
<tr>
<td>Maternal age at registration (years)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;20</td>
<td>10644</td>
<td>2.60</td>
</tr>
<tr>
<td>20–24</td>
<td>14801</td>
<td>2.60</td>
</tr>
<tr>
<td>25–29</td>
<td>8346</td>
<td>3.02</td>
</tr>
<tr>
<td>30–34</td>
<td>4433</td>
<td>4.44</td>
</tr>
<tr>
<td>≥35</td>
<td>2865</td>
<td>4.29</td>
</tr>
<tr>
<td>Parity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>15367</td>
<td>2.71</td>
</tr>
<tr>
<td>1</td>
<td>8804</td>
<td>2.54</td>
</tr>
<tr>
<td>2</td>
<td>6046</td>
<td>3.06</td>
</tr>
<tr>
<td>≥3</td>
<td>10794</td>
<td>3.69</td>
</tr>
<tr>
<td>Maternal socioeconomic index†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0–3.9</td>
<td>15575</td>
<td>3.00</td>
</tr>
<tr>
<td>4.0–5.9</td>
<td>12304</td>
<td>3.02</td>
</tr>
<tr>
<td>6.0–9.5</td>
<td>12008</td>
<td>2.68</td>
</tr>
<tr>
<td>Smoking during pregnancy (cigarettes/d)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>21958</td>
<td>2.86</td>
</tr>
<tr>
<td>1–9</td>
<td>8036</td>
<td>2.65</td>
</tr>
<tr>
<td>10–19</td>
<td>5090</td>
<td>3.22</td>
</tr>
<tr>
<td>≥20</td>
<td>5681</td>
<td>3.68</td>
</tr>
<tr>
<td>Maternal pre-pregnancy BMI (kg/m²)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;20</td>
<td>10411</td>
<td>2.77</td>
</tr>
<tr>
<td>20–24.9</td>
<td>19890</td>
<td>2.76</td>
</tr>
<tr>
<td>25–29.9</td>
<td>5412</td>
<td>3.25</td>
</tr>
<tr>
<td>≥30</td>
<td>2141</td>
<td>3.78</td>
</tr>
<tr>
<td>Gestational age at delivery (weeks)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;28</td>
<td>599</td>
<td>62.27</td>
</tr>
<tr>
<td>28–31</td>
<td>916</td>
<td>21.18</td>
</tr>
<tr>
<td>32–36</td>
<td>5019</td>
<td>5.30</td>
</tr>
<tr>
<td>37–38</td>
<td>7805</td>
<td>1.56</td>
</tr>
<tr>
<td>39–40</td>
<td>15620</td>
<td>0.83</td>
</tr>
<tr>
<td>≥41</td>
<td>11130</td>
<td>1.34</td>
</tr>
</tbody>
</table>

*The sum for each characteristic may not be 41,089 due to missing values.
†Socioeconomic index is a composite index for education, occupation, and family income. A low score corresponds to a low socioeconomic status.
association between low maximum DBP across the whole pregnancy and perinatal death likely reflects bias due to differential-by-outcome mean lengths of gestation and the corresponding differential mean number of BP measurements. Thus, our analysis suggests that the previous findings\textsuperscript{1,2} likely resulted from an artifact due to the failure to account for reverse causation related to gestational length.

Using mean gestational-age-specific DBP z-score should produce a marker of DBP that eliminates the artifactual correlation between maximum DBP and gestational length. Our analysis based on the z-score showed no important association between lower DBP during pregnancy and perinatal mortality, although the lowest 1% of z-score came close to statistical significance. Whether this may have clinical relevance in a subgroup of women remains to be determined. The higher mortality with higher maternal blood pressure (Fig. 3) is consistent with the well-known risks of hypertensive disorders in pregnancy.\textsuperscript{8}

Previous studies have suggested that maternal hypotension during pregnancy (often defined as BP ≤110/65 mm Hg) may be associated with reduced uteroplacental perfusion, prematurity, and low birth weight.\textsuperscript{9 –13} However, these studies often included few women or lacked adjustment for potential confounders.\textsuperscript{14} Zhang and Klebanoff\textsuperscript{15} argued that the adverse outcomes in women with low DBP were due to their characteristics, such as low prepregnancy BMI and low social status, rather than to low DBP per se.

This was a large prospective cohort with multiple blood pressure measurements. However, our study has limitations. Recruitment occurred between 1959 and 1966, and perinatal death has substantially decreased since. Nonetheless, in replicating the analysis of Steer et al\textsuperscript{1} on this large cohort, we found an association when using maximum DBP throughout pregnancy, which disappeared after taking into consideration the fact that shorter gestations tended to have a low maximum DBP.

Overall, our analysis suggests that a low maternal blood pressure during pregnancy is not predictive of increased risk of perinatal death at the population level, and that maximum DBP 

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Perinatal mortality (diamonds) and 95% confidence interval (vertical bars) by maximum DBP during pregnancy in 41,089 pregnant women, Collaborative Perinatal Project, 1959–1966. The line is the smoothed spline of estimated perinatal mortality. Also, mean gestational length and mean number of BP measurements at each maximum DBP.}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Maximum DBP until a specific gestational week in ongoing pregnancies, stratified by whether the baby ultimately experienced perinatal death or not.}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Adjusted odds ratios and 95% confidence interval for perinatal death by percentiles of mean DBP z score. Also, perinatal mortality at each mean DBP z score.}
\end{figure}
DBP throughout pregnancy is subject to bias as a predictor of poor perinatal outcome.

ACKNOWLEDGMENTS

We thank Clarice R. Weinberg for invaluable advice on data analysis and writing, and Dale P. Sandler and Matthew P. Longnecker for comments on an earlier draft of the manuscript.

REFERENCES

How Much Would Closing Schools Reduce Transmission During an Influenza Pandemic?

Kathryn Glass and Belinda Barnes

Background: When deciding whether to close schools during an influenza pandemic, authorities must weigh the likely benefits against the expected social disruption. Although schools have been closed to slow the spread of influenza, there is limited evidence as to the impact on transmission of disease.

Methods: To assess the benefits of closing schools for various pandemic scenarios, we used a stochastic mathematical model of disease transmission fitted to attack rates from past influenza pandemics. We compared these benefits with those achieved by other interventions targeted at children.

Results: Closing schools can reduce transmission among children considerably, but has only a moderate impact on average transmission rates among all individuals (both adults and children) under most scenarios. Much of the benefit of closing schools can be achieved if schools are closed by the time that 2% of children are infected; if the intervention is delayed until 20% of children are infected, there is little benefit. Immunization of all school children provides only a slight improvement over closing schools, indicating that schools are an important venue for transmission between children. Relative attack rates in adults and children provide a good indication of the likely benefit of closing schools, with the greatest impact seen for infections with high attack rates in children.

Conclusions: Closing schools is effective at reducing transmission between children but has only a moderate effect on average transmission rates in the wider population unless children are disproportionately affected.

(Epidemiology 2007;18: 623–628)

In the event of an influenza pandemic, health authorities can choose from among a number of interventions to reduce the spread of infection. Closing schools is one intervention that has been used during past influenza pandemics.1 However, the disruption to the community that results2 makes its use controversial. Before a decision to close schools is taken, it is essential to assess its likely impact, as well as the circumstances under which it is most effective.

There is evidence that interventions targeted at children are effective for endemic influenza. In Japan, excess mortality rates dropped after the introduction of a policy of vaccinating school children and increased again when the vaccination policy was discontinued. The changes in mortality rates were largely attributable to adult deaths from influenza and pneumonia.3 In Israel, the incidence of respiratory illness in school children decreased by around 40% during a 2-week period in which schools were closed because of a labor dispute by teachers.4 However, there is less evidence supporting closure of schools during an influenza pandemic, where adults may have little or no immunity to the pathogen. A recent review of pandemic influenza interventions5 noted that “data on the effectiveness of school closures are limited”; the review cited situations in which closing schools was believed to have been effective at reducing transmission as well as situations in which case numbers were higher when children were not at school. Modeling studies have suggested that vaccination of children could be an effective intervention during an influenza pandemic.5 Modeling analyses of school closures have variously shown that closing schools may have primary use in flattening a pandemic, rather than reducing the overall attack rate,6 or that complete isolation of school children can be very effective at reducing the attack rate.7

In this paper, we assess the impact of closing schools on the attack rate in children and the overall attack rate in the population. We do so using a stochastic household model of 1 million households with sizes and composition of adults and children matching Australian census data. This model incorporates heterogeneity in household size, and takes account of the fact that transmission between children is an important route through which large households infect one another. The main focus of this work is to calculate the likely impact of closing schools during an influenza pandemic. In the process, we consider a range of reproduction numbers and levels of immunity in adults and children that is wider than the range considered plausible for pandemic influenza. This lets us compare the effect of closing schools during an influenza pandemic with closing schools during an outbreak of an endemic infection.
METHODS

Stochastic Simulation Model
We use a stochastic model of transmission in 1 million households with 2 types of individuals: adults and school children. Household sizes and compositions are based on data from the 2001 Australian census, as outlined in eTable 1 (available with the online version of this article at www.epidem.com.) School children make up 18% of the population, and their households have 3.1 other members on average; adults have an average of 1.6 other individuals in their household. The population contains around 2.6 million individuals in total. Within the household, disease spreads according to a Reed-Frost model. Mixing time outside the household is split between time spent at work or school, and time spent mixing in the community. Mixing at school or work is assumed to be with individuals of the same type, while mixing within the community includes both types. Full details of the model structure and assumptions are provided in the online Appendix, together with a discussion of alternative mixing assumptions, and the sensitivity of the model to the main parameters.

Modeling Interventions
Closing schools is modeled by eliminating all transmission among children at school, while increasing transmission between school children within the household in proportion to the extra time spent at home. This represents an “ideal” school closure, in which children are kept at home during school hours. Similar modifications are made to take account of some adults staying home to care for children, assuming that one adult stays home from work in every household with school children and no nonworking adult. As we do not distinguish between adults in the household, this role might be assumed by one adult, or shared by more than one. Immunization of children is implemented by assuming that all school children are immune before the epidemic begins. Again, this represents an ideal case, used for the purposes of comparison.

In the early stages of the outbreak, it may be possible to isolate all diagnosed cases (both children and adults) in hospitals, thus greatly reducing the spread of infection after isolation. Isolation of cases was incorporated by expanding the model to identify health care workers who care for diagnosed and isolated cases. This expanded model is considered both with and without the use of additional protection for health care workers. Where protection is included, transmission between the case and health care worker is reduced to 2% of the unprotected value, representing a combination of antivirals for both case and caregiver and personal protective equipment for health care workers.

Model Reproduction Number and Attack Rate Approximations
To obtain general results for a wide range of reproduction numbers and immunity levels, we used theoretical results that allow a reproduction number for the household model to be expressed in a relatively simple form. This reproduction number was used to compare the effectiveness of interventions under different levels of adult and child immunity. The expected attack rates without intervention were derived analytically from an analogous differential equation model, again allowing for differing levels of immunity in adults and children. See the online Appendix for details of these calculations.

Parameter Values
An analysis of influenza data from Tecumseh was used to calibrate the relative amount of transmission between and within households. Outside the household, individuals are assumed to spend 60% of their time at work or school, although results remained broadly similar across the range 40–80%. (See the Appendix for a discussion of the sensitivity of the model to these mixing assumptions, available with the electronic version of this article.)

The model includes a school-mixing parameter to take account of extra transmission between children at school. This parameter was estimated by fitting the model to age-specific attack rates from Kansas City from the 1957 and 1968–1969 influenza pandemics, adjusting for the fact that all adults included in this study were family members of school children. The age-specific attack rates for the 2 pandemic years show different patterns: the 1957 data have relatively high attack rates in children, while the 1968–1969 data have a fairly flat age-specific attack rate curve. We estimated 2 values of the school-mixing parameter corresponding to the 2 pandemics, and used the model with these values to compare the 1957 and 1968–1969 scenarios. We then cross-checked the estimated values by calculating the percentage of households in which a school-aged child was the index case. The model showed good agreement with the data, although the percentage of child index cases was slightly lower in the model because of smaller household sizes in Australia. See the online Appendix for further details.

It seems likely that immunity from previously circulating strains of influenza partly accounts for the lower attack rates in adults seen in the 1957 pandemic (as has been assumed previously). When adult immunity was incorporated into the model, closing schools under the 1957 scenario became less effective, although still more effective than under the 1968–1969 scenario. For reasons of space, we present results only from the 1957 scenario without immunity. However, we note that this scenario represents an extreme case, where heightened attack rates in school children are entirely attributed to extra mixing at school. Unless otherwise stated, we assume that all individuals in the population are susceptible.

The Reproduction Number
The reproduction number is a measure of the infectiousness of a disease. In a population where individuals mix homogeneously, it is defined as the mean number of infections generated by a single infected individual over the course of his or her infectious period. The basic reproduction number (\(R_0\)) is in the absence of control measures, while the effective reproduction number (\(R\)) is calculated once control measures have been introduced. Estimates of the basic reproduction number for pandemic influenza vary, with recent
estimates lying between 1.4 and 3.75. Although high attack rates in past pandemics suggest that the basic reproduction number in some isolated communities may be much higher, it seems likely that the reproduction number in a typical western population lies between 1.5 and 3.5. When we consider the impact of school closure on infections spread person-to-person in general, we use a wider range of possible basic reproduction numbers. This allows us to compare the pandemic influenza scenario with that of an endemic childhood infection.

RESULTS

Figure 1 shows the effect of closing schools at the start of the outbreak on the epidemic curve for values of the basic reproduction number ($R_0$) of 1.5 and 2.5, comparing the model calibrated to age-specific attack rates from the 1968–1969 and the 1957 pandemics. As the model is stochastic, there is some variability in the timing and the size of the epidemic peak. In each plot, the curve shows the median of 100 simulations, and the shaded area contains 90% of the simulations, with the upper and lower 5th percentiles omitted. This shaded region indicates the variability seen over different realizations of the stochastic model. The case with no intervention (red dotted line) is compared with the case where schools are closed (blue solid line), and the case where schools are closed and some parents stay home from work to care for their children (black dashed line).

Closing schools has most effect when $R_0$ is low, and is more effective if there is a high attack rate in children (as in 1957) than if attack rates in adults and children are similar (as in 1968–1969). When $R_0$ is 2.5, closing schools has only a moderate effect on the epidemic curve, and the effect is even less noticeable for higher values of the basic reproduction number. The situation in which parents stay home from work to care for children has little impact on the spread of the infection in the model. Although the final attack rates under no intervention are the same for the 1957 and 1968–1969 scenarios, the baseline epidemics (red dotted lines) under the 1957 scenario take off faster and have a higher epidemic peak than those under the 1968–1969 scenario. This is a consequence of the greater spread of infection between children in the 1957 scenario, which boosts the initial take-off of the outbreak.

Figure 2 shows the effect of the timing of school closure on the final attack rate in adults (dashed lines), children (solid lines), and overall (dotted lines) for a range of values of the basic reproduction number, and for the 1957 and the 1968–1969 scenarios. In each plot, the horizontal axis is the prevalence in school children at the time when schools are

FIGURE 1. The impact of closing schools on the epidemic curves for $R_0$ values of 1.5 and 2.5, with transmission rates fitted to the 1968–1969 and the 1957 pandemic attack rates. Each plot compares the case with no intervention (red dotted line), the case where schools are closed at the start of the epidemic (blue solid line), and the case where schools are closed and parents stay home to look after the children (black dashed line). In each case, the lines represent the median of 100 stochastic simulations, and the shaded area contains 90% of the simulations, with the upper and lower 5th percentiles removed. The model used is the stochastic transmission model containing 1 million households, corresponding to a population of around 2.6 million individuals.

FIGURE 2. The effect of the timing of school closure on the final attack rate in a population of 1 million households containing around half a million children. In each plot, the horizontal axis shows the number of current cases in school children that will trigger school closure, using a log scale. In the case where this value exceeds the number of children infected, schools remain open throughout the epidemic. Each curve shows the median of 100 simulations of the stochastic model.
closed in the population of 1 million households. This axis is shown on a log scale, and extends to 0.8 for visual clarity although the prevalence never reaches that level. Where the intervention threshold for closing schools (shown on the horizontal axis) exceeds the maximum prevalence achieved during the epidemic, the intervention is never introduced.

The model indicates that most of the effect of closing schools can be achieved in this population if schools are closed by the time the prevalence reaches around 2% in school children; in contrast, if the intervention is delayed until 20% of children are infected, most of the effect of closing schools will be lost. If the model is run with fewer households, the threshold is still 2%. The horizontal axis represents the true prevalence in the population. If there is a high proportion of asymptomatic or undiagnosed cases, the number of diagnosed cases will underestimate this true prevalence, and the timing of school closure should be adjusted accordingly.

As there is extra mixing between children in the model, attack rates in children are higher than in adults if schools are kept open, and the difference is particularly large under the 1957 mixing assumptions. For values of the basic reproduction number of 2.5 or 3.5, nearly all children become infected even though the average attack rates for the population are 89% and 97%, respectively. Figure 2 shows that closing schools reduces the attack rate in school children to a much greater extent than the overall average attack rate. The greatest effect is seen for a basic reproduction number of 1.5 under the 1957 scenario, where closing schools reduces the attack rate in children from 84% to 32%, and the overall attack rate from 58% to 38%. This is the most optimistic scenario for closing schools, as it assumes a high level of mixing in schools, and a low basic reproduction number. The other cases considered here show a reduction in the attack rate in children of 10–35%, and a reduction in the overall attack rate of 3–15%. All plots in Figure 2 were created under the assumption that, once closed, schools remained closed until the epidemic was over. We also considered the effect of reopening schools as the epidemic dies out. Provided the prevalence in children is less than 1% when schools are reopened, most of the benefit of having closed the schools is retained.

In Figure 3, the effect of closing schools is compared with that of successfully immunizing all school children. While such an intervention is unlikely to be achievable in practice, it represents the effect of completely removing all transmission from children. Figure 3 compares these interventions under the 1968–1969 scenario, assuming that all diagnosed cases (both adults and children) are isolated, with Figure 3A representing the case where patients and health-care workers caring for these isolated cases are not protected, and Figure 3B the case where patients and health-care workers are given antivirals, and health-care workers are given personal protective equipment. Both plots show the contours where the effective reproduction number is equal to 1 under the interventions, so that the regions below the curves include all parameter values for which the intervention will bring transmission under control. In each plot, the vertical axis shows the basic reproduction number, and the horizontal axis represents the fraction of infectiousness experienced by cases before isolation. The position on the horizontal axis will depend both on the timeliness of isolation and on the profile of infectiousness in an individual after infection. If the typical infectiousness profile for influenza resembles that estimated by Ferguson et al, then most of an individual’s infectious-

![Figure 3](image-url)

**FIGURE 3.** A comparison of interventions targeted at children if all cases are isolated and cared for by health-care workers upon diagnosis. The graphs compare closing schools (solid lines) and immunizing children (dashed line) to no further intervention (dotted line) under the 1968–1969 scenario. Each plot shows the parameter values for which the effective reproduction number is equal to 1, so that the region under each curve represents parameter values for which the intervention will eliminate the infection. The horizontal axis shows the fraction of infectiousness experienced by a case before he or she is isolated, while the vertical axis gives the basic reproduction number \( R_0 \). Plot (A) assumes that there is no protection for health-care workers caring for isolated patients, and plot (B) assumes that health-care workers are protected by antivirals and personal protective equipment. The assumptions concerning the level of protection provided to health-care workers are described in the text.
ness will occur before isolation. See the Appendix for a more detailed discussion of this point.

Figure 3 clearly shows the effectiveness of isolating cases and providing protection for health-care workers if individuals can be isolated before they have experienced much of their infectivity. In reality, values on the horizontal axis below 0.5 are unlikely to be achievable in practice, and even values in the range 0.5–0.8 may be impractical if individuals are very infectious in the early stages of infection. Figure 3 confirms that immunizing all school children is more effective than closing schools, but suggests that much of the benefit of eliminating transmission from children can be achieved by closing schools, provided that school children can be kept at home when schools are closed. Each intervention is only moderately effective at reducing transmission under the 1968–1969 scenario, indicating that if attack rates in children are similar to those in adults, interventions targeted at children will have limited ability to reduce transmission. The model fitted to 1957 attack rates shows a greater absolute effect for both interventions, but shows similar relative effects of immunization and closing schools.

In Figure 4, we broaden the study to look at the effect of closing schools on the effective reproduction number of a pathogen for which there may be existing immunity in the population and the level of immunity may differ between adults and school children. One example of such an alternative scenario is a local outbreak of measles in a country in which measles is not fully controlled by vaccination. Under this scenario, the basic reproduction number is likely to be high, and adult immunity would be expected to be very much greater than immunity in children. In Figure 4, the vertical axis shows the fractional reduction in the reproduction number when schools are closed, while the horizontal axis indicates the extent to which children are disproportionately affected before intervention; a value of 1 indicates that attack rates are equal in adults and children, and a value close to 0 indicates that hardly any cases occur in adults. To create the graph, we compared the reproduction number before and after schools are closed for values of the initial reproduction number ranging up to 18, and initial immunity in adults ranging from 0% to 90%, assuming that immunity in children is no greater than in adults. Each cross on the graph represents one such calculation. As this model is intended to compare the effect of closing schools for a range of infectious diseases spread in a similar manner to influenza, we performed a thorough sensitivity analysis on the parameters in the model that remained fixed, and also tested the effect of clustered immunity in households. This analysis confirmed that the results presented here are robust to changes to the mixing and immunity assumptions.

Figure 4 shows a clear association between relative attack rates in adults and children and the effect of closing schools; furthermore, this association is robust across this wide range of parameter values. If attack rates in the 2 groups are similar, the reproduction number is reduced by around 20% by closing schools, whereas if the attack rate in children is much higher than in adults, closing schools can reduce the reproduction number by as much as 75%. The ratio of attack rates in adults to attack rates in children calculated from pandemic influenza data from 1957 and 1968–1969 lies between 0.5 and 0.9, indicating that closing schools is only moderately effective for pandemic influenza. In comparison, closing schools could be very effective in reducing the reproduction number of an endemic childhood infection where cases are predominantly among children.

DISCUSSION

Closing schools is moderately effective at reducing transmission during an influenza pandemic, particularly if attack rates in children are high in comparison to adults, or if the basic reproduction number is low. Even when the intervention has little effect on the overall attack rate, closing schools can be helpful in protecting children from infection. To provide maximum benefit, schools should be closed by the time that around 2% of children are currently infected; if the intervention is delayed until 20% of children are infected, very little will be gained. In agreement with other modeling work, we find that as the epidemic dies out, it is safe to reopen schools when the prevalence in children has returned to low levels (less than 1%, say).

Although closing schools does not reduce transmission as effectively as immunizing all school children, much of the benefit of full immunization is achieved by closing schools.
This indicates that a large proportion of transmission between children occurs at school. However, to implement school closure successfully, children must be kept at home during school hours, which may require parents to stay home from work. A table giving the distribution of working and nonworking adults in Australian households is provided in the online Appendix (Table 2, available with the electronic version of this article at www.epidem.com). Roughly 50% of households containing children have no nonworking adult. If one adult stayed home from work in each of these households, about 14% of the workforce would be unavailable for work. Although this would cause considerable disruption, it does not lead to much additional benefit in terms of reduced transmission. Many adults do not work, so the change in behavior affects only around 8% of the total adult population (that is, both working and nonworking adults). Recent modeling work has found large differences in the effectiveness of closing schools according to the extent to which children can be segregated from the community. Here, we have assumed that the level of community mixing is unchanged, as this seems more likely to be achievable. If families succeed in isolating themselves from the community, these measures would be more effective, but if community mixing were to increase following closure of schools, the benefit of closing schools is reduced. When assessing the most effective means of protecting children, the benefits of closing schools must be weighed against the likely disturbance to the community. Should a vaccine effective for school children become available early enough in the pandemic, immunization of children may provide a less disruptive method of protecting children from infection.

Previous modeling studies of pandemic influenza identify some differences in the likely benefit of closing schools. Ferguson et al. found that school closure can reduce peak attack rates by up to 40%, but has little impact on overall attack rates. In contrast, Glass et al. found that closing schools with 90% compliance reduces the overall attack rate by 22%. It seems likely that assumptions about baseline age-specific attack rates will at least partly explain this discrepancy. Glass and colleagues assume a baseline attack rate in children of 78% compared with an attack rate of 44% in adults, while the Ferguson model does not appear to include enhanced mixing among children, suggesting that baseline attack rates in children are similar to adults. In this paper, we show that a comparison of attack rates in adults and children provides a good indication of the likely benefits of closing schools, with much greater benefits seen when children are disproportionately affected. Monitoring age-specific attack rates during the early stages of an influenza pandemic will provide valuable information concerning the likely benefit of closing schools. This finding is robust over a wide range of parameter values and levels of immunity, and thus has more general implications for infectious diseases that are similar to influenza. In general, closing schools will be more effective against infections that show high attack rates in children, including childhood infections (to which most adults are immune), than against novel infections for which all age groups are similarly susceptible.

ACKNOWLEDGMENTS We thank Niels Becker, Jodie McVernon, and James McCaw for their valuable feedback on the manuscript.

REFERENCES

Cigarette Smoking and Risk of Breast Carcinoma In Situ

Amy Trentham-Dietz,*† Hazel B. Nichols,* Kathleen M. Egan,‡ Linda Titus-Ernstoff,§ John M. Hampton,* and Polly A. Newcomb¶

Background: Although the associations with cigarette smoking have been explored extensively for invasive breast cancer, the relation to in situ cancer has not previously been examined in depth.

Methods: We analyzed data from a population-based case-control study of women living in Wisconsin, Massachusetts, and New Hampshire. Eligible cases of incident breast carcinoma in situ were reported to statewide registries in 1997–2001 (n = 1878); similarly aged controls (n = 8041) were randomly selected from population lists. Smoking history and other risk factor information were collected through structured telephone interviews. Odds ratios (ORs) and 95% confidence intervals (95% CIs) were calculated from logistic regression models adjusting for potential confounders.

Results: In multivariate models, the OR for breast carcinoma in situ among current smokers was 0.8, compared with never-smokers (95% CI = 0.7–1.0). Risk estimates increased towards the null with greater time since smoking cessation. Odds ratios were also less than 1.0 among women who initiated smoking in adolescence (OR = 0.8) or after a full-term birth (OR = 0.7), relative to women who never smoked. The reduced odds ratios associated with current smoking were strongest among women with annual screening mammograms (OR = 0.7, 95% CI = 0.6–0.9). Odds ratios were not less than 1.0 among current smokers without a recent screening mammogram (1.3; 0.9–2.0).

Conclusions: Our findings suggest an inverse association between current smoking and risk of breast carcinoma in situ among women undergoing breast cancer screening.

(Epidemiology 2007;18: 629–638)

Cigarette smoking is consistently associated with most malignancies except breast cancer.1,2 Smoking behavior has generally been explored in relation to invasive breast cancer, with modest effects observed according to age at initiation, duration, and cessation of smoking. These subtle changes in risk may be due, in part, to the joint antiestrogenic and carcinogenic properties of cigarette smoke and underscore the importance of accurately capturing the window of exposure.3–6

Since the advent of population mammographic screening in the 1980s, the incidence of breast carcinoma in situ has increased more than 7-fold. This increase has been observed predominately among women age 50 years and older who follow the recommended screening guidelines.7 An estimated 62,000 cases of breast carcinoma in situ will be diagnosed in the United States in 2007.8 Ductal carcinoma in situ, in particular, is often identified through the use of mammograms,9 whereas lobular carcinoma in situ may more likely be identified as an incidental finding.

Although survival rates exceed 95%,9 breast carcinoma in situ has clear links to invasive disease. Women with a history of breast carcinoma in situ have a 2- to 9-fold excess risk of developing invasive breast cancer.10,11 Women with breast carcinoma in situ may suffer substantial worry and concern after their diagnosis, and many experience a second breast cancer diagnosis despite aggressive treatments.12,13 Identification of factors related to incidence is therefore of significant public health interest.

Although the associations with cigarette smoking have been explored extensively for invasive breast cancer,5–6 the relation to in situ cancer has less frequently been studied. To further investigate these associations, we examined data from a population-based case-control study.

This analysis was performed using data from the Three State Study, a population-based case-control study conducted at the University of Wisconsin, Madison, Wisconsin; the Harvard School of Public Health, Boston, Massachusetts; and Dartmouth University, Lebanon, New Hampshire. The study was conducted according to institutionally approved protocols at each study site.

Selection of Cases

Women ages 20–74 years residing in Wisconsin, Massachusetts (excluding metropolitan Boston), and New Hampshire with a new diagnosis of breast carcinoma in situ (ICD-O version 2 C50.0-C50.9) reported to each state’s cancer registry during 1997–2001 were eligible for this study. The physician of record for each eligible case was contacted by mail and provided an opportunity to actively refuse participation on the case patient’s behalf. Interviews were conducted from February 1997 to May 2001. Eligibility was limited to case subjects with listed telephone numbers, driver’s licenses verified by self-report (if less than 65 years of age),
and known dates of diagnosis. A total of 2269 cases were eligible for the study. Of these cases, physicians refused contact with 58 (2.6%), 17 (0.7%) were deceased, 63 (2.8%) could not be located, and 244 (10.8%) refused to participate. Overall, 1887 (83%) women were interviewed. Participation was similar between states. Nine interviewed cases were considered unreliable by the interviewers, leaving 1878 cases available for analysis. Of the interviewed cases, 99% were confirmed by histology, cytology, or other means according to the registry reports. For this study, in situ cases were distinguished according to the fifth digit behavior code (in situ = 2) of the morphology code. Subtypes of breast carcinoma in situ were defined as lobular morphology (8520) and ductal/nonlobular (8500, 8501, 8503, 8504, 8010, and 8140). Histologic classification (ductal or lobular) was not available for 10% (n = 184) of cases. Of the remaining 1694 cases, 87% (n = 1471) were ductal/nonlobular and 13% (n = 223) were lobular.

**Selection of Controls**

Community controls were randomly selected during 1997–2001 in each state using 2 sampling frames: those under 65 years of age were selected from lists of licensed drivers, and those 65 to 74 years of age were selected from a roster of Medicare beneficiaries compiled by the Centers for Medicare & Medicaid Services. Controls were selected at random within 5-year age strata to yield an age distribution similar to the cases enrolled in each state. Controls were required to have no personal history of breast cancer, and a listed telephone number. Of the 10,690 potential controls approached for participation, 86 (0.8%) were deceased, 475 (4.4%) could not be located, and 2074 (19.4%) refused to participate. Interviews were obtained for 8055 (75%) of these women. Fourteen interviewed controls were considered unreliable by the interviewers; hence, 8041 controls were available for analysis.

**Data Collection**

Case subjects and controls were sent letters briefly describing the study before they were contacted by trained telephone interviewers. Interviews for case and control participants were conducted contemporaneously. The 40-minute interview elicited detailed information on smoking, screening mammography, reproductive/menstrual history, use of exogenous hormones (oral contraceptives and postmenopausal therapy), recreational physical activity, occupational history, alcohol consumption, height and weight, personal and family medical history, and demographic factors. Information about the woman’s personal and family history of cancer was obtained at the end of the interview to maintain blinding. For 95% of cases and 93% of controls, the interviewers reported being unaware of the woman’s case-control status until the end of the interview.

**Exposure Definitions**

Participants who indicated during the telephone interview that they had smoked more than 100 cigarettes before the referent date were classified as smokers. Those who had not smoked more than 100 cigarettes were classified as never-smokers. We categorized study participants who reported smoking during the year preceding the referent date as “current smokers.” Duration of smoking was defined as the number of years participants reported smoking regularly. Pack-years were calculated as the average number of packs of cigarettes smoked per day times the smoking duration in years. We created categories of years of smoking, pack-years and age of smoking initiation based on approximate quartiles among the controls. Categories of cigarettes per day and years since smoking cessation were created in 10-unit intervals.

We defined a woman as postmenopausal if she reported a natural menopause (no menstrual periods for at least 6 months) before the referent date. Women who reported taking postmenopausal hormones and still having periods were classified as (1) premenopausal if their referent ages were in the first decile of age at natural menopause among the controls (<41 years of age for current smokers and <43 years of age for nonsmokers), (2) postmenopausal if their referent ages were in the highest decile for age at natural menopause in the control group (≥54 years of age for current smokers and ≥56 years of age for nonsmokers) with age at menopause defined as unknown, and (3) unknown menopausal status if age at intermediate ages (second to ninth decile). Menopausal status was unknown for 527 controls (7%) and 132 cases (7%); reasons included report of a hysterectomy but not bilateral oophorectomy (n = 361) or current use of postmenopausal hormones (n = 198) among women age 43–55 (age 41–53 for current smokers), reports of being unsure whether a menstrual period had occurred within 6 months (n = 68), and reports of unspecified female surgery (n = 28) or never having periods (n = 4). We evaluated alcohol use as frequency of beer, red wine, white wine, or liquor consumption during the year previous to the referent date. Weight gain (or loss) was determined by subtracting each participant’s weight at age 18 from her weight one year before the referent date (“recent weight”). Body mass index (BMI) was calculated as recent weight (kg)/tallest adult height (m).²

**Statistical Analysis**

For each case, a referent date was defined as the registry-supplied date of breast carcinoma in situ diagnosis. For comparability, the control subjects were assigned a referent date that corresponded to the average diagnosis date for cases of similar ages (within 5-year strata by state). Referent age was defined as the woman’s age at the referent date. Only exposures that occurred before the assigned referent date were included in analyses.

We considered the following potential confounders: age at menarche, age at first full-term pregnancy, parity, menopausal status, age at menopause, oral contraceptive and postmenopausal hormone use, family history of breast cancer, education, body mass index, weight change since age 18, alcohol consumption, personal history of benign breast disease, and mammography use. Variables that were associated with risk of breast carcinoma in situ (P ≤ 0.05 in age and state-adjusted models) were included in multivariate models. Our study population was 94.8% white (1799 cases and 7605 controls); thus, we did not evaluate breast carcinoma in situ risk associations by race or ethnicity.
Multivariate models were adjusted for age (<40, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, ≥70 years), state (MA, NH, WI), age at menarche (<12, 12, 13, ≥14, unknown), age at first birth (<20, 20–24, 25–29, ≥30, unknown), parity (≤1, 2, ≥3, unknown), menopausal status (premenopausal, postmenopausal, unknown), age at menopause (<45, 45–49, 50–54, ≥55, unknown), oral contraceptive use (never, ever), postmenopausal hormone use (never, former, current), family history of breast cancer (yes, no, unknown), education (less than high school diploma, high school diploma, some college, college diploma, unknown), weight at age 18 (continuous), height (continuous), weight change since age 18 (weight loss, weight gains of 0–15 lbs, 16–30 lbs, 31–50 lbs, >50 lbs, unknown), personal history of benign breast disease (yes, no, unknown), alcohol consumption (0 drinks per week, <7, 7–13, ≥14, unknown), and number of mammograms within 5 years before the referent date (none, less than five, 5 or more, unknown). Although age at menarche and alcohol consumption were not associated with breast carcinoma in situ in preliminary models, these 2 variables were included in multivariate models as decided a priori.

Odds ratios (ORs) and 95% confidence intervals (CIs) for breast cancer were produced using multivariate logistic regression models adjusted for the covariates described above. To obtain P-values for trend, we included select variables as continuous linear terms in regression models. Effect modification was evaluated by inclusion of cross-product interaction terms in logistic models and measuring the change in the log-likelihood using χ² tests. The analyses described above were performed using SAS version 9.1 software (SAS Institute, Cary, NC).

## RESULTS

Table 1 presents the distribution of age at diagnosis, histologic type, and the method of detection for breast carcinoma in situ cases. As expected, the majority (77.5%; n = 1456) of breast carcinoma in situ cases were discovered by screening mammogram.

In multivariate models, ORs for breast carcinoma in situ increased with older age at first birth and menopause, low parity, oral contraceptive and postmenopausal hormone use, a personal history of benign breast disease, greater frequency of mammography screening, a family history of breast cancer, and (among postmenopausal women) weight gain since age 18 (Table 2). Compared with premenopausal women, postmenopausal women had a lower odds of breast carcinoma in situ. Odds did not appear to be strongly associated with age at menarche, education, or alcohol consumption.

Among women in our study, approximately half reported ever smoking (49% of cases; 52% of controls). Of these, 68% of cases and 59% of controls were categorized as former smokers and 32% of cases and 41% of controls as current smokers. Smoking history was strongly related to frequency of mammography among control women: annual screening examinations were reported by 47% of former smokers, 42% of never-smokers, and 32.2% of current smokers. Current smokers (22%) were twice as likely as former smokers (11%) to report no screening mammograms in this period, with intermediate rates in never-smokers (14%; P for trend <0.001).

In the adjusted models, we observed a lower odds ratio for breast carcinoma in situ (OR = 0.82; 95% CI = 0.70–0.96) among current, but not former (1.00; 0.88–1.13) smokers, compared with never-smokers (Table 3). We did not observe a clear pattern of risk associated with number of cigarettes per day, duration of smoking in years, or pack-years of smoking. Odds ratios were less than 1.0 among women who initiated smoking at younger ages. Women who started smoking at age 18 or younger had 0.81 times the risk compared with never-smokers. Risk increased towards the null with greater time since smoking cessation; women who smoked more recently appeared to be at decreased risk compared with never-smokers. In models restricted to parous women, we also observed lower odds ratios among women who initiated smoking habits after the first birth, compared with never-smokers (0.74; 0.57–0.96). The direction of the association was the same for both premenopausal and postmenopausal women (OR = 0.73, 95% CI = 0.53–1.01 and OR = 0.84, 95% CI = 0.67–1.07, respectively).

Heterogeneity of the relation between cigarette smoking and breast carcinoma in situ was explored according to age as well as the factors shown in Table 2 (data not shown). Among these, interaction was suggested only with frequency of mammography screening in the 5 years preceding the referent date (χ² = 16.092, 2 df; P < 0.001 for never/ever smoking versus any mammography and χ² = 19.673, 4 df; P < 0.001 for never/former/current smoking versus 0/<5/5 mammograms).
### TABLE 2. Association of Selected Characteristics With Breast Carcinoma In Situ

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Cases (n = 1878)</th>
<th>Controls (n = 8041)</th>
<th>OR (95% CI)</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. (%)</td>
<td>No. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at menarche</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;12§</td>
<td>358 (19.1)</td>
<td>1573 (19.5)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>12</td>
<td>480 (25.6)</td>
<td>1846 (23.0)</td>
<td>1.13 (0.97–1.32)</td>
<td>1.12 (0.96–1.32)</td>
</tr>
<tr>
<td>13</td>
<td>521 (27.7)</td>
<td>2193 (27.3)</td>
<td>1.04 (0.89–1.20)</td>
<td>1.03 (0.88–1.21)</td>
</tr>
<tr>
<td>≥14</td>
<td>477 (25.4)</td>
<td>2244 (27.9)</td>
<td>0.91 (0.78–1.07)</td>
<td>0.93 (0.80–1.10)</td>
</tr>
<tr>
<td>P for trend</td>
<td></td>
<td></td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>Parity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nulliparous§</td>
<td>271 (14.4)</td>
<td>967 (12.0)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1–2</td>
<td>818 (43.6)</td>
<td>3174 (39.9)</td>
<td>0.90 (0.77–1.06)</td>
<td>0.92 (0.78–1.08)</td>
</tr>
<tr>
<td>3–4</td>
<td>629 (33.5)</td>
<td>2840 (35.1)</td>
<td>0.77 (0.65–0.90)</td>
<td>0.79 (0.67–0.94)</td>
</tr>
<tr>
<td>5 or more</td>
<td>144 (7.7)</td>
<td>1011 (12.5)</td>
<td>0.48 (0.38–0.60)</td>
<td>0.54 (0.42–0.69)</td>
</tr>
<tr>
<td>P for trend</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age at first birth§</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;20§</td>
<td>235 (14.8)</td>
<td>1371 (19.3)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>20–24</td>
<td>733 (46.1)</td>
<td>3399 (48.2)</td>
<td>1.24 (1.06–1.46)</td>
<td>1.22 (1.02–1.44)</td>
</tr>
<tr>
<td>25–29</td>
<td>435 (27.3)</td>
<td>1580 (22.7)</td>
<td>1.59 (1.34–1.90)</td>
<td>1.57 (1.29–1.92)</td>
</tr>
<tr>
<td>≥30</td>
<td>186 (11.7)</td>
<td>647 (9.4)</td>
<td>1.63 (1.31–2.03)</td>
<td>1.61 (1.25–2.08)</td>
</tr>
<tr>
<td>P for trend</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Menopausal status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premenopausal§</td>
<td>717 (38.2)</td>
<td>2936 (36.8)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Postmenopausal</td>
<td>1029 (54.8)</td>
<td>4578 (56.2)</td>
<td>0.87 (0.72–1.04)</td>
<td>0.75 (0.61–0.91)</td>
</tr>
<tr>
<td>Unknown</td>
<td>132 (7.0)</td>
<td>527 (7.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at menopause§</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;45§</td>
<td>171 (16.6)</td>
<td>876 (19.7)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>45–49</td>
<td>213 (20.7)</td>
<td>974 (21.6)</td>
<td>1.03 (0.84–1.26)</td>
<td>1.02 (0.83–1.26)</td>
</tr>
<tr>
<td>50–54</td>
<td>301 (29.3)</td>
<td>1348 (29.0)</td>
<td>1.10 (0.91–1.33)</td>
<td>1.10 (0.90–1.34)</td>
</tr>
<tr>
<td>≥55</td>
<td>118 (11.5)</td>
<td>448 (20.1)</td>
<td>1.33 (1.04–1.71)</td>
<td>1.35 (1.04–1.76)</td>
</tr>
<tr>
<td>P for trend</td>
<td></td>
<td></td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>Oral contraceptive use</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never§</td>
<td>876 (46.6)</td>
<td>3920 (48.9)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Ever§</td>
<td>977 (52.0)</td>
<td>3995 (49.6)</td>
<td>1.13 (1.01–1.27)</td>
<td>1.12 (1.00–1.26)</td>
</tr>
<tr>
<td>Missing/unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postmenopausal hormone use§</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never§</td>
<td>438 (24.6)</td>
<td>2361 (52.4)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Former</td>
<td>104 (10.1)</td>
<td>419 (9.3)</td>
<td>1.30 (1.02–1.66)</td>
<td>1.29 (1.00–1.67)</td>
</tr>
<tr>
<td>Current</td>
<td>476 (46.3)</td>
<td>1742 (37.0)</td>
<td>1.62 (1.40–1.89)</td>
<td>1.38 (1.17–1.64)</td>
</tr>
<tr>
<td>Benign breast disease or fibrocystic breasts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No§</td>
<td>1205 (64.2)</td>
<td>6183 (76.8)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Yes</td>
<td>635 (33.8)</td>
<td>1702 (23.1)</td>
<td>1.95 (1.74, 2.18)</td>
<td>1.64 (1.46–1.85)</td>
</tr>
<tr>
<td>Mamnographic examination within 5 yr</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No§</td>
<td>196 (10.4)</td>
<td>1191 (14.4)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Yes</td>
<td>1657 (88.2)</td>
<td>6720 (83.9)</td>
<td>1.40 (1.20–1.64)</td>
<td>1.15 (0.96–1.38)</td>
</tr>
<tr>
<td>&lt;5 mammograms</td>
<td>495 (26.4)</td>
<td>3079 (38.8)</td>
<td>0.97 (0.81–1.18)</td>
<td>0.85 (0.70–1.04)</td>
</tr>
<tr>
<td>&gt;5 mammograms</td>
<td>1023 (54.5)</td>
<td>3328 (41.3)</td>
<td>2.14 (1.77–2.58)</td>
<td>1.64 (1.35–2.01)</td>
</tr>
<tr>
<td>Family history of breast cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No§</td>
<td>1380 (73.5)</td>
<td>6752 (83.9)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Yes</td>
<td>427 (22.7)</td>
<td>1019 (12.7)</td>
<td>2.02 (1.78–2.30)</td>
<td>1.84 (1.61–2.11)</td>
</tr>
</tbody>
</table>

(Continued)
Most women reported annual screening mammography prior to the referent date. The association between smoking history for women who were frequently screened was similar to the results overall. In multivariate models restricted to women who reported having annual mammograms in the 5-year period, ORs for breast carcinoma in situ were less than 1.0 among current smokers (OR = 0.70; 95% CI = 0.56–0.89) and former smokers (0.87; 0.74–1.03) compared with never-smokers (Table 4). Among women at least 50 years of age who reported annual mammograms, the odds ratio associated with current smoking was similar in multivariate models, both adjusted (61; 0.46–1.00) and unadjusted (0.72; 0.58–0.89) for pack years of smoking. In the models restricted to women who reported annual mammography, odds ratios were lowest among women who started smoking at younger ages, and who started smoking after their first full-term pregnancy. However, these trends were not stable nor was there a clear dose-response pattern. No consistent associations were observed among women undergoing less frequent mammogram screening during this period (OR = 0.82, 95% CI = 0.62–1.09 for current smoking; OR = 1.20, 95% CI 0.96–1.51 for past smoking).

For the relatively few women (n = 196 cases; 1191 controls) reporting no mammograms in the preceding 5 years, we observed ORs greater than 1.0 for former (1.69 [1.13–2.53]) and current (1.29 [0.85–1.96]) smoking (Table 4). In this group, smoking more than a pack per day and smoking for longer periods were associated with increased risk.

**DISCUSSION**

Cigarette smoking has rarely been considered in studies that have sought to determine whether risk factor associations are similar for invasive and in situ breast cancer.16–18 We are aware of only 3 studies that examined breast carcinoma in situ risk according to cigarette smoking. In a 1995 combined analysis of invasive and in situ breast cancer, Brinton et al reported a null association for smoking and breast cancer risk (OR = 0.56; 0.39–0.80), compared with never-smokers. In a 1999 study by Brinton et al,17 smoking more than 5 cigarettes per day and smoking for 2 to 34 years were associated with increased risk. Meanwhile, the study by Gammon et al18 observed a reduced risk of breast carcinoma in situ among never-smokers (n = 1969 cases; 986 controls; OR = 0.82; 0.69–0.97) compared with former smokers (1.20; 0.96–1.52) or current smokers (2.33; 1.78–3.07). The number of in situ cases was not reported. A 2001 study by Claus et al19 did not detect an association between smoking (ever/never) and risk of ductal cancer in situ (n = 838 cases, 986 controls; OR = 1.01; 95% CI = 0.82–1.26) or lobular cancer in situ (n = 238 cases, 986 controls; OR = 1.02; CI = 0.81–1.33). The studies by Brinton and Gammon et al did not evaluate mammographic screening; Claus’ findings adjusted

**TABLE 2. (Continued)**

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Cases (n = 1878)</th>
<th>Controls (n = 8041)</th>
<th>OR (95% CI)</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>133 (7.1)</td>
<td>632 (7.9)</td>
<td>0.89 (0.72–1.10)</td>
<td>1.04 (0.81–1.33)</td>
</tr>
<tr>
<td>High school diploma*</td>
<td>700 (37.3)</td>
<td>3138 (38.8)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Some college</td>
<td>472 (25.1)</td>
<td>2110 (26.3)</td>
<td>1.01 (0.89–1.15)</td>
<td>1.04 (0.90–1.20)</td>
</tr>
<tr>
<td>College graduate</td>
<td>551 (29.3)</td>
<td>2053 (25.6)</td>
<td>1.22 (1.07–1.38)</td>
<td>0.98 (0.84–1.13)</td>
</tr>
<tr>
<td>Weight loss/gain since age 18*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lost weight</td>
<td>73 (7.1)</td>
<td>371 (8.2)</td>
<td>1.05 (0.76–1.43)</td>
<td>1.15 (0.83–1.59)</td>
</tr>
<tr>
<td>0–15 lbs gain</td>
<td>146 (14.2)</td>
<td>758 (16.6)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>15–30 lbs gain</td>
<td>224 (21.8)</td>
<td>991 (21.6)</td>
<td>1.14 (0.91–1.44)</td>
<td>1.18 (0.93–1.50)</td>
</tr>
<tr>
<td>31–50 lbs gain</td>
<td>286 (27.8)</td>
<td>1221 (26.7)</td>
<td>1.20 (0.96–1.50)</td>
<td>1.26 (1.00–1.59)</td>
</tr>
<tr>
<td>&gt;50 lbs gain</td>
<td>269 (26.1)</td>
<td>1073 (23.4)</td>
<td>1.29 (1.03–1.61)</td>
<td>1.51 (1.19–1.92)</td>
</tr>
<tr>
<td>P for trend</td>
<td></td>
<td>0.01</td>
<td>0.00</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Alcohol consumption (drinks per week)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0†</td>
<td>341 (18.2)</td>
<td>1481 (18.5)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>&lt;7</td>
<td>1275 (67.9)</td>
<td>5467 (67.9)</td>
<td>1.03 (0.90–1.17)</td>
<td>0.93 (0.81–1.07)</td>
</tr>
<tr>
<td>7–13</td>
<td>163 (8.7)</td>
<td>707 (8.8)</td>
<td>1.02 (0.83–1.26)</td>
<td>0.90 (0.72–1.12)</td>
</tr>
<tr>
<td>≥14</td>
<td>80 (4.3)</td>
<td>295 (3.6)</td>
<td>1.24 (0.94–1.63)</td>
<td>1.15 (0.86–1.53)</td>
</tr>
<tr>
<td>P for trend</td>
<td></td>
<td>0.5</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

*Due to missing values, some categories do not sum to 100%. Percentages for controls are age-adjusted to the age distribution of the cases.

†Adjusted for age and state.

‡Adjusted for age, state, weight at age 18, height, and all factors in the table. Parity and age at first birth are not adjusted for simultaneously when evaluating the effect of either variable.

§Reference category.

¶Among postmenopausal women only (1029 cases; 4578 controls).

© 2007 Lippincott Williams & Wilkins

633

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.
for history of at least one screening mammogram in the 5-year period 1 year before the referent date.

The relation of cigarette smoking with invasive breast cancer is complex; changes in risk associated with initiation, duration, and cessation of smoking have been reported.\textsuperscript{4–6} Cigarette smoking is thought to exert conflicting forces on breast cancer risk by applying both antiestrogenic and carcinogenic effects. Within the invasive breast cancer literature, studies have reported an increase in breast cancer risk associated with initiating smoking

| TABLE 3. Association of Smoking History With Breast Carcinoma In Situ |
|--------------------------|-----------------|------------------|-----------------|
| Cigarette Smoking        | Cases (n = 1878) | Controls (n = 8041) | OR (95% CI)\textsuperscript{*} | OR (95% CI)\textsuperscript{†} |
| Smoking status           |                 |                  |                 |                 |
| Never\textsuperscript{‡} | 939             | 3752             | 1.00            | 1.00            |
| Ever                     | 919             | 4199             | 0.90 (0.81–0.99) | 0.93 (0.84–1.04) |
| Smoking status           |                 |                  |                 |                 |
| Former                   | 621             | 2482             | 1.02 (0.91–1.15) | 1.00 (0.88–1.13) |
| Current                  | 295             | 1700             | 0.72 (0.62–0.83) | 0.82 (0.70–0.96) |
| Cigarettes per day       |                 |                  |                 |                 |
| ≤10 cigarettes           | 352             | 1595             | 0.90 (0.78–1.03) | 0.89 (0.77–1.03) |
| 11–20 cigarettes         | 389             | 1819             | 0.88 (0.77–1.01) | 0.93 (0.81–1.07) |
| ≥21 cigarettes           | 172             | 739              | 0.97 (0.81–1.16) | 1.06 (0.88–1.29) |
| P for trend\textsuperscript{‡} | 0.3 | 0.9 |
| Years of smoking         |                 |                  |                 |                 |
| ≤10 yr                   | 230             | 982              | 0.97 (0.82–1.14) | 0.94 (0.79–1.11) |
| 11–20 yr                 | 208             | 868              | 1.00 (0.85–1.19) | 0.96 (0.80–1.15) |
| 21–30 yr                 | 210             | 964              | 0.87 (0.74–1.03) | 0.93 (0.76–1.10) |
| ≥31 yr                   | 268             | 1367             | 0.80 (0.69–0.93) | 0.91 (0.76–1.08) |
| P for trend\textsuperscript{‡} | 0.003 | 0.3 |
| Pack-years                |                 |                  |                 |                 |
| ≤5.25                    | 226             | 1036             | 0.89 (0.76–1.05) | 0.87 (0.74–1.04) |
| 5.25–17.24               | 249             | 1036             | 0.99 (0.85–1.16) | 0.97 (0.82–1.14) |
| 17.25–33.99              | 223             | 1008             | 0.89 (0.76–1.05) | 0.95 (0.80–1.13) |
| ≥34.00                   | 215             | 1072             | 0.83 (0.70–0.98) | 0.97 (0.82–1.16) |
| P for trend\textsuperscript{‡} | 0.2 | 0.6 |
| Age first started smoking\textsuperscript{‡} | 303 | 1518 | 0.84 (0.70–1.00) | 0.81 (0.67–0.98) |
| Age 16 or younger         | 244             | 1186             | 0.84 (0.71–1.00) | 0.81 (0.67–0.97) |
| Age 17–18                 | 192             | 653              | 1.18 (0.97–1.43) | 1.14 (0.93–1.39) |
| Age 19–20                 | 177             | 825              | 0.86 (0.71–1.04) | 0.85 (0.70–1.03) |
| Age 20 or older           |                 |                  |                 |                 |
| P for trend\textsuperscript{‡} | 1.0 | 0.9 |
| Time since cessation\textsuperscript{‡} | 295 | 1700 | 0.64 (0.53–0.77) | 0.69 (0.57–0.85) |
| Current smoker            | 154             | 655              | 0.87 (0.70–1.10) | 0.84 (0.67–1.06) |
| Smoked 1–9 yr ago         | 210             | 856              | 0.94 (0.78–1.13) | 0.88 (0.73–1.07) |
| Smoked 10–19 yr ago       | 257             | 971              | 1.05 (0.89–1.23) | 0.99 (0.84–1.17) |
| Smoked 20 or more years ago |                |                  |                 |                 |
| P for trend\textsuperscript{‡} | 0.0001 | 0.008 |
| Starting smoking in relation to first full-term birth\textsuperscript{***} | 797 | 3255 | 1.00 | 1.00 |
| Never smokers             | 699             | 3195             | 0.92 (0.82–1.03) | 0.93 (0.80–1.08) |
| Before                    | 86              | 500              | 0.70 (0.55–0.90) | 0.74 (0.57–0.96) |

\textsuperscript{*}Adjusted for age and state.
\textsuperscript{†}Adjusted for age, state, age at menarche, parity, age at first birth, oral contraceptive use, menopausal status, age at menopause, postmenopausal hormone use, weight at age 18, height, weight gain since age 18, family history of breast cancer, education, history of benign breast disease, alcohol consumption, and mammography.
\textsuperscript{‡}Reference category unless otherwise specified.
\textsuperscript{§}P value includes 0 for never smokers.
\textsuperscript{¶}P value calculated for smokers only. In time since cessation analyses, current smokers = 0 yr since cessation.
\textsuperscript{#}Among parous women only (1591 cases 7025 controls).
\textsuperscript{**}Reference category.
habits at an early age,22,23 before the first birth,6,23 and among premenopausal women in particular.22 In a 2000 report, Manjer et al24 observed an inverse relationship between invasive breast cancer risk and greater time since smoking cessation. In our study, risk estimates for breast carcinoma in situ increased toward the null with greater time since smoking cessation; however, this relation was less apparent after restricting models to women with recent

### TABLE 4. Association of Smoking History and Recent Screening Mammography With Breast Carcinoma In Situ

<table>
<thead>
<tr>
<th>Smoking status</th>
<th>No Screening (n = 196 Cases, 1191 Controls)</th>
<th>Fewer Than Annual Mammograms (n = 495 Cases, 3079 Controls)</th>
<th>Annual Mammograms (n = 1023 Cases, 3328 Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR (95% CI)*</td>
<td>OR (95% CI)*</td>
<td>OR (95% CI)*</td>
</tr>
<tr>
<td>Never*</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Ever</td>
<td>1.48 (1.05–2.10)</td>
<td>1.04 (0.85–1.28)</td>
<td>0.82 (0.70–0.95)</td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Former</td>
<td>1.69 (1.13–2.53)</td>
<td>1.20 (0.96–1.51)</td>
<td>0.87 (0.74–1.03)</td>
</tr>
<tr>
<td>Current</td>
<td>1.29 (0.85–1.96)</td>
<td>0.82 (0.62–1.09)</td>
<td>0.70 (0.56–0.89)</td>
</tr>
<tr>
<td>Cigarettes per day</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤10 cigarettes</td>
<td>1.57 (1.02–2.43)</td>
<td>0.99 (0.76–1.30)</td>
<td>0.75 (0.62–0.92)</td>
</tr>
<tr>
<td>11–20 cigarettes</td>
<td>1.18 (0.75–1.84)</td>
<td>1.07 (0.82–1.38)</td>
<td>0.88 (0.72–1.07)</td>
</tr>
<tr>
<td>≥21 cigarettes</td>
<td>2.31 (1.34–3.99)</td>
<td>1.15 (0.81–1.65)</td>
<td>0.87 (0.66–1.16)</td>
</tr>
<tr>
<td>P for trend‡</td>
<td>0.007</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>Years of smoking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤10 yr</td>
<td>1.07 (0.64–1.79)</td>
<td>1.11 (0.83–1.51)</td>
<td>0.82 (0.64–1.05)</td>
</tr>
<tr>
<td>11–20 yr</td>
<td>2.05 (1.21–3.47)</td>
<td>1.09 (0.78–1.51)</td>
<td>0.78 (0.61–1.01)</td>
</tr>
<tr>
<td>21–30 yr</td>
<td>1.31 (0.75–2.29)</td>
<td>1.00 (0.73–1.40)</td>
<td>0.86 (0.67–1.10)</td>
</tr>
<tr>
<td>≥31 yr</td>
<td>1.84 (1.06–3.17)</td>
<td>0.97 (0.69–1.36)</td>
<td>0.83 (0.67–1.03)</td>
</tr>
<tr>
<td>P for trend‡</td>
<td>0.02</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Pack-years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;5.25</td>
<td>1.17 (0.70–1.95)</td>
<td>1.00 (0.74–1.35)</td>
<td>0.74 (0.58–0.95)</td>
</tr>
<tr>
<td>5.25–17.24</td>
<td>1.60 (0.96–2.67)</td>
<td>1.22 (0.90–1.64)</td>
<td>0.81 (0.64–1.02)</td>
</tr>
<tr>
<td>17.25–33.99</td>
<td>1.28 (0.73–2.22)</td>
<td>0.97 (0.71–1.34)</td>
<td>0.90 (0.71–1.14)</td>
</tr>
<tr>
<td>≥34.00</td>
<td>2.22 (1.32–3.74)</td>
<td>1.01 (0.71–1.44)</td>
<td>0.87 (0.68–1.11)</td>
</tr>
<tr>
<td>P for trend‡</td>
<td>0.01</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Age first started smoking‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 16 or younger</td>
<td>1.48 (0.88–2.48)</td>
<td>0.80 (0.56–1.13)</td>
<td>0.73 (0.56–0.95)</td>
</tr>
<tr>
<td>Age 17–18</td>
<td>0.91 (0.51–1.60)</td>
<td>0.94 (0.67–1.32)</td>
<td>0.72 (0.56–0.92)</td>
</tr>
<tr>
<td>Age 19–20</td>
<td>1.98 (1.09–3.60)</td>
<td>1.55 (1.09–2.23)</td>
<td>0.84 (0.63–1.12)</td>
</tr>
<tr>
<td>Age 20 or older</td>
<td>1.06 (0.56–2.02)</td>
<td>1.03 (0.72–1.48)</td>
<td>0.74 (0.56–0.97)</td>
</tr>
<tr>
<td>P for trend‡</td>
<td>0.5</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Time since cessation‡</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current</td>
<td>0.97 (0.57–1.65)</td>
<td>0.66 (0.46–0.95)</td>
<td>0.57 (0.42–0.76)</td>
</tr>
<tr>
<td>Smoked 1–9 yr ago</td>
<td>1.40 (0.75–2.60)</td>
<td>0.71 (0.44–1.16)</td>
<td>0.76 (0.55–1.05)</td>
</tr>
<tr>
<td>Smoked 10–19 yr ago</td>
<td>1.78 (1.01–3.12)</td>
<td>1.13 (0.80–1.59)</td>
<td>0.68 (0.52–0.90)</td>
</tr>
<tr>
<td>Smoked ≥20 yr ago</td>
<td>1.11 (0.51–2.43)</td>
<td>1.30 (0.96–1.77)</td>
<td>0.85 (0.68–1.06)</td>
</tr>
<tr>
<td>P for trend‡</td>
<td>0.2</td>
<td>0.02</td>
<td>0.1</td>
</tr>
<tr>
<td>Starting smoking in relation to first full term birth§</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smokers§</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Before</td>
<td>1.17 (0.73–1.86)</td>
<td>1.04 (0.80–1.37)</td>
<td>0.83 (0.67–1.02)</td>
</tr>
<tr>
<td>After</td>
<td>1.18 (0.57–2.47)</td>
<td>0.92 (0.56–1.53)</td>
<td>0.63 (0.43–0.91)</td>
</tr>
</tbody>
</table>

*Adjusted for age, state, age at menarche, parity, age at first birth, oral contraceptive use, menopausal status, age at menopause, postmenopausal hormone use, weight at age 18, height, weight gain since age 18, family history of breast cancer, education, history of benign breast disease, and alcohol consumption.

†Reference category unless otherwise specified.

‡P value includes 0 for never smokers.

§Also adjusted for pack-years.

§P value calculated for smokers only. In time since cessation analyses, current smokers = 0 yr since cessation.

‖Among parous women only (1591 cases, 7025 controls).

*Reference category.
annual mammograms. Additionally, women who started smoking at younger ages or after a first birth had the lowest odds ratios for breast carcinoma in situ, compared with never-smokers.

Differences in the observed relations suggest that selection bias may have influenced the association between smoking and risk of breast carcinoma in situ. Smoking may be associated with both the frequency of health screening behaviors and the willingness to participate in health surveys. National surveys have demonstrated that current smokers are half as likely to take advantage of cancer screening such as mammography. In the current study, control women who smoked cigarettes were substantially less likely to receive routine mammography than former and never-smokers. The diagnosis of breast carcinoma in situ is more common in women with higher socioeconomic status. Differential response rates according to smoking history and breast screening may have distorted associations of smoking and breast carcinoma in situ risk in the current study. Although participation rates in both cases and controls were relatively high, and our risk estimates were adjusted for several factors that are often correlated with health-seeking behavior (such as educational level and postmenopausal hormone therapy), the results may still be influenced by residual bias.

Previous research has suggested that studies of breast carcinoma in situ should most appropriately use data from women who have undergone mammographic screening to address this concern. The presence of undiagnosed breast carcinoma in situ within the control group potentially could bias associations towards the null. Because odds ratio estimates among unscreened women were greater than 1.0, the influence of selection bias due under-participation from control women who smoke is likely greater than any bias due to undiagnosed disease among controls. The lack of a detectable dose-response according to the amount, duration, or recency of smoking among women who reported annual mammograms also reinforces concerns that the data were affected by bias.

Chance or uncontrolled confounding as an explanation for our findings cannot be ruled out. Alternatively, it may be possible that the long-term carcinogenic effects of smoking early in life dominate in relation to invasive breast cancer risk, while the more immediate antiestrogenic effects of current or recent smoking dominate the relation with risk of in situ breast cancer. The window of exposure is likely important; full differentiation of breast tissue occurs following the first full-term pregnancy and lactation, and may result in a reduced susceptibility to tumor initiating effects. Researchers have theorized that the carcinogenic effects of cigarette smoke are maximized in premenopausal women who initiate smoking around menarche and before the first birth (a period of high endogenous estrogen when antiestrogenic effects associated with cigarette smoking would be relatively weak), while antiestrogenic effects would be most pronounced among postmenopausal women who commenced smoking after the first full-term birth, especially those who have gained weight in adulthood. In postmenopausal women, endogenous estrogen levels are in part derived from androgens in adipose tissue, and may be more sensitive to antiestrogenic effects of cigarette smoke. In our data, we observed lower odds ratios for breast carcinoma in situ among women who started smoking before age 19, or during a period of potential high susceptibility for undifferentiated breast tissue. However, among both postmenopausal and premenopausal women, odds ratios for breast cancer were also lower among women who initiated smoking after their first full-term birth, relative to parous never-smokers.

The antiestrogenic effect attributed to cigarette smoking may vary according to exogenous hormone use, menopausal status, or body mass index. We did not observe effect modification by these factors within our study population. However, our results did suggest that the association between risk of breast carcinoma in situ and smoking status depended on mammography screening behavior. Reductions in breast carcinoma in situ risk associated with smoking status were limited to women who received a mammogram in the 5-year period 1 year before breast cancer diagnosis in the cases or comparable period in controls. In contrast, former smoking was associated with elevated odds ratios among women reporting no mammograms during this period. As noted above, selection bias may have influenced the results, especially among women who did not report any recent mammograms. In our study, case women were more likely than controls to report annual mammograms in the 5 years before diagnosis. This is consistent with the evidence that breast carcinoma in situ is usually diagnosed through mammography (78% of cases in the current study), and women undergoing regular breast screening would be more likely to have a prevalent cancer diagnosed.

Women ages 20–74 years participated in our study; during this period screening was advised only for women ages 50 and older. In some circumstances, women may have included diagnostic mammography when reporting frequency of screening mammograms, which would contribute to an elevated risk associated with the procedure. To ensure that screening mammograms reported by younger women were not a reflection of subclinical disease symptoms, detection of a lump, or a family history, we restricted smoking analyses to women ages 50 and older who had had at least one mammogram in the 5-year period of interest. In this group, the relation of current smoking and decreased risk of breast carcinoma in situ remained unchanged. We also performed analyses of age at smoking initiation, time since cessation, and smoking in relation to the first full-term birth both adjusted and unadjusted for pack-years. In the full study population, the pattern of association was unaltered.

Associations observed in this study with factors other than smoking are similar to previously published reports. Our results are consistent regarding increased risk of breast carcinoma in situ associated with older ages at first birth and at menopause, a personal history of benign breast disease, a family history of breast cancer, and higher education. Similar associations are observed for invasive breast cancer risk. Two recent case-control studies reported that
women with breast cancer in situ were less likely to be postmenopausal than similarly-aged controls.\textsuperscript{29,32} A previous study reported an inverse association between body mass index and risk of breast carcinoma in situ;\textsuperscript{17} while our findings did not indicate an association with body mass index, we did observe an increase in situ cancer risk in relation to weight gain. We observed a null association for risk of breast cancer in situ in relation to alcohol consumption as have other investigators.\textsuperscript{17,21,33}

Our study findings are strengthened by the large study population and the use of standardized instruments, including self-reported mammographic screening information. In addition, we were able to categorize a wide range of smoking measures\textsuperscript{6} and adjust for known breast cancer risk factors. However, some studies have suggested that not accounting for passive smoking could weaken the assumed positive association between (invasive) breast cancer and smoking,\textsuperscript{22} and we were unable to account for exposure to passive cigarette smoke. Recall bias is an inherent concern with the use of self-reported, retrospective data. Reassuringly, self-reported lifetime smoking has been highly reproducible in other studies and we have previously established the strong reliability of covariate data from our survey instrument.\textsuperscript{34–37}

In conclusion, our findings suggest that current smoking is inversely associated with risk of breast carcinoma in situ. This association was restricted to women who reported frequent screening mammography. The interaction with mammography frequency reaffirms the importance of capturing screening information in studies of breast carcinoma in situ to adequately address the potential for selection bias and the necessity of an appropriate comparison group.

**ACKNOWLEDGMENTS**

We thank Henry Anderson, Patrick L. Remington, Meir J. Stamper, Walter C. Willett, John A. Baron, and E. Robert Greenberg; Laura Stephenson and the staff of the Wisconsin Cancer Reporting System; Susan T. Gershman and the staff of the Massachusetts Tumor Registry; Marguerite Stevens and the staff of the New Hampshire Cancer Registry; and Mary Pankratz, Linda Haskins, Jerry Phipps, Heidi Judge, Laura Mignone, and Shafika Abrahams-Gessel, along with the study interviewers in all 3 states for assistance with data collection. We are especially grateful to the study participants, whose generosity made this research possible.

**REFERENCES**


Cigarette Smoking and Lung Cancer
Modeling Effect Modification of Total Exposure and Intensity

Jay H. Lubin,* Neil Caporaso,† H. Erich Wichmann,‡ Angelika Schaffrath-Rosario,§ and Michael C. R. Alavanja¶

Background: A recent analysis indicates that the excess odds ratio for lung cancer by smoking is described by a function that is linear in pack-years and exponential in the logarithm of smoking intensity and its square (Cancer Epidemiol Biomarkers Prev. 2006;15:517–523). The model suggests that below 15–20 cigarettes per day there is a “direct exposure rate” effect, ie, the excess odds ratio per pack-year for higher intensity (and shorter duration) smokers is greater than for lower-intensity (and longer duration) smokers. Above 20 cigarettes per day, there is an “inverse-exposure-rate” effect, ie, the excess odds ratio per pack-year for higher intensity smokers is smaller than for lower-intensity smokers.

Methods: Using pooled data from 2 large case-control studies of lung cancer (the European Smoking and Health Study and the German Radon Study), we evaluated effect modification of the association between smoking and lung cancer.

Results: Interaction effects are very specific. Variations in risk of lung cancer with years since cessation of smoking, age, method of inhalation, and type of cigarette result from interactions with smoking intensity, and not total pack-years. In contrast, risk variations by sex result from the interaction with total pack-years, while intensity effects are homogeneous. Risk variations by age at which smoking started result from interactions with both total pack-years and intensity. All intensity interactions are homogeneous across studies.

Conclusions: The specificity of the interactions may provide clues for the molecular basis of the smoking and lung cancer relationship.

(Epidemiology 2007;18: 639–648)

Understanding the relationship of lung cancer and cigarette smoking and modifiers of that relationship provides a framework for evaluating the molecular basis of carcinogenesis, including activation and detoxification of tobacco carcinogens and DNA repair. A more detailed understanding of etiology may improve risk assessment and public health decision-making. Specifically, in studies of lung cancer and bladder cancer, odds ratios (ORs) increase with smoking intensity (cigarettes smoked per day), but then often level off at high intensities. This pattern has been described but not adequately quantified.

Models that incorporate smoking duration and intensity are problematic for exploring this pattern due to changing total exposures. For example, in a logistic model, the intensity parameter represents the In/OR per cigarette per day at a fixed duration. A similar interpretation applies for the duration parameter at a fixed intensity. Because duration is fixed, ORs at 2 different intensities therefore reflect not only the different intensities but also different total pack-years. For 30 years' duration, ORs at 20 and 30 cigarettes per day embed effects of different total exposures, ie, 30 and 45 pack-years, respectively. Thus, the intensity parameter does not represent a “pure” intensity effect, but includes the effect of pack-years. In contrast, ORs for 20 and 30 cigarettes per day obtained from a model that includes pack-years and intensity reflect on intensity effect unconfounded by total exposure.

Using data on never, current, and recent (within 5 years) former smokers, investigators have developed an excess odds ratio (EOR) model for lung cancer. The model is linear in pack-years and exponential in the logarithm of intensity and its square, which quantifies the effects of intensity on the EOR per pack-year. The model isolates the intensity effects for fixed total pack-years, thus enabling the comparison of ORs for total exposure delivered at low intensity (for long duration) and at high intensity (for short duration). Below 15–20 cigarettes per day, there is a direct exposure rate effect (increased potency or exposure enhancement effect), whereby the EOR per pack-year increases with increasing intensity, ie, for equal pack-years, increasing intensity (decreasing duration) increases risk. Above 20 cigarettes per day, there is an inverse exposure rate effect (reduced potency or wasted exposure effect), whereby the EOR per pack-year decreases with increasing intensity, ie, for equal pack-years increasing intensity decreases risk.
Using pooled data from 2 large studies of lung cancer, we evaluated variations of the linear pack-years effect and exponential intensity effects by years since smoking cessation, attained age, age started smoking, sex, inhalation and type of cigarette (filter or nonfilter/mixed) smoked.

**METHODS**

**Studies of Smoking and Lung Cancer**

The European Smoking and Health Study was a hospital-based case-control study of lung cancer conducted between 1976 and 1980 at 7 European centers (Glasgow, Hamburg, Heidelberg, Vienna, Paris, Milan, and Rome).4,5 The study enrolled 7,804 cases and 15,207 controls, frequency-matched by age, sex, and center. We excluded 712 subjects who smoked cigars or pipes, 199 subjects who started smoking after age 40, and 36 subjects with missing data. The basic dataset included 4988 cases (4336 men and 652 women) and 8389 controls (7142 men and 1247 women), who were age 50 to 75 years. Participants were never-smokers, current smokers, or former smokers who stopped smoking within 5 years of enrollment.

The German Radon Study was a population-based lung cancer case–control study conducted between 1990 and 1997 in 23 regions of Germany.6–8 The study enrolled 4071 cases age 75 years and younger and 4628 controls, frequency-matched on age, sex, and region. We excluded smokers who started smoking after age 40, smoked cigars or pipes, or had temporarily stopped smoking for more than 5 years, leaving 3212 cases (2589 men and 614 women) and 3809 controls (2932 men and 877 women). The basic dataset included 2256 cases (1796 men and 460 women) and 2157 controls (1512 men and 645 women) age 50 to 75, who were never-smokers, current smokers, or former smokers who stopped smoking within 5 years of enrollment. Initial analysis indicated that the inclusion of 158 men (70 cases and 88 controls) who smoked up to 8 cigarettes per day resulted in models that slightly underestimated the EOR per pack-year at moderate and high intensities. Although inference is unaffected, we omitted these subjects.

We imposed the age restriction to reduce the potential impact of any genetic cancer predisposition in younger cases or diagnostic ambiguity in elderly cases, and omitted cigar or pipe smokers to allow consistent analyses of cigarette exposure. Both studies were approved by appropriate internal review boards.

**Models**

We defined I intensity categories and indicator variables $n_i$, $i = 1, \ldots, I$, where $n_i = 1$ for intensities within the $i$th category and zero otherwise and fit the model

$$\text{OR}(d) = 1 + \sum \gamma_i n_i d$$

(1)

where $d$ is pack-years. Within category $i$, ORs are linear in pack-years (ie, OR = 1 + $\gamma_i d$). The slope $\gamma_i$ defines the EOR per pack-year. For calculating estimates, $\exp(\gamma_i)$ replaces $\gamma_i$. Factoring out $\gamma_i$, model (1) becomes

$$\text{OR}(d) = 1 + \gamma_i d \sum \left(\gamma_i/\gamma_t\right) n_t$$

Thus, a natural modeling for continuous intensity, $n$, is

$$\text{OR}(d) = 1 + \beta d g(n)$$

(2)

where $g(.)$ represents intensity effects and $\beta$ represents the EOR per pack-year at $g(n) = 1$. We set $g(n) = \exp\{d_1 \ln(n) + d_2 \ln(n)^2\}$. This form provided a better fit than $g(n) = \exp\{d_1 \ln(n) + d_2\}$ or $g(n) = \exp\{d_1 n + d_2 n^2\}$. Adding $\ln(n) \times \ln(n)^2$ or $n^2$ did not further improve fit. The parameters $d_1$ and $d_2$ define the modulation of the EOR per pack-year with intensity, and their relative size identifies the maximum EOR per pack-year, $\beta g(n_{\text{max}})$, where for $d_1 > 0$ and $d_2 < 0$ $n_{\text{max}} = \exp(-d_1/2d_2)$. The first derivative of the EOR per pack-year function, $d[\beta g(n)]/dn = \beta g(n)(d_1 + 2d_2 n \ln(n))\ln(n)$, describes changes in the function with intensity.

We evaluated effect modification by graphing models (1) and (2) within categories of the factor; we tested statistical significance using variants of the EOR in model (2), namely, $\beta g(n_0) + \beta_1 d g(n)$, $\beta_1 d g(n)$, and $\beta_2 d g(n)$, where “$f$” denotes separate parameter ($\beta$) or set of parameters ($d_1$ and $d_2$) for each level of the factor. The difference of deviances for nested models defines a test of homogeneity of the pack-year and intensity effects across levels of the factor. Degrees of freedom equal the difference in numbers of parameters.

We evaluated homogeneity over study, $s$, using $\beta_{s,f} d \ln(g(n))$, where “$s\times f$” denotes all levels of s and f. Similarly, “$s,f$” denotes separate parameters for levels of f and one additional parameter for study effect. For all factors, intensity effects are homogeneous across studies, ie, inclusion of $g_{s,f}(n)$ or $g(n)$ provides comparable fit. We therefore present results based on $\beta_{s,f} d g(n)$. For the figures, we summarized models by setting the study variable to values +1 (ESHS) or −1 (GRS).

We used the binary outcome module in Epicure (Hirossoft International Corp., Seattle, WA), with stratification on center/region, sex, and age (five levels: 50–54, . . ., 70+).

**RESULTS**

Among never-smokers and cigarette-only smokers, percentages of ever-smokers for male and female cases and controls were similar in both studies. The German study had fewer current smokers, whereas smokers in the European study consumed at higher intensities for longer durations (Table 1).

**Model for Pack-Years and Smoking Intensity**

Applying the analytic methods used previously with the European study, we computed ORs for the German study by categories of pack-years and intensity relative to never-smokers. ORs by pack-years are approximately linear within intensity categories. Slope estimates ($\gamma_i$’s) vary with intensity. As seen below, the fitted model (2) closely conforms to the $\gamma_i$ estimates. Study-specific estimates of EOR per pack-year ($\beta$) differ significantly ($P < 0.01$), whereas homogeneity of $d_1$ and $d_2$ over study is not rejected ($P = 0.17$), indicating similar intensity patterns (Table 2). Estimates of $\beta$ are 0.0047 for the European study and 0.0146 for the German study, and summary estimates of $d_1$ and $d_2$ are 2.86 and −0.495, respectively.
TABLE 1. Summary of Smoking Data* for Case-Control Studies by Sex

<table>
<thead>
<tr>
<th></th>
<th>% Ever Smoked</th>
<th>% Current Smokers(^1)</th>
<th>Duration (Mean No. Years)</th>
<th>Intensity (Mean No. Cigarettes/d)</th>
<th>Mean Pack-Years</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>European Smoking and Health Study</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases (n = 4958)</td>
<td>97.7</td>
<td>75.3</td>
<td>41.3</td>
<td>23.6</td>
<td>48.6</td>
</tr>
<tr>
<td>Controls (n = 9081)</td>
<td>78.1</td>
<td>62.0</td>
<td>37.2</td>
<td>20.3</td>
<td>37.8</td>
</tr>
<tr>
<td>Women</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases (n = 683)</td>
<td>62.4</td>
<td>84.5</td>
<td>39.3</td>
<td>17.5</td>
<td>34.7</td>
</tr>
<tr>
<td>Controls (n = 1320)</td>
<td>31.0</td>
<td>71.4</td>
<td>34.4</td>
<td>14.6</td>
<td>25.3</td>
</tr>
<tr>
<td><strong>German Radon Study</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases (n = 2308)</td>
<td>97.9</td>
<td>67.6</td>
<td>38.9</td>
<td>19.4</td>
<td>37.9</td>
</tr>
<tr>
<td>Controls (n = 2613)</td>
<td>69.3</td>
<td>32.4</td>
<td>28.8</td>
<td>15.3</td>
<td>23.4</td>
</tr>
<tr>
<td>Women</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cases (n = 518)</td>
<td>60.8</td>
<td>74.3</td>
<td>34.7</td>
<td>14.8</td>
<td>26.9</td>
</tr>
<tr>
<td>Controls (n = 753)</td>
<td>29.5</td>
<td>43.7</td>
<td>26.4</td>
<td>10.8</td>
<td>15.9</td>
</tr>
</tbody>
</table>

*Includes never and cigarette-only smokers between ages 50 and 75 yr.
\(^1\)Current smokers or stopped smoking within 2 yr of enrollment.

Evaluation of Effect Modification

Table 2 shows estimates for the \( \beta_{x \times f} d_g(n) \) model, maxima for the EOR functions, deviance changes relative to the study-specific \( \beta_g d_g(n) \) model, and \( P \) values for 4 tests of interaction for pack-years and 3 tests of interaction for intensity and \( f \) (see table footnote).

We added former smokers to the basic dataset to evaluate smoking cessation. Effect modification by time since smoking cessation derives from an interaction with intensity and not pack-years. The intensity and cessation interaction is statistically significant, regardless of the adjustment for the pack-years and cessation interaction \( (P = 0.01) \), comparing \( \beta_{x \times f} d_g(n) \) with \( \beta_g d_g(n) \); \( P = 0.04 \), comparing \( \beta_{x \times f} d_g(n) \) with \( \beta_{x \times g} d_g(n) \); and \( P < 0.01 \), comparing \( \beta_{x \times f} d_g(n) \) with \( \beta_{f \times g} d_g(n) \). Adjusting for the intensity and cessation interaction, the pack-years and cessation interaction is not statistically significant \( (P = 0.14) \), comparing \( \beta_{x \times f} d_g(n) \) with \( \beta_g d_g(n) \); and \( P = 0.78 \), comparing \( \beta_{x \times f} d_g(n) \) with \( \beta_{g \times f} d_g(n) \). The model \( \beta_g d_g(n) \) closely describes the intensity-specific EOR per pack-year estimates \( (\gamma_i) \) within smoking cessation categories (Fig. 1). First derivatives suggest that differences in EOR per pack-year functions result from a reduced modulation with intensity by increasing time since cessation (Fig. 1D).

We added subjects under age 50 to the basic dataset to assess age. The intensity and age interaction is significant, after adjusting for the pack-years and age interaction (Table 2). Controlling for the intensity and age interaction, pack-year effects are statistically homogeneous across age groups. As with cessation, differences in the EOR per pack-year functions (Fig. 2) result from differential modulation with intensity by age, primarily a reduced variation in the youngest age group.

Interactions between the age at which smoking began and intensity or pack-years are each statistically significant after controlling for the other interaction. Parameter estimates and fitted EOR per pack-year maxima are different for people who start smoking at young ages, particularly at very low intensities of smoking (Fig. 3).

Differences in the EOR for smoking by sex result from variations in total pack-years rather than intensity. Controlling for the interaction of pack-years and sex, the interaction of intensity and sex was not statistically significant \( (P = 0.30) \), while the interaction of pack-years and sex was significant (Fig. 4). Notably, the interaction by pack-years and sex differs by study (Table 2: \( \chi^2(1) = 47.1–38.5 = 8.6 \) comparing \( \beta_{x \times f} d_g(n) \) with \( \beta_{x \times f} d_g(n) \); \( P < 0.01 \)). Estimates from the model \( \beta_{x \times f} d_g(n) \) are \( \beta_{ESHS,male} = 0.0121, \beta_{ESHS,female} = 0.0062, \beta_{GRS,male} = 0.0507, \beta_{GRS,female} = 0.0109, \phi_1 = 2.43 \) and \( \phi_2 = -0.432 \). Relative pack-year effects for men and women are 2.0 in the European study and 4.7 in the German study.

Effect modification by frequency of inhalation (data only from the European study) and depth of inhalation results from variations with intensity and not pack-years. The frequency of inhalation and intensity interaction is statistically significant \( (P < 0.01) \), even after controlling for the interaction of frequency and pack-years \( (P = 0.03) \). Controlling for intensity, the interaction of pack-year and frequency of inhalation is not significant \( (P = 0.78) \). Results are similar for depth of inhalation. Differences in the EOR per pack-year by inhalation pattern derive from a reduced modulation with intensity in less vigorous inhalers (Fig. 5). More frequent and deeper inhalers have higher EOR per pack-year maxima that occur at lower intensities.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimates for $β_{\text{ESHS}, t}$, $d$, $g(n)$</th>
<th>Maximum EOR</th>
<th>Evaluation of Effect Modifier ($f$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$β_{\text{ESHS}, t}$, $β_{\text{GRS}, t}$, $Φ_{t1}$, $Φ_{t2}$, $n_{\text{max}}$, ESHS, GRS</td>
<td>$β_{d} d(g(n))$</td>
<td>$P (β)^{b}$, $P (g)^{f}$</td>
</tr>
<tr>
<td>Overall</td>
<td>0.0047, 0.0146, 2.86, -0.495, 18.0, 0.29, 0.91</td>
<td>375.0 (12)</td>
<td>0.14, 0.01</td>
</tr>
<tr>
<td>ESHS</td>
<td>0.0072, 0.0071, 3.14, -0.504, 22.5, 0.30, 0.94</td>
<td>360.2 (8)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>GRS</td>
<td></td>
<td>368.5 (10)</td>
<td>0.08, 0.04</td>
</tr>
<tr>
<td>Years since cessation$^{h}$</td>
<td></td>
<td>358.5 (6)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>&lt;5</td>
<td>0.0049, 0.0150, 2.81, -0.485, 18.1, 0.29, 0.88</td>
<td>42.8 (16)</td>
<td>0.09, 0.01</td>
</tr>
<tr>
<td>5–14</td>
<td>0.0046, 0.0107, 2.72, -0.501, 15.1, 0.19, 0.43</td>
<td>20.8 (10)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>15+</td>
<td>0.0017, 0.0033, 3.50, -0.716, 11.5, 0.12, 0.24</td>
<td>35.1 (13)</td>
<td>0.35, 0.02</td>
</tr>
<tr>
<td>Age (yrs)$^{a}$</td>
<td></td>
<td>13.9 (7)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>&lt;50</td>
<td>0.0005, 0.0043, 3.42, -0.523, 26.3, 0.14, 1.15</td>
<td>25.2 (16)</td>
<td>0.06, 0.03</td>
</tr>
<tr>
<td>50–59</td>
<td>0.0010, 0.0031, 3.68, -0.593, 22.3, 0.31, 0.92</td>
<td>11.9 (10)</td>
<td>0.06, 0.04</td>
</tr>
<tr>
<td>60–69</td>
<td>0.0070, 0.0222, 2.82, -0.517, 15.2, 0.32, 1.02</td>
<td>23.8 (13)</td>
<td>0.01, 0.04</td>
</tr>
<tr>
<td>70–74</td>
<td>0.0111, 0.0342, 2.27, -0.427, 14.3, 0.23, 0.70</td>
<td>10.9 (7)</td>
<td>0.01, 0.04</td>
</tr>
<tr>
<td>Age started smoking (yrs)</td>
<td></td>
<td>12.9 (10)</td>
<td>0.04</td>
</tr>
<tr>
<td>&lt;15</td>
<td>0.1365, 0.3786, 0.87, -0.209, 7.9, 0.33, 0.93</td>
<td>47.1 (8)</td>
<td>0.002, 0.30</td>
</tr>
<tr>
<td>15–19</td>
<td>0.0037, 0.0113, 3.02, -0.517, 18.5, 0.30, 0.93</td>
<td>44.7 (6)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>20–24</td>
<td>0.0116, 0.0367, 2.21, -0.370, 19.7, 0.31, 0.98</td>
<td>38.5 (7)</td>
<td>0.06, 0.30</td>
</tr>
<tr>
<td>25+</td>
<td>0.0001, 0.0002, 5.71, -0.962, 19.4, 0.27, 0.96</td>
<td>36.1 (5)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td>35.0 (6)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Males</td>
<td>0.0156, 0.0656, 2.26, -0.402, 16.5, 0.37, 1.56</td>
<td>21.6 (9)</td>
<td>0.78, 0.03</td>
</tr>
<tr>
<td>Females</td>
<td>0.0002, 0.0004, 5.01, -0.909, 15.7, 0.21, 0.37</td>
<td>11.2 (5)</td>
<td>0.01</td>
</tr>
<tr>
<td>Frequency of inhalation (ESH only)</td>
<td></td>
<td>21.1 (7)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Rarely/part of the time</td>
<td>0.0066, 2.18, -0.325, 28.9, 0.26</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Most of the time</td>
<td>0.0475, 1.41, -0.270, 13.6, 0.30</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>All of the time</td>
<td>0.0120, 2.41, -0.442, 15.3, 0.32</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Depth of inhalation</td>
<td></td>
<td>20.3 (8)</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Slightly/never</td>
<td>0.0028, 0.0087, 2.91, -0.466, 22.7, 0.26, 0.81</td>
<td>20.8 (11)</td>
<td>0.92, 0.01</td>
</tr>
<tr>
<td>Moderately</td>
<td>0.0153, 0.0487, 1.98, -0.330, 20.0, 0.30, 0.94</td>
<td>8.3 (7)</td>
<td>0.04</td>
</tr>
<tr>
<td>Deeply$^{**}$</td>
<td>0.0170, 2.24, -0.426, 13.8, 0.32</td>
<td>20.8 (10)</td>
<td>0.78, 0.01</td>
</tr>
<tr>
<td>Type of cigarette brand</td>
<td></td>
<td>8.2 (6)</td>
<td>0.02</td>
</tr>
<tr>
<td>Filter only</td>
<td>0.0229, 0.0758, 1.28, -0.178, 36.0, 0.23, 0.75</td>
<td>21.4 (8)</td>
<td>0.55, 0.12</td>
</tr>
<tr>
<td>Mixed/nonfilter</td>
<td>0.0050, 0.0152, 2.86, -0.490, 18.6, 0.33, 1.00</td>
<td>17.2 (6)</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

*Model: EOR = $β_{s1} d g(n) (e)$ with $g(n) = \exp(\phi_1 \log(n) + \phi_2 \log(n)^2)$, where $d$ is pack-years, $n$ is cigarettes smoked per day. Subscripts denote separate effects for study (s) and/or factor (t). All models adjust for study or center, age and sex.

†Maximum EOR/pack-year, $β_{s1} g(n) (n_{\text{max}})$, where $n_{\text{max}} = \exp(-\phi_2/2\phi_1)$ is the number of cigarettes smoked per day at the maximum.

‡Deviance change relative to the 4-parameter model $β_{d} d g(n)$ and number of parameters in the EOR.

§Values for likelihood ratio tests: $β_{s1} d g(n)$ with $β_{s1} d g(n)$; $β_{s2} d g(n)$ with $β_{s2} d g(n)$; $β_{s3} d g(n)$ with $β_{s3} d g(n)$; $β_{s4} d g(n)$ with $β_{s4} d g(n)$.

||$P (β)^{b}$, $P (g)^{f}$|
FIGURE 1. (A–C) Estimated excess odds ratio (EOR) per pack-year for categories of smoking intensity (square symbol) within categories of years since cessation of smoking and the fitted EOR per pack-year based on the model $\beta_n g(n)$ (solid line) and the pointwise 95% prediction interval (dashed line). Results are summarized over study. (D) first derivatives of the fitted models.

FIGURE 2. (A–D) Estimated excess odds ratio (EOR) per pack-year for categories of smoking intensity (square symbol) within categories of attained age and the fitted EOR per pack-year based on the model $\beta_n g(n)$ (solid line) and the pointwise 95% prediction interval (dash line). Results are summarized over study. (E) First derivatives of the fitted models.
FIGURE 3. (A–D) Estimated excess odds ratio (EOR) per pack-year for categories of smoking intensity (square symbol) within categories of age started smoking and the fitted EOR per pack-year based on the model \( \beta_{\text{sex}} g(n) \) (solid line) and the pointwise 95% prediction interval (dashed line). Results are summarized over study. (E) First derivatives of the fitted models.

FIGURE 4. (A, B) Estimated excess odds ratio (EOR) per pack-year for categories of smoking intensity (square symbol) within sex and the fitted EOR per pack-year based on the model \( \beta_{\text{sex}} g(n) \) (solid line) and the pointwise 95% prediction interval (dashed line). Results are summarized over study. (C) First derivatives of the fitted models.
The interaction of intensity and type of cigarette (filter-only or nonfilter-only/mixed) is statistically significant \((P < 0.1)\), although p-values diminish after controlling for the interaction of pack-years and type of cigarette \((P = 0.12\) using \(\beta_{\text{nf,f}}\) and \(P = 0.13\) using \(\beta_{\text{s,f}}\)). Controlling for the interaction of intensity and type of cigarette, the interaction of pack-years and type of cigarette is not significant. Analysis suggests that effect modification with type of cigarette results from variation with smoking intensity (Fig. 6).

Finally, first-derivative plots showed that effect modifiers have their greatest impact on the EOR per pack-year function under 20–25 cigarettes per day.

**DISCUSSION**

Variations in smoking-related ORs with time since smoking cessation, age, frequency and depth of inhalation, and type of cigarette derive from interactions with smoking intensity, rather than pack-years. In contrast, variations in smoking-related ORs with sex result from interactions with total pack-years, while smoking intensity effects are similar in men and women. Variations in ORs with the age at which smoking began result from interactions with both intensity and pack-years.

The modeling of the EOR per pack-year suggests enhanced carcinogenic potency of cigarettes at low intensities and a relatively reduced potency at moderate and high intensities, both overall and within levels of effect modifiers. The latter pattern parallels relationships reported in biomarker studies of smoking.\(^2,9,10\) Polycyclic aromatic hydrocarbons (PAH), many of which are known to be carcinogenic, occur in tobacco smoke, and after metabolic activation they form DNA and protein adducts.\(^11\) Individuals exposed at environmental levels have higher DNA-adduct levels in white blood cells per unit exposure to PAHs than coke oven workers exposed at high levels, suggesting reduced carcinogenic potency at high exposures. Similarly, among never-smokers and current smokers, the ratio of serum carboxyhemoglobin to number of cigarettes smoked per day decreased with increasing smoking intensity.\(^12\) Studies of PAHs and serum carboxyhemoglobin have limitations for evaluating tobacco effects, since they can arise from nontobacco sources. In contrast, 4-(methylnitrosamino)-(3-pyridyl)-1-butanone (NNK) is a tobacco-specific carcinogen, and levels of its metabolites are markers of tobacco effects.\(^13,14\) Using data from by Joseph et al.,\(^15\) an ongoing analysis finds a decline in the ratio of NNK metabolites to urinary cotinine (a marker of tobacco exposure) with increasing cotinine levels, which is consistent with decreasing potency (J. H. Lubin and S. S. Hecht, personal communications). Intensity patterns may thus reflect biologic phenomena, with enhanced potency resulting from reduced DNA repair capacity,\(^16,17\) and with reduced potency resulting from an increased repair capacity in heavy tobacco users,\(^18–21\) saturation of activation pathways,\(^9–11\) or an increased induction of detoxification enzymes.\(^22\)

Although patterns and effect modifications of the EOR per pack-year may reflect biologic phenomena, these patterns may also reflect influences of nicotine satiation, whereby carcinogenic yield per cigarette decreases with increasing intensity as smokers seek to maintain addiction-sufficient nicotine levels, such that the number of cigarettes per day increasingly overestimates the internal exposure rate. How-
ever, in the European study, there was no evidence of a relationship between frequency or depth of inhalation and intensity after controlling for total pack-years. In contrast, a study of 190 smokers did find increased plasma cotinine and nicotine levels with increased intensity, and a marginally significant ($P = 0.08$) decline in “nicotine boost”, ie, an increase in blood plasma nicotine per cigarette.

We evaluated the potential influence of nicotine satiation in 2 ways, and found that, while cigarettes smoked per day may overestimate internal exposure rate at higher intensities, it is unlikely that nicotine satiation fully explains the inverse exposure rate pattern and the complex patterns of effect modification. We first analyzed effect modification within categories of inhalation, assuming that “overestimation” of internal exposure rate by cigarettes per day was greater in smokers who moderately or deeply inhale than in smokers who slightly inhale. The patterns of effect modification by age, age started smoking, sex and smoking cessation were similar in the 2 inhalation groups.

We next conducted a sensitivity analysis to evaluate the effects of overestimation of internal exposure rate by cigarettes per day, using cotinine as a marker of internal exposure rate. In smokers, cotinine levels increase approximately linearly with smoking intensity up to about 20–30 cigarettes per day. At higher intensities, cotinine has been variously shown to increase without diminution, to increase but at a diminished rate, or to increase until there is a leveling and possibly even a decline. For the sensitivity analysis, we created an adjusted intensity variable, $n_{adj}$, then recomputed pack-years and refitted models. We assumed that observed intensity $n$ accurately characterized internal exposure rate up to $n_o$ cigarettes per day, ie, the regression of $n_{adj}$ on $n$ increased from zero with slope one. We considered 2 adjustment schemes. One scheme specified a piecewise linear relationship with a reduced slope above $n_o$, namely, $n_{adj} = n$ for $n \leq n_o$ and $n_{adj} = n_o + (n - n_o) \times (1 - K)$, for $n > n_o$, where $K$ is the reduction fraction. A second scheme applied a proportionality factor that increasingly deviates from one. The regression equation was $n_{adj} = n$ for $n \leq n_o$ and $n_{adj} = n_o \times \exp\{\ln(0.5) \times (n - n_o)/T\}$ for $n > n_o$, where $T$ is analogous to a “half-life” and represents the number of cigarettes per day that results in a 50% adjustment. Above $n_o$, $n_{adj}$ increases to a maximum then declines. For example, with $n_o = 25$ and $T = 20$, values $n = 30$, 45 and 80 cigarettes per day adjust to $n_{adj} = 25 + (30-25) \times 0.5^{0.25} = 29.2$, $25 + (45-25) \times 0.5^{1.0} = 35.0$ and $25 + (80-25) \times 0.5^{2.75} = 33.2$ cigarettes per day, respectively.

We evaluated $n_o = 20$, 25, 30 with $K = 0.3$, 0.7 for scheme 1 and $T = 20$, 30, 40 for scheme 2. We found only minimal changes in statistical inference on effect modification, due to the high proportion of light and moderate smokers in the data, with 76% of smokers consuming 25 or fewer cigarettes per day, and 86% smoking 30 or fewer. For various $n_o$ and $T$ or $K$, the adjusted estimates of EOR per pack-year

---

**FIGURE 6.** (A, B) Estimated excess odds ratio (EOR) per pack-year for categories of smoking intensity (square symbol) within categories of type of cigarette smoked and the fitted EOR per pack-year based on the model $\beta_n g(n)$ (solid line) and the pointwise 95% prediction interval (dashed line). Results are summarized over study. (C) First derivatives of the fitted models.
within intensity categories increased (since observed cigarettes per day reflects an “overestimate” of internal exposure rate). However, the shape of the EOR function remained largely unchanged, since simultaneously \( n_{adj} \) for higher intensities (x-axis values) shifted towards lower intensities. Nevertheless, since ranges for intensities are reduced, power to detect variations declines.

Our analysis of effect modification finds that ORs for lung cancer decline with time since cessation of smoking.1 More specifically, intensity plays a major role in determining risk in active smokers (by modulating the EOR per pack-year), but a more limited role in long-term former smokers (Fig. 1). Thus, factors related to reduced potency at high intensities have diminished impact with longer time since carcinogenic challenge, a pattern consistent with greater DNA repair capacity in current, as compared with former or never smokers, and in heavier smokers.20

Variations in smoking-related ORs with frequency and depth of inhalation and type of cigarette are mediated through smoking intensity. Smokers who never or only slightly inhale, or who smoke only filtered cigarettes, incur less intensity-dependent modulation of the EOR per pack-year compared with more vigorous inhalers or nonfilter/mixed smokers. This implies that, for equal total pack-years, the level of lung cancer risk is relatively less influenced by intensity in infrequent or slight inhalers, and relatively more influenced by intensity in frequent or deep inhalers.

There is debate about the magnitude of ORs for lung cancer by sex for comparable cigarette exposure.31–42 Evidence suggests that women and men differ in their expression of phase I and II enzymes, NNK metabolism, DNA repair capacities, and growth and hormonal factors,19,43,44 but also suggests comparable susceptibility.45 In our data, intensity effects are similar for men and women, suggesting that delivery and processing of tobacco carcinogens are comparable in the 2 sexes. However, estimates of EOR per pack-year are higher in men than in women, indicating differential consequences of total exposure or duration of exposure.

Patterns of EORs per pack-year differ for subjects younger than age 50, consistent with differential risk at younger ages.46 However, the number of subjects under age 50 and the range for pack-years are limited and results are therefore uncertain.

In the pooled data, 1675 smokers (854 cases and 821 controls) started smoking before age 15 years, and 574 smokers (287 cases and 287 controls) started before age 13 years. Figure 3 suggests that smoking at low intensities may be particularly deleterious for those who start smoking in their preteen and early teen years.

The fitted EOR per pack-year maximum is 0.91 for the German study, which is 3 times the 0.29 value for the European study. In the 7 European centers maxima range from 0.11 to 0.54, with a value of 0.16 for Hamburg and Heidelberg centers combined, all lower than the value for the German study. These differences occur in both men and women. Reasons for the differences are not clear, particularly since smoking exposures are greater in the European study (and in Hamburg and Heidelberg) than in the German study.

Lung cancer rates vary widely throughout Europe, with male rates declining from the mid-1970s through 1990s, depending on country, and female rates increasing, except in the United Kingdom and Ireland.47–49 The European study was hospital-based, and an over-representation of smokers in control subjects could explain the difference. However, results did not materially change when we accounted for concurrent diseases among the controls.

In summary, we find that variations of lung cancer risk with time cessation of smoking, age, frequency and depth of inhalation and type of cigarette smoked are mediated through smoking intensity, rather than total pack-years. In contrast, differences in lung cancer risk by sex are mediated through total exposure, with intensity effects similar by sex. Implications of these findings relative to the molecular basis of smoking risk need further elucidation.

ACKNOWLEDGMENTS

We thank Montserrat Garcia-Closas of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, for useful discussions on this and related topics.

REFERENCES


© 2007 Lippincott Williams & Wilkins


Models of Smoking and Lung Cancer Risk

A Means to an End

Jonathan M. Samet,* Michael J. Thun,† and Amy Berrington de Gonzalez*  

Abstract: This commentary provides some historical context to the analysis of smoking and lung cancer risk by Lubin and colleagues in this issue of EPIDEMIOLOGY. It also considers the potential utility of ongoing efforts to apply complex mathematical models to epidemiologic data on smoking and lung cancer risk. We conclude that the work of Lubin and colleagues adds to the models already developed and points to some potential complexities that models should incorporate.

More than a half century has passed since the first epidemiologic studies provided strong evidence that cigarette smoking causes lung cancer. Since then, numerous case-control and cohort studies have characterized how risk varies with the number of cigarettes smoked and the duration of smoking, as well as time since cessation for former smokers. The epidemiologic data have been used to test and refine models of carcinogenesis, to predict individual risk of developing disease for clinical purposes, and to estimate and refine models of carcinogenesis, to predict individual risk of carcinogenesis to lung cancer. In 1978, Doll and Peto fit a multistage model of carcinogenesis to lung cancer incidence in the British Doctors’ Study and demonstrated that risk increased by the fourth or fifth power of the duration of smoking but only by the second power of the number of cigarettes smoked per day. The results imply that the temporal trend of the time toward increasingly younger age at initiation and more prolonged duration of smoking would lead to substantially increased risk. Furthermore, research analyses that combine number of cigarettes smoked and duration of smoking into a single cumulative measure (eg, pack-years) may misrepresent the relative importance of the 2 factors.

More recently, researchers have examined other models of lung cancer risk in smokers, some replicating the findings of Doll and Peto and others offering new approaches. These new papers are based in large cohort and case-control studies, and model development and estimation are facilitated by new statistical methods and software. The models vary in their dependence on assumptions about the underlying biology of carcinogenesis in the situation of interest. Some, known as biologically motivated models, include parameters for tissue growth, cell kinetics, and number of stages—none of which can be verified or quantified at present. Others, known as empirical risk models, rely more directly on the observational data, although they too make implicit assumptions about the shape of the dose-response relationship and how model parameters should be specified.

In this issue of EPIDEMIOLOGY, Lubin and colleagues further examine the relationship between lung cancer risk—or, more precisely, the excess relative risk of lung cancer—and several measures of smoking. Using a model they proposed previously, they describe a set of analyses on modification of the effects of a cumulative exposure indicator (pack-years of smoking) and of an indicator of “intensity” of smoking (cigarettes smoked per day) by attained age and sex, as well as additional descriptors of smoking. The underlying goal of this analysis, like that of Doll and Peto in their 1978 paper, is to gain insight about the biologic actions of smoking on lung cancer risk. Despite its laudable aim, however, their approach does not resolve some of the inherent challenges encountered when using epidemiologic data to test and refine models of carcinogenesis.

Lubin et al pool data from 2 large case-control studies—the European Smoking and Health Study, a multicenter study conducted in the late 1970s, and the German Radon Study, carried out in the 1990s. In classifying exposure to cigarette smoking, the authors recognize the limitations of standard approaches that specify either duration or cigarettes/
day separately or the combined variable pack-years; such approaches preclude the ability to separate dose rate from total dose. Pack-years alone cannot distinguish a history of smoking 40 cigarettes/day for 10 years from smoking 20 cigarettes/day for 10 years. Similarly, an increase in relative risk associated with smoking more cigarettes/day for a fixed duration cannot be interpreted as representing solely the effect of the increase in cigarettes smoked per day, if cumulative exposure is assumed to affect risk independently. The authors attempt to resolve this problem by including indices of both cumulative exposure (pack-years of smoking) and “intensity” (cigarettes smoked per day) in the model. However, the attempted solution creates a new problem, since cigarettes smoked per day is represented twice in the model. For a fixed number of pack-years, an increase in the number of cigarettes/day requires a decrease in the duration of smoking. A cumulative exposure of 20 pack-years of smoking could represent either 20 cigarettes/day for 20 years or 40 cigarettes/day for 10 years.

The difficulties in classifying these various dimensions of exposure are further complicated by the number of interaction terms tested in the models. Testing for effect-modification of both smoking intensity and total pack-years results in 5 interaction tests for each potential effect modifier. Not only does the large number of interaction tests increase the potential for false-positive results, but the observed interactions are complex and interpreted based primarily on the level of statistical significance. Most of the interaction terms considered, such as years since cessation, appear to modify the association with intensity. The opposite is reported for sex, which modifies the association of lung cancer with total exposure (pack-years) but not with cigarettes per day. It is difficult to interpret these findings in a biologic framework, given the potential for false positives, the complexity of the associations, and unexplained heterogeneity of results between the 2 studies.

Despite these concerns, several findings are intriguing and may in fact provide insight into underlying biologic processes. Both this study and a previous analysis of European case-control study data report that lung cancer risk does not increase in a linear fashion with the number of cigarettes smoked per day. Instead, the excess relative risk per cigarette smoked diminishes above approximately 20 cigarettes per day. This observation meshes nicely with studies of some biomarkers of smoke components for nicotine specifically, the intake per cigarette is less for smokers of above 20 per day, compared with those consuming fewer. Type of cigarette smoked has little impact on risk, even though the contrast is between filter and nonfilter cigarettes smoked decades previously, particularly in the European study. The latter finding is consistent with other epidemiologic studies and with studies comparing biomarkers in smokers of cigarettes with differing machine-measured yields of tar and nicotine. These observations may relate more to the dosimetry of exposure to carcinogens from cigarette smoke than to underlying mechanisms of carcinogenesis.

Another reason to develop models of lung cancer risk in relation to smoking is to predict individual risk for clinical purposes such as counseling about smoking cessation or selecting high-risk individuals for participation in clinical trials of chemoprevention or screening. Models developed for other diseases are exemplary; the risk prediction model based on the Framingham Heart Study data has been widely applied in clinical practice, as has the breast cancer model developed by Gail and colleagues. The Bach lung cancer model was developed primarily to estimate expected lung cancer incidence and mortality rates in heavy smokers to compare these with the observed rates in observational studies of computed tomography screening for lung cancer.

For clinical purposes, we need accepted and validated models for individual risk predictors. Cigarette smoking is a remarkably strong cause of lung cancer, and there are numerous large data sets available for model development and validation. The work of Lubin and colleagues add to the models already developed and points to some potential complexities that models should incorporate. A next step might be collaboration among modelers to compare their approaches and assess generalizability of their risk estimates with the goal of offering the best models for clinical and public health application. After all, these models are just a means to the end – ending the epidemic of lung cancer deaths.

ABOUT THE AUTHORS

JONATHAN SAMET is professor and chair of the Department of Epidemiology at the Johns Hopkins Bloomberg School of Public Health and director of the School’s Institute for Global Tobacco Control. The epidemiology of lung cancer has been a major focus of his work, with emphasis on tobacco smoking and radon. MICHAEL THUN is vice-president of Epidemiology and Surveillance Research at the American Cancer Society, overseeing both cancer surveillance and studies on the causes and prevention of cancer. AMY BERRINGTON DE GONZALEZ is an assistant professor of epidemiology and biostatistics at the Johns Hopkins Bloomberg School of Public Health where her research is focused on cancer epidemiology and prevention and includes the development of statistical methods for epidemiologic studies.

REFERENCES


© 2007 Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.


Stephan Karl Maria Weiland, 1958–2007

Ulrich Keil

Stephan Weiland died suddenly on 19 March 2007 after a full day’s work at his institute in Ulm and an evening swim. Stephan was a well-trained sportsman—he was a marathon runner, a mountain climber, and a dedicated cyclist. At the time of his death he was preparing for a triathlon.

Stephan was born in Münster on 25 December 1958. He studied medicine at the University of Cologne from 1978 to 1985, followed by 2 years of surgery training at the University Hospital of Cologne. From 1987 to 1989 he studied at McGill University in Montreal, where he completed an MSc degree in epidemiology.

He joined the Department of Social Medicine and Epidemiology at Ruhr-University Bochum in 1990. In 1993 he accepted a position at the University of Münster, and in 2002 he moved to the University of Ulm where he was appointed Professor and Chair of the Department of Epidemiology.

Stephan started his career in the field of occupational epidemiology. He created a large historical cohort study and a surveillance project in the German rubber industry, which over the years produced a number of landmark papers. Early in his career, he also became interested in the epidemiology of asthma and allergies. Together with colleagues in England, New Zealand, and Germany, he started the International Study of Asthma and Allergies in Childhood—the well-known ISAAC Project. Throughout his career, Stephan was particularly dedicated to this project, and became one of the most important members of the ISAAC family. The international community will perhaps remember him best for his contributions to this worldwide endeavor. He had a sincere interest in science, and was always a pleasant and stimulating person to work with.

Stephan Weiland viewed epidemiology as a contributor to clinical medicine and as a basic science of public health. One of his last papers is titled “Increasing life expectancy in Germany.” His paper begins by reporting the impressive increases of life expectancy during the last decades. It goes on to note that, despite this progress, nearly 20% of men and 10% of women die before reaching the age of 65. It is heart-breaking to realize that Stephan belongs to this group. He was beloved by his friends and colleagues, and will be sorely missed.