STATISTICS IN MEDICINE, VOL. 11, 167–178 (1992)

STATISTICAL VALIDATION OF INTERMEDIATE ENDPOINTS FOR CHRONIC DISEASES

LAURENCE S. FREEDMAN AND BARRY I. GRAUBARD

Biometry Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD 20892, U.S.A.

AND

ARTHUR SCHATZKIN

Cancer Prevention Studies Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD 20892, U.S.A.

SUMMARY

We discuss the implementation of a criterion due to Prentice for the statistical validation of intermediate endpoints for chronic disease. The criterion involves examining in a cohort or intervention study whether an exposure or intervention effect, adjusted for the intermediate endpoint, is reduced to zero. For example, to examine whether serum cholesterol level is an intermediate endpoint for coronary heart disease (CHD), we may investigate the effect of the cholesterol lowering drug cholestyramine on CHD incidence adjusted for serum cholesterol levels. We show that use of this criterion will usually demand some form of model selection. When the unadjusted exposure or treatment effect is less than four times its standard error, the analysis can usually lead only to a weak form of validation, a conclusion that the data are not inconsistent with the validation criterion. More significant unadjusted exposure effects offer the potential for stronger types of validation statement such as 'the intermediate endpoint explains at least 50 per cent (or 75 per cent) of the exposure effect'.

INTRODUCTION

A major obstacle in the study of the etiology of chronic diseases and the development of effective prevention is the long latent period between the initiation of the disease and its diagnosis. Prospective studies relating possible risk factors to disease or investigating the effects of an intervention on disease incidence therefore require extended periods of follow-up. Such studies are costly. In addition the protracted period during which information accumulates places a strain on funding organizations, who prefer to see a quick return on their investments.

Intermediate endpoints (IE's) are biological markers or events that may be assessed or observed prior to the clinical appearance of the disease, and that bear some relationship to the development of that disease. They are 'intermediate' in the sense of occurring sometime between a given exposure or intervention that affected the disease process and the time of clinical diagnosis of the disease (Figure 1). In view of the difficulties outlined above, intermediate endpoints (IE's) in a chronic disease process are of great interest. If we could study the effect of a risk factor or an intervention on a valid intermediate precursor of the disease then the duration of prospective studies could be shortened. From this perspective we are viewing intermediate endpoints as

0277–6715/92/020167–12$06.00

Received January 1991
Revised May 1991

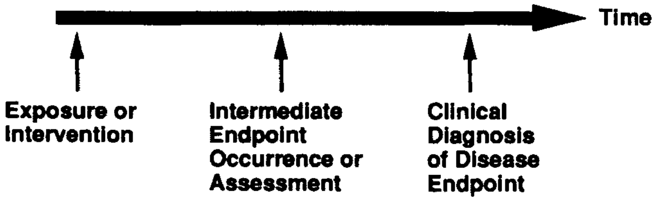
Time
Exposure or Intervention
Intermediate Endpoint Occurrence or Assessment
Clinical Diagnosis of Disease Endpoint

Figure 1. Time sequence of exposure, intermediate endpoint and clinical diagnosis

potential surrogate endpoints for disease incidence in prevention trials. However, the discovery of such an intermediate endpoint would also add importantly to our knowledge of the course of development of the disease and may suggest new approaches to prevention. For example, suppose we show that human papilloma virus infection is an intermediate endpoint for cervical cancer. This would suggest that besides continuing efforts to reduce cervical cancer mortality by screening for cervical dysplasia or early carcinoma, we should explore the avenue of detecting and treating the viral infection. Thus validation of an intermediate endpoint has implications far beyond improving the efficiency of prevention trials.

Schatzkin *et al.*[1] discuss how intermediate endpoints may be studied and validated. They make the following points:

(i) Intermediate endpoints should usually be validated within prospective studies, either observational cohort studies or experimental intervention trials.

(ii) In a cohort study we need to examine the exposure–IE–disease relationship; in an intervention study the intervention–IE–disease relationship should be examined.

(iii) Intermediate endpoints for a disease can only be validated in reference to a given exposure (or intervention). Once validated for that exposure the IE may be considered valid for other exposures supposed to affect the disease through the same pathway.[2]

(iv) The criterion for validation is that the exposure (or intervention) effect on disease, adjusted for the IE, is equal to zero. Susser[3] and Prentice[2] have described similar criteria, Susser in the context of 'intermediate' variables in epidemiological studies and Prentice in the context of surrogate endpoints for clinical trials.

In this paper we expand upon Prentice's work with respect to the criterion stated in (iv), by describing and discussing the statistical implementation of this criterion. We give an example of a 'validation analysis' of serum cholesterol as an intermediate endpoint for coronary heart disease.

OPERATIONAL CRITERIA FOR AN INTERMEDIATE ENDPOINT FOR A BINARY OUTCOME VARIABLE

Following three discussion papers on the use of surrogate endpoints in clinical trials for various diseases,[4–6] Prentice[2] gives general operational criteria for the identification of surrogate endpoints. Since we are interested in identifying intermediate endpoints which may serve as surrogates for the later occurrence of disease, this work is directly relevant. Generally the term intermediate endpoint has been used in the epidemiology literature and the term surrogate endpoint in clinical trials. In the remainder of this paper the two are used synonymously.

Prentice's development is in terms of the *hazard rate* of an event which is the true outcome. The introduction of the time dimension into the problem, although of practical importance, complicates the description of the operational criteria and may mask some of the inherent simplicity of the underlying concepts. In this paper we reproduce the development in Prentice's Section 3, using his notation, but choose the outcome T to be binary ($T = 0$ or 1) instead of time to an event. The treatment (or exposure) group is indexed by X which we assume a categorical variable. In a two-group trial X will equal 1 or 2. The intermediate endpoint S is measured during the trial after the treatment has been given but before the outcome T is determined. The variable S may be discrete or continuous.

In the trial we will be interested primarily in estimating $P(T=1\,|\,X)$ the proportion of cases (individuals with disease) ($T=1$) in each group X. A link between $P(T=1\,|\,X)$ and the intermediate endpoint S is given by:

$$P(T=1\,|\,X) = \int P(T=1\,|\,X,S)f(S\,|\,X)\mathrm{d}S \tag{1}$$

where $f(S\,|\,X)$ is the probability density function of S conditional on X. In this notation $f(S\,|\,X)$ is shorthand for $f(S=s\,|\,X=x)$ and $P(T=1\,|\,X)$ means $P(T=1\,|\,X=x)$.

The notion that a surrogate for T should be able to capture the dependence of T on X can be expressed as

$$P(T=1\,|\,S,X) = P(T=1\,|\,S). \tag{2}$$

We now show that expression (2) provides a criterion for validation of an intermediate endpoint. Readers willing to take this on trust may skip to the next section.

Suppose the intermediate endpoint is independent of treatment, so that

$$f(S\,|\,X) = f(S). \tag{3}$$

Then (3), together with (1) and (2), implies that

$$\begin{aligned} P(T=1\,|\,X) &= \int P(T=1\,|\,X,S)f(S\,|\,X)\mathrm{d}S \quad &&\text{(Step 1; from (1))}\\ &= \int P(T=1\,|\,S)f(S\,|\,X)\mathrm{d}S &&\text{(Step 2; from (2))}\\ &= \int P(T=1\,|\,S)f(S)\mathrm{d}S, &&\text{(Step 3; from (3))}\end{aligned}$$

that is,

$$P(T=1\,|\,X) = P(T=1). \tag{4}$$

Thus any departure from the null hypothesis (4) will be reflected in a departure from the null hypothesis (3), as long as criterion (2) holds. However for departure from (3) to imply departure from (4) we need a further condition besides (2). Specifically, we require that

$$P(T=1\,|\,S) \neq P(T=1) \tag{5}$$

so that the intermediate endpoint is prognostically related to the outcome T.

If on the contrary $P(T=1\,|\,S) = P(T=1)$ then at step 3 above we could write

$$\int P(T=1\,|\,S)f(S\,|\,X)\mathrm{d}S = \int P(T=1)f(S\,|\,X)\mathrm{d}S = P(T=1)$$

and we would not require the relationship (3) to hold. Thus (5) is required for (3) to imply (4).

Finally, Prentice discusses restricting the class of alternatives to (3) to certain distributions $f(S \mid X)$. His restriction (6) is equivalent to including only distributions $f(S \mid X)$ which lead to

$$P(T=1 \mid X) \neq P(T=1). \tag{6}$$

Thus we consider only distributions $f(S \mid X)$ which lead to an overall non-zero difference between any pair of treatment groups.

VALIDATION OF AN INTERMEDIATE ENDPOINT

Having established an operational criterion for an intermediate endpoint, namely that given in (2), we now consider how such a criterion might be employed. Specifically, how can we use (2) to validate a potential intermediate endpoint?

To simplify the discussion we assume that there are two treatment groups ($X = 1$ or 2). In this case criterion (2) may be written

$$P(T=1 \mid S, X=1) = P(T=1 \mid S, X=2). \tag{7}$$

In words, we require that the proportion of cases ($T = 1$), while dependent upon the intermediate endpoint value S, is the same in the two treatment groups *for any value of* S.

If S is discrete, taking k values $s_1, \ldots, s_k$, and $P(T=1 \mid S=s_i, X=j) = p_{ij}$ ($i = 1, \ldots, k$; $j = 1, 2$) then (7) may be written:

$$p_{i1} = p_{i2} \qquad (i = 1, \ldots, k). \tag{8}$$

Thus our criterion defines a composite null hypothesis involving the equality of k pairs of proportions. This hypothesis may be tested, using a test statistic which under the null hypothesis has a χ^2 distribution on k degrees of freedom.

While such a test directly addresses the generality of criterion (2), it may not always be the best approach to validation of the intermediate endpoint. Specific departures from the null hypothesis (8) may occur for which this general test has rather poor statistical power. The problem of power becomes greater the larger the value of k. Moreover if S is a continuous variable the general χ^2 test of criterion (2) can only be performed if S is grouped into a discrete number of intervals.

Another approach is to restrict attention to patterns of departure from the null hypothesis (8) which are thought to be the most likely to occur. These patterns of departure may be described in statistical models linking outcome variable T to S and X. A class of model which may be particularly useful is

$$g(p(T=1 \mid S, X=j)) = h(S) + \tau_j. \tag{9}$$

In this model the effects of S and X are additive for a chosen transformation g of the probability of disease. For example, g may be the logistic transformation. The term $h(S)$ represents some function of S, including parameters which are estimated from the data. For example, if S is discrete, $h(S)$ may equal $\mu + \sigma_i$ when $S = s_i$ ($i = 1, \ldots, k$). The parameter μ represents the overall mean and is estimated from the data. The parameters σ_i are estimated under the usual restriction that $\sigma_1 + \ldots + \sigma_k = 0$. The parameter τ_j represents the jth treatment effect with the restriction that $\tau_1 + \tau_2 = 0$.

As indicated above a specific case of (9) is given by

$$\log\left[\frac{p(T=1 \mid S=s_i, X=j)}{1 - p(T=1 \mid S=s_i, X=j)}\right] = \mu + \sigma_i + \tau_j \tag{10}$$

which is the familiar linear logistic model.

If we consider that model (9), or more specifically model (10), provides a reasonable model for the data, then criterion (2) is now satisfied as long as $\tau_1 = \tau_2 (=0)$. Hence a test of the criterion (2) is provided by testing the null hypothesis that $\tau_1 = \tau_2 (=0)$. (In this situation the appropriate statistical test is the Mantel–Haenszel test.[7]) Thus by restricting our attention to classes of model represented by (9) we need investigate the equality of only *one* pair of parameters (τ_1, τ_2) rather than k pairs $(p_{i1}, p_{i2}$ for $i = 1, \ldots, k)$. Provided the model is well chosen we should thereby considerably increase the power to detect departures from criterion (2).

Model (10) represents a model with no interaction between intermediate endpoint S and treatment X. Introduction of the interaction terms $(\sigma\tau)_{ij}$ into model (10) would lead to the new model

$$\log\left[\frac{p(T=1 \mid S=s_i, X=j)}{1-p(T=1 \mid S=s_i, X=j)}\right] = \mu + \sigma_i + \tau_j + (\sigma\tau)_{ij}. \qquad (11)$$

If this model were appropriate then criterion (2) is now satisfied only if $\tau_j = 0$ and $(\sigma\tau)_{ij} = 0$. The test for this hypothesis is equivalent to the test of $p_{i1} = p_{i2}$ $(i = 1, \ldots, k)$. Thus, simplifying the test of criterion (2) by using model (10) is based on assuming no interaction between intermediate endpoint and treatment. This assumption may itself be subjected to statistical testing.

The following procedure for validation of an intermediate endpoint therefore suggests itself:

Step A Test for interaction between intermediate endpoint and treatment. If a significant interaction is found there is strong evidence against criterion (2) and the procedure may stop.

Step B If there is no significant interaction, adopt a no interaction model and test for a treatment effect. If there is a significant treatment effect there is strong evidence against criterion (2).

The above procedure is based on significance tests of a null hypothesis derived from criterion (2). A statistically significant result gives evidence against the use of the candidate intermediate endpoint. However, lack of statistical significance does not necessarily constitute strong evidence for the intermediate endpoint. It is therefore helpful to look for additional measures of evidence for validating the endpoint.

One such measure is to examine the reduction in the estimate of τ_1 when changing from model (10) with $\sigma_i = 0\,(i = 1, \ldots, k)$ to model (10) with no restriction on σ_i except $\sigma_1 + \ldots + \sigma_k = 0$. If the difference between the two estimates is statistically significant this yields evidence that at least part of the effect of treatment X is explained by its effect on the intermediate endpoint. In the Appendix we present a statistical method for this significance test.

A more direct approach is to estimate the proportion of the exposure effect that is explained by the intermediate endpoint. A natural estimate is given by $1 - (\hat{\tau}_{1a}/\hat{\tau}_1)$, where $\hat{\tau}_{1a}$ is the adjusted estimate of τ_1 and $\hat{\tau}_1$ is the unadjusted estimate. If criterion (2) were satisfied we should expect this estimated proportion to equal 1. We may satisfy ourselves of the importance of the intermediate endpoint by showing the lower limit of the 95 per cent confidence interval of this estimate to be greater than some critical value, say 0·5 (or 0·75). In other words we calculate confidence limits for $1 - (\hat{\tau}_{1a}/\hat{\tau}_1)$ to see if the data provide convincing evidence that the intermediate endpoint explains at least 1/2 (or 3/4) of the exposure effect. An asymptotic method, based on Fieller's Theorem,[8] for calculating the confidence limits is outlined in the Appendix.

SERUM CHOLESTEROL AND CORONARY HEART DISEASE

The Lipid Research Clinics Coronary Primary Prevention Trial[9] is a multicentre, randomized, double-blind study which has tested the efficacy of a cholesterol lowering drug, cholestyramine, for reducing the risk of coronary heart disease (CHD) in 3806 asymptomatic middle-aged men with hypercholesterolemia (Lipid Research Clinics Program). The treatment group received cholestyramine and the control received placebo for an average of 7·4 years.

Table I shows data on combined CHD mortality and morbidity in the trial, according to treatment group and serum cholesterol levels at 1 year after entry to the trial. The number of persons at risk in each group is adjusted for person years of follow-up. Only CHD events occurring after 1 year are included. We assume as an approximation that each subject has the same length of follow-up, and accordingly we treat the number of events as binomially distributed.

For these data there are five cholesterol groups ($k = 5$) and two treatment groups. The event is either death from CHD or the occurrence of a myocardial infarction. It is clear from Table I that condition (5) holds, so that cholesterol level is a strong prognostic factor. It is also clear from the data that condition (6) holds, namely that the cholestyramine treatment reduces CHD events. In fact the reduction is significant at the 5 per cent level ($\chi^2 = 4{\cdot}64$ on 1 d.f., $P = 0{\cdot}03$). Thus we may apply the procedure outlined at the end of Section 3.

Model (11) represents the saturated model for these data and therefore has zero deviance and zero degrees of freedom. To test the importance of the interaction terms we may fit the model without interaction (Model (10)). The deviance is now 3·737 on 4 degrees of freedom. Since the expected value of the deviance under the null hypothesis of no interaction is 4·0 (equal to the degrees of freedom) there is no evidence here of the presence of an interaction term. This completes step A of the procedure.

In step B we fit the model with the treatment terms (τ_j) but no cholesterol terms (σ_i), and compare the results with those for Model (10) which includes the cholesterol terms. In the model with no cholesterol terms, the estimate of the treatment effect is $-0{\cdot}2610$ with standard error 0·1216. This, as mentioned earlier, represents a statistically significant treatment difference ($P = 0{\cdot}03$). When cholesterol terms (σ_i) are added to the model the estimate of the treatment effect is reduced to $-0{\cdot}1310$ with standard error 0·1340. Thus the change in cholesterol levels explains half ($-0{\cdot}1310/-0{\cdot}2610$) of the observed treatment difference in the CHD event rate. Since the cholesterol-adjusted treatment effect is not statistically significant ($P = 0{\cdot}3$) we conclude that the data are consistent with criterion (2), namely that cholesterol level satisfies the requirements for a surrogate endpoint for a CHD event in this trial. The full results of the model fitting procedure are shown in Table II.

The rather weak statement that the 'data are consistent with criterion (2)' may be typical of what can be achieved in practice in such validation analyses. Ideally we would like to see the estimated treatment difference change from highly significant to vanishingly small after adjustment for an intermediate endpoint. However, sampling variation will often interfere with such a desirous state of affairs. If the unadjusted treatment difference is only a little more than 2 standard errors distant from the null value then we cannot expect high precision in the validation of the intermediate endpoint. The situation will be much clearer when unadjusted treatment effects are large, say 4 standard errors or more from the null value.

As shown above the estimate of the treatment effect changes from $-0{\cdot}2610$ to $-0{\cdot}1310$ when serum cholesterol is included in the model. The change in the regression coefficient is therefore $-0{\cdot}1310 - (-0{\cdot}2610) = 0{\cdot}1300$. The variance of the change may be calculated from methods described in the Appendix. The covariance of the two estimates of the treatment effect is

Table I. Definite CHD mortality or myocardial infarction events in the Lipids Research Clinics Coronary Primary Prevention Trial according to randomized treatment group (P = placebo, C = cholestyramine)

	Number of patients at risk*		Number of events		Percent of events†	
Cholesterol (mg/dl) at year 1	P	C	P	C	P	C
< 180	7	106	0	9	0·0	8·5
180–230	91	675	8	34	8·8	5·0
230–280	1069	742	78	54	7·3	7·3
280–330	636	304	64	23	10·1	7·6
> 330	115	61	18	10	15·7	16·4
Total	1918	1888	168	130	8·8	6·9

* Adjusted for person-years follow-up. Because of this adjustment the numbers of patients in the different subgroups differ slightly from those entered in the trial, although the total number 3806 remains the same

† (Number of events/Number of patients at risk) × 100

Table II. Intermediate endpoint validation analysis for the LRC Trial data

Model	Deviance	Degrees of freedom	Estimated treatment effect	Standard error
$\ln(p/1-p)=\mu+\sigma_i+\tau_j+(\sigma\tau)_{ij}$	0·00	0	—	—
$\ln(p/1-p)=\mu+\sigma_i+\tau_j$	3·74	4	− 0·13	0·13
$\ln(p/1-p)=\mu+\tau_j$	22·35	8	− 0·26	0·12
$\ln(p/1-p)=\mu$	26·99	9	—	—

calculated to be 0·01475; hence the variance of the change equals $0{\cdot}1216^2 + 0{\cdot}1340^2 - 2\times 0{\cdot}01475 = 0{\cdot}003243$, and the standard error equals 0·0569. Thus the change in regression coefficient (0·1300) is 2·28 times its standard error (0·0569) with a P-value of 0·024. This indicates that the effect of treatment on serum cholesterol does explain at least some of the observed reduction in coronary heart disease mortality seen in the trial.

The proportion of the treatment effect which is explained by serum cholesterol is estimated by $1-(-0{\cdot}131/-0{\cdot}261)=0{\cdot}4981$. Using the asymptotic methods given in the Appendix, the 95 per cent confidence interval is (0·07, 5·91). This interval is too wide to convey any useful information, but agrees with our previous analysis in which we have shown that serum cholesterol does explain at least some of the treatment effect.

POWER TO CONCLUDE THAT THE INTERMEDIATE ENDPOINT EXPLAINS AT LEAST A FRACTION f OF THE EXPOSURE EFFECT

As noted earlier, the validation would carry more weight if it could be shown that the greater part of the exposure effect were explained by the IE. Formally we would like the lower 95 per cent confidence limit for $1-(\tau_{1a}/\tau_1)$ to be greater than a chosen fraction f (for example, 0·5 or 0·75).

We may calculate the probability of obtaining this result under the hypothesis $\tau_{1a}=0$ (that the IE explains all of the exposure effect). Assuming confidence limits are calculated using Fieller's theorem and using a result published by Yee,[10] the probability is

$$\Phi[(1-f)(\hat{\tau}_1/\text{S.E.}(\hat{\tau}_1))/(2(1-\rho)(1-f)+f^2)^{1/2}-1{\cdot}96] \qquad (12)$$

Table III. Probability* of concluding that the intermediate endpoint explains at least a certain fraction (f) of the exposure effect, as a function of its relative standard error and the correlation, ρ, between the adjusted and unadjusted estimates of the effect

		Ratio of estimated exposure effect to its standard error			
	f	2	4	6	8
$\rho = 0{\cdot}9$	0·5	0·39	0·92	0·999	1·00
	0·75	0·09	0·25	0·48	0·73
$\rho = 0{\cdot}7$	0·5	0·28	0·77	0·98	1·00
	0·75	0·09	0·22	0·43	0·66
$\rho = 0{\cdot}5$	0·5	0·21	0·64	0·93	0·996
	0·75	0·08	0·20	0·38	0·60

* Under the assumption that the intermediate endpoint truly explains all of the exposure effect

where Φ is the standard normal integral, ($\Phi(1{\cdot}96) = 0{\cdot}975$), and ρ is the correlation between $\hat{\tau}_1$ and $\hat{\tau}_{1a}$. The probability is a function of the ratio of the unadjusted exposure effect to its standard error and also the correlation ρ. More details are supplied in the Appendix. Typically we would expect ρ to be positive and large; in our example above, ρ is approximately 0·9. In Table III we tabulate the probability for $f = 0{\cdot}5, 0{\cdot}75$; $\hat{\tau}_1/\mathrm{S.E.}(\hat{\tau}_1) = 2, 4, 6, 8$; and $\rho = 0{\cdot}9, 0{\cdot}7, 0{\cdot}5$.

The results show that the probability of concluding that $1 - (\tau_{1a}/\tau_1)$ is significantly greater than 0·5 is greater than 0·9 when the ratio $\hat{\tau}_1/\mathrm{S.E.}(\hat{\tau}_1)$ is 4 or larger. However the probability of concluding that $1 - (\tau_{1a}/\tau_1)$ is significantly greater than 0·75 remains low even for very highly significant exposure effects ($\hat{\tau}_1/\mathrm{S.E.}(\hat{\tau}_1) = 6$ and 8).

DISCUSSION

Intermediate endpoints or surrogate endpoints are currently of interest in the study of several chronic diseases. Besides the cancer and heart disease examples mentioned in this paper, there is much interest in their use for studying the treatment of AIDS.[11,12] In this paper we have attempted to clarify the criteria which may be used to validate an intermediate endpoint and to describe how the validation analysis may be conducted.

While endeavouring to do this we have realized that the rather general criterion (2) is difficult to test in practice. If the intermediate endpoint is continuous then a statistical model must be chosen to test whether (2) is satisfied. If the IE is discrete with a finite number k of categories then the exposure (or intervention) effect should be zero within each IE category for criterion (2) to be satisfied. However as k increases the power for testing such a hypothesis becomes weaker. Simplifying models may be required to recover some of the lost power. Thus the validation analysis will tend to involve some aspects of model selection.

Secondly, as noted earlier, the test of criterion (2) involves a test of the hypothesis that a certain intervention or exposure effect, adjusted for the IE, is zero. If this hypothesis were true, then criterion (2) could be accepted and the IE validated. However we are well aware that non-significant results do not prove the null hypothesis. We can only say that the data are 'not inconsistent' with the null hypothesis. This is a rather weak form of validation.

The validation analysis would carry more weight if it could be shown that the greater part of the exposure or intervention effect were explained by the IE. For this to be possible the unadjusted exposure effect would need to be large in relation to its standard error. Table III gives the probability of concluding that the IE explains at least 1/2 or at least 3/4 of the exposure effect when the exposure effect is 2, 4, 6, or 8 times its standard error. It appears that to make reasonably precise estimates of the proportion of the effect explained by the intermediate endpoint, we need to be explaining unadjusted exposure effects which are at least 4 times their standard errors.

Such large effects are quite commonly found in observational cohort studies. However the intervention effects found in long-term prevention studies have generally been smaller. For this reason and because cohort studies are more often conducted than prevention trials, the best opportunities for validating IE's are likely to be through incorporating measurement of IE's in prospective cohort studies.

On the other hand, much may be gained also from selectively including IE measurements in some intervention studies. First we gain the opportunity to assess in a randomized study the effect of the intervention on an IE which is thought to relate to the disease endpoint; secondly we can study prospectively that hypothesized relationship; and thirdly some form of validation exercise for the IE, as described above, may be possible. In particular, a very highly significant intervention effect could permit a strong form of validation. The difficulty is that we cannot tell at the outset of the study whether a highly significant intervention effect will emerge.

We should note that our chosen measure of the proportion of the exposure effect explained by the IE, $1 - (\hat{\tau}_{1a}/\hat{\tau}_1)$, is not unique. The value of $2\tau_1$ represents the log odds ratio of disease given exposure. Other measures of disease effect include the excess relative odds given by $\exp(2\tau_1) - 1$. Use of this would lead to a different measure of the proportion of exposure effect explained by the IE, namely $1 - (\exp(2\hat{\tau}_{1a}) - 1)/(\exp(2\hat{\tau}_1) - 1)$. For large exposure effects the difference between these two measures can be considerable. For example if $\hat{\tau}_1 = 1$ and $\hat{\tau}_{1a} = 0{\cdot}5$ then according to the 'log odds' measure the proportion of exposure effect explained is 0·5, whereas according to the 'excess odds' measure the proportion explained is 0·73. The choice of measure can sometimes be guided by knowledge regarding the additive or multiplicative nature of the risk. Often, however, insufficient data exist to choose between the additive and multiplicative model and in these cases the choice of measure remains somewhat arbitrary.

With reference to the example, we regard the result of our analysis as encouraging news for the potential use of serum cholesterol level as an intermediate endpoint for coronary heart disease, at least for evaluating drugs with a mechanism similar to cholestyramine. The unadjusted treatment effect was halved when adjusted for serum cholesterol and was not statistically significant after adjustment. A separate analysis shows that the difference between the adjusted and unadjusted effect was statistically significant, so the change in serum cholesterol level distribution does appear to explain at least some of the treatment effect. However, as mentioned above, when the unadjusted treatment effect is itself only just significant ($P = 0{\cdot}03$) then the validation analysis is not all that convincing. Indeed we can think of three reasons why the treatment effect should not be completely explained by serum cholesterol. First, the cholesterol level may itself be a surrogate for the real determinate of the disease process, which could perhaps be LDL cholesterol level, HDL cholesterol level or a combination of lipid levels. Secondly, a 1 year cholesterol level will probably not fully represent the effect over 7·4 years of follow-up.

Thidly, cholesterol levels vary substantially within the same subject. This variation would cause the treatment effect adjustment, based as it was on a single measurement of serum cholesterol on each individual, to be incomplete. Such adjustment would fail to reduce the treatment effect to zero even if criterion (2) were true. The larger the variation in IE within an individual compared to the between individual variation, the less effective will be the adjustment. Thus analyses of IE's will need to take account of errors in the measurement of the IE. For our example, work by Carroll *et al.*[13] will be helpful in correcting estimates for this error measurement problem.

Finally, we note that in circumstances where many intervention trials or cohort studies have been conducted in which both intermediate endpoint and outcome have been measured, we can take other approaches to validation of the intermediate endpoint. A'hern *et al.*,[14] for example, have correlated the complete response rate with the length of survival over a series of trials of treatments for advanced breast cancer. With the help of such data we can develop a model for predicting the effect on length of survival of a treatment observed to increase the complete response rate by a certain amount. By contrast the method we have discussed in this paper uses data on individuals in a single study, rather than aggregate data over many studies. It is possible that the two approaches could be combined to advantage if the individual data within the many studies were available.

APPENDIX: FIRST-ORDER TAYLOR APPROXIMATION OF THE COVARIANCE BETWEEN ESTIMATED BETAS FROM TWO LOGISTIC REGRESSION MODELS APPLIED TO THE SAME DATA SET

Suppose we have two logistic regression models relating disease occurrence to covariates for n subjects. Let

$$\boldsymbol{\beta}^{(1)} = (\beta_1^{(1)}, \ldots, \beta_p^{(1)})' \quad \text{and} \quad \boldsymbol{\beta}^{(2)} = (\beta_1^{(2)}, \ldots, \beta_q^{(2)})'$$

denote $p \times 1$ and $q \times 1$ vectors of unknown parameters, respectively, for the two models. The outcome data consists of $\mathbf{d} = (d_1, \ldots, d_n)'$ a vector of indicator variables $d_j = 1$ if the jth subject develops the disease and $d_j = 0$ otherwise.

Let $X^{(1)}$ and $X^{(2)}$ denote $n \times p$ and $n \times q$ matrices of covariates, respectively, for the two models. We have two logistic regression models ($i = 1, 2$) where the probability of disease for each observation is given by

$$\boldsymbol{\Pi}_i(\boldsymbol{\beta}^{(i)}) = (\Pi_{i1}, \ldots, \Pi_{in})'$$

where

$$\Pi_{ij} = \frac{\exp(\mathbf{X}_j^{(i)'} \boldsymbol{\beta}^{(i)})}{\{1 + \exp(\mathbf{X}_j^{(i)'} \boldsymbol{\beta}^{(i)})\}}$$

and $\mathbf{X}_j^{(i)'}$ is the jth row of $X^{(i)}$, $i = 1, 2$.

The log-likelihood for the logistic regression is equal to

$$\sum_{j=1}^{n} d_j \ln \Pi_{ij} + (1 - d_j) \ln(1 - \Pi_{ij}).$$

The first derivative with respect to $\boldsymbol{\beta}^{(i)}$ of the log-likelihood gives the following set of estimating equations (when $i = 1$ there are p such equations and when $i = 2$ there are q equations):

$$X^{(i)'}(\mathbf{d} - \boldsymbol{\Pi}_i(\hat{\boldsymbol{\beta}}^{(i)})) = 0,$$

where $\hat{\boldsymbol{\beta}}^{(i)}$ is the solution.

Taking a first-order Taylor expansion about $\boldsymbol{\beta}^{(i)}$ of $\boldsymbol{\Pi}_i(\boldsymbol{\beta}^{(i)})$ and substituting into the estimating equations gives

$$(\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}^{(i)}) = (X^{(i)'} C(\boldsymbol{\beta}^{(i)}) X^{(i)})^{-1} X^{(i)'}(\mathbf{d} - \boldsymbol{\Pi}_i(\boldsymbol{\beta}^{(i)})) \tag{13}$$

where $C(\boldsymbol{\beta}^{(i)}) = \operatorname{diag}(\Pi_{i1}(1 - \Pi_{i1}), \ldots, \Pi_{in}(1 - \Pi_{in}))$. We need to determine the variance of the difference between two elements of $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$, respectively. For convenience, we call them elements $\hat{\beta}_2^{(1)}$ and $\hat{\beta}_2^{(2)}$ ($\hat{\beta}_1^{(1)}$ will normally represent the intercepts). In the text these same estimates are referred to as $\hat{\tau}_1$ and $\hat{\tau}_{1a}$, respectively. Since

$$\operatorname{var}[\hat{\beta}_2^{(1)} - \hat{\beta}_2^{(2)}] = \operatorname{var}\hat{\beta}_2^{(1)} + \operatorname{var}\hat{\beta}_2^{(2)} - 2\operatorname{cov}(\hat{\beta}_2^{(1)}, \hat{\beta}_2^{(2)})$$

we need to determine the covariance term. Note that in practice estimates of the variance terms will be provided by standard logistic regression analyses of the respective models.

To determine the covariance term we note that in equation (13) the only quantity which is a random variable is **d**. We write the inverse of the matrix $X^{(i)\prime}C(\boldsymbol{\beta}^{(i)})X^{(i)}$ as $E^{(i)}$ and denote its elements by $e_{jk}^{(i)}$. The covariance term may now be written as

$$\sum_{j=1}^{p}\sum_{k=1}^{q} e_{2j}^{(1)}\, e_{2k}^{(2)} \left\{ \sum_{r=1}^{n} X_{jr}^{(1)}\, X_{kr}^{(2)} \operatorname{var}(d_r) \right\}.$$

The variance of d_r is calculated as $\Pi_{ir}(1 - \Pi_{ir})$ under Model i. For the purposes of calculating the covariance term we used the Model 2 (which contains Model 1) estimates of $\boldsymbol{\Pi}$.

Using Fieller's Theorem, the 95 per cent confidence limits of $(1 - \beta_2^{(2)}/\beta_2^{(1)})$ may be obtained from the two quantities given by:

$$1 - \{((\hat{\beta}_2^{(1)}\hat{\beta}_2^{(2)} - z^2\operatorname{cov}(\hat{\beta}_1^{(2)}, \hat{\beta}_2^{(2)})) \pm [(\hat{\beta}_2^{(1)}\hat{\beta}_2^{(2)} - z^2\operatorname{cov}(\hat{\beta}_2^{(1)}, \hat{\beta}_2^{(2)}))^2$$
$$- ((\hat{\beta}_2^{(1)})^2 - z^2\operatorname{var}\hat{\beta}_2^{(1)})((\hat{\beta}_2^{(2)})^2 - z^2\operatorname{var}\hat{\beta}_2)]^{1/2})/((\hat{\beta}_2^{(1)})^2 - z^2\operatorname{var}\hat{\beta}_2^{(1)})\}$$

where z is the upper 97·5 percentile of the normal distribution, that is, $z = 1{\cdot}96$, and the variance and covariance terms are estimated by the methods discussed previously.

Suppose $\hat{\Theta}_u$ is the Fieller upper 95 per cent confidence limit of $(1 - \beta_2^{(2)}/\beta_2^{(1)})$. Then Kee[10] shows that

$$Pr[\hat{\Theta}_u \geqslant 1 - c] = \Phi[z + (\beta_2^{(2)}/\beta_2^{(1)} - c)\hat{\beta}_2^{(1)}/(\operatorname{var}(\hat{\beta}_2^{(1)})(1 - 2c\rho + c^2))^{1/2}]$$

where $\rho = \operatorname{cov}(\hat{\beta}_2^{(1)}, \hat{\beta}_2^{(2)})/(\operatorname{var}(\hat{\beta}_2^{(1)})\operatorname{var}(\hat{\beta}_2^{(2)}))^{1/2}$.

Putting $1 - c = f$, $\beta_2^{(2)} = 0$ and $\hat{\beta}_2^{(1)} = \hat{\tau}_1$ we obtain the result in equation (12).

ACKNOWLEDGEMENTS

We thank Dr. C. E. Davis for providing the data from the LRC trial. Drs. M. Schiffman and S. Dawsey were closely involved in clarifying the logic of how intermediate endpoints may be validated in observational and intervention studies. Drs. T. Chen, M. Gail, E. Korn and R. Simon shared illuminating comments with us on the problems of validation.

REFERENCES

1. Schatzkin, A., Freedman, L. S., Schiffman, M. H. and Dawsey, S. M. 'Validation of intermediate endpoints in cancer research', *Journal of the National Cancer Institute*, **82**, 1746–1752 (1991).
2. Prentice, R. L. 'Surrogate endpoints in clinical trials: definitions and operational criteria', *Statistics in Medicine*, **8**, 431–440 (1989).
3. Susser, M. (ed.) *Causal Thinking in the Health Sciences*, Oxford University Press, New York, 1973.
4. Ellenberg, S. S. and Hamilton, J. M. 'Surrogate endpoints in clinical trials: cancer', *Statistics in Medicine*, **8**, 405–413 (1989).
5. Wittes, J., Lakatos, E. and Probstfield, J. 'Surrogate endpoints in clinical trials: cardiovascular disease', *Statistics in Medicine*, **8**, 415–425 (1989).
6. Hillis, A. and Siegel, D. 'Surrogate endpoints in clinical trials: ophthalmologic disorders', *Statistics in Medicine*, **8**, 427–430 (1989).
7. Mantel, N. and Haenszel, W. 'Statistical aspects of the analysis of data from retrospective studies of disease', *Journal of the National Cancer Institute*, **22**, 719–748 (1959).

8. Fieller, E. C. 'The biological standardization of insulin', *Journal of the Royal Statistical Society*, **7**, Supplement 1–15 (1940).
9. Lipid Research Clinics Program. 'The Lipid Clinics coronary primary prevention trial results. I. Reduction in incidence of coronary heart disease', *Journal of the American Medical Association*, **251**, 351–364 (1984).
10. Yee, K. F. 'The calculation of probabilities in rejecting bioequivalence', *Biometrics*, **42**, 961–965 (1986).
11. Ellenberg, S. S. 'Surrogate endpoints in clinical trials', *British Medical Journal*, **302**, 63–64 (1991).
12. Machado, S. G., Gail, M. H. and Ellenberg, S. S. 'On the use of laboratory markers as surrogates for clinical endpoints in the evaluation of treatment for HIV infection', *Journal of Acquired Immune Deficiency Syndromes*, **3**, 1065–1073 (1990).
13. Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T. and Abbott, R. D. 'On errors-in-variables for binary regression models', *Biometrika*, **71**, 19–25 (1984).
14. A'hern, R. P., Ebbs, S. R. and Baum, M. B. 'Does chemotherapy improve survival in advanced breast cancer? A statistical overview', *British Journal of Cancer*, **57**, 615–618 (1988).