Trend Tests in Epidemiology: P-Values or Confidence Intervals?

LUDWIG A. HOTHORN
University of Hannover
Germany

Summary
Frequently, p-values are used in reporting epidemiological trend data. Because a p-value is a confounded mixture of effect size and sample size in dichotomous data, LANG, ROTHMAN, and CANN (1998) recommended the slope of a trend line with its standard error and a graphical presentation containing the rate ratios as a function of mid-exposure levels. However, the slope contains the assumption of a dose-response function. This article discusses a proposal based on odds ratios and the corresponding one-sided lower confidence intervals for pair-wise comparisons (‘exposure levels with zero exposure’) as well as comparisons between incremental exposure levels. The proposed method allows both decisions on the global trend, the lowest-observed-adverse-effect-level (LOAEL) and the non-observed-adverse-effect-level (NOAEL), and a simple exploratory analysis.

Key words: Trend test; Dose-response analysis; Exposure study; NOAEL; LOAEL.

1. Introduction

Recently, CANTOR, LYNCH, et al. (1998) published an epidemiological exposure study for the risk of developing bladder cancer with increasing duration of exposure to all chlorinated water sources. The results of potential trends are summarized by p-values. In a comment to this paper, LANG et al. (1998) recommended not to report a p-value as the result of a trend test because it is a confounded mixture of magnitude of the underlying measure (size of effect) and the precision of the measure (sample sizes in dichotomous data). A further argument against the use of p-values was that they “are often based on assumptions the reader can not easily judge”. Instead of p-values, they recommend the determination of the slope for the trend line with its standard error or confidence interval (CI) and a figure presenting the rate ratios as a function of mid-exposure levels.

2. The Problem

Here, the arguments against the use of significance tests (and p-values) in epidemiology will not be repeated. In confirmatory randomized studies, an a priori sam-
ple size estimation (based on certain assumptions) is performed, making significance tests appropriate as long as the assumptions are not seriously violated in the data. This is not the case in exploratory studies, for example in many epidemiological observational studies. However, in this article the Neyman-Pearson or Fisher test theory will not be questioned, but see further references (for example Greenland, 1998 or Goodman, 1998).

Lang’s et al. (1998) second argument, that p-values may be based on assumptions the reader cannot judge, is particularly important for trend tests. The p-value of a trend test includes not only effect size and sample size, but also the information on the shape of the dose-response relationship. Trend tests are commonly based on a total order assumption (for proportions $p_i$ of $k + 1$ exposure levels $E_i$, where $E_0$ denotes a non-exposed or negative control group):

$$H_0 : p_0 = p_1 = \ldots = p_k,$$

$$H_1 : p_0 \leq p_1 \leq \ldots \leq p_k \mid p_0 < p_k.$$

The null-hypothesis is rejected (and hence a low p-value results) if any of the inequalities are valid, that is at least one or as many as all. This means that any monotonic shape of the dose-response relationship, for example linear, quadratic, convex, or concave satisfies this alternative. However, epidemiologists have serious concerns with this property, see the discussion by Maclure and Greenland (1992) or Perez-Hoyos and Benavides (1997).

Connected with this problem is that no uniformly most powerful test for all possible ordered alternatives exists. This is reflected by the simple fact that different trend test statistics (even those that differ only by contrast coefficients or by exposure scores) result in different p-values for a selected shape. Beyond monotonicity, the shape is a priori unknown; it is simply an outcome of the study. In this case of a-priori unknown shapes, linear test statistics are recommended by Graubard and Korn (1987). However, according to Neuhaüsler (1996), the Pitman efficiency (in a balanced parametric set-up) of the linear contrast test is reduced to 0.60 for the frequently occurring convex shape with respect to the Helmert contrast, which is optimal in this situation. The Pitman efficiency is roughly the ratio of the slopes of the power functions of two tests near $\alpha$, assuming infinite sample sizes. Here, contrast tests

$$T_C = \sum_{i=0}^{k} c_i p_i \quad \text{(with } \sum_i c_i = 0)$$

are considered. For example, the coefficients $c_i$ for the Helmert contrast are defined as $[c_i = -1, \ i = 0, 1, \ldots, k - 1 \text{ and } c_k = k]$, that is, for three exposure groups and an unexposed group, simply $T_C = -p_0 - p_1 - p_2 + 3p_3$.

Model-based approaches are questionable, usually because no valid assumptions on the function – not only on the shape – are available a priori (alternatives are fractional polynomial regression and regression splines according to Greenland
Furthermore, exposure information is frequently vague in epidemiology. Therefore, the choice of appropriate exposure scores is difficult; an example of this is the open-ended category problem according to Gautam (1997).

Several trend test approaches are available which are ‘average power optimal’, that is, never the worst or the best, but a good power test for all possible shapes of the dose-response relationship. Examples include the likelihood ratio test according to Bartholomew (1959) (although difficult for practical use, even in the parametric set-up), multiple contrast tests (according to Mukerjee, Robertson, and Wright, 1986 and Hothorn, Neuhäuser, and Koch, 1997) and tests with optimal scores according to Podgor et al. (1996).

An important question is, what is the objective of dose-response analysis in epidemiology? One of the basic problems is to test the causality of a hypothetical effect. According to Hill (1971) and Weed and Gorelic (1996) one argument for causality is dose dependence, which means that a global trend either exists or does not exist. But if a global trend is detected, then two questions follow: what is the highest level of exposure that will not yield any effect, and at what level does exposure begin to produce an effect? The model behind these questions is, according to Paracelsus “All things are poisons, their effect depends only on dose.” The difference between these questions seems only one of formulation. But this reflects the selection of either a test for equivalence (proof of safety for identifying the non-observed-adverse-effect-level: NOAEL according to Hauschke and Hothorn, 1998) or a test for the difference for identifying the minimum effective level (minimum unsafe exposure or lowest-observed-adverse-effect-level: LOAEL). Most epidemiological exposure studies use the latter approach. However, some questions seem more appropriate formulated as a proof of safety, such as whether a side effect of a new technology or chemical is safe at all (global safety) or safe up to a certain exposure level (NOAEL).

The problem of whether to make a one- or two-sided test (or to use one- or two-sided confidence intervals) has to be discussed. The ordered alternative hypothesis shown above can also be formulated as a two-sided test problem. But a scenario in which monotonicity is a priori assumed, regardless of whether an increase or a decrease exists, is difficult to imagine. Therefore, the desired direction should be defined a priori and one-sided tests or confidence intervals should be preferred.

Sometimes epidemiologists define a trend only if all increments of exposure cause monotonic increases. All the above-mentioned trend tests are more or less robust against such non-monotonicity, which seems unacceptable for some epidemiological problems. A proposal for strictly incremental tests was recently published by Dosemeci and Benichou (1998). A very strict definition of trend is that all pair-wise increments ‘Ei vs. Ei−1’ are significant. According to the inter-section union principle (Casella and Berger, 1982) the global null-hypothesis is rejected if all elementary hypotheses are rejected at level α (or equivalently $p_{max} < \alpha$). For
the more appropriate confidence interval approach, this requires that all lower
limits of the one-sided \((1 - \alpha)\) confidence intervals to be larger than 1 (odds ratio)
or to be larger than 0 (difference of proportions) (see Figure 2).

According to Bauer (1997), step-wise procedures based on non-pair-wise con-
trasts for estimating the minimum effective level do not control directional error;
that is, ineffective levels can be falsely declared as effective if intermediate non-
monotonicity occurs. Only pair-wise contrasts ‘\(E_i\) vs. \(E_0\)’ (exposure level \(i\) vs.
zero exposure) \(i \in (1, \ldots, k)\) (pair-wise tests or confidence intervals) are appro-
priate for estimating the lowest-observed-adverse-effect-level (LOAEL) or the
NOAEL.

3. An Approach

The consequence of the above arguments for an appropriate trend test in epide-
miology is the exclusive use of pair-wise tests — or better, confidence intervals —
‘\(E_i\) versus \(E_0\)’ (exposure level \(i\) versus zero exposure), \(i \in (1, \ldots, k)\) and ‘\(E_i\) versus
\(E_{i-1}\)’, \(i \in (0, \ldots, k)\) (testing the increments). A possible argument against this
proposal is that information is lost because the intermediate exposure levels are
ignored. Indeed, ensuring a level \(\alpha\)-test (particularly for possible non-monotonic-
ity) will cause a power loss, but this loss is not as dramatic as is generally be-
lieved. The Pitman efficiency (versus the actual optimal test) of the pair-wise con-
trast is 0.90 for a linear shape and 0.67 for a convex shape (balanced parametric
set-up, Neuhäuser, 1996).

In summary, I propose the following approach, for example based on odds
ratios commonly used in epidemiology:

Use point estimators and their \((1 - \alpha)\) one-sided confidence intervals ‘\(E_i\) versus
\(E_0\)’ to control the directional error of global trend tests and NOAEL or LOAEL.

3.1 Identifying lowest-adverse-effect-level (LOAEL)

Start with the comparison ‘\(E_k\) vs. \(E_0\)’. If the lower limit of its one-sided \((1 - \alpha)\)
certainty interval is larger than 1 (or a pre-defined efficacy threshold \(\delta > 1\)),
then go to the next step. Otherwise, stop with the conclusion that no LOAEL (and
no global trend) exists. In the next step, compare ‘\(E_{k-1}\) vs. \(E_0\)’. If the lower limit
of its one-sided \((1 - \alpha)\) confidence interval is larger than 1 (or a pre-defined
‘efficacy’ threshold \(\delta > 1\)), then go to the following step. Otherwise, stop with the
conclusion that level \(k\) is the LOAEL (and a global trend exists). Continue down
to the lowest experimental exposure level until a lower limit is smaller than 1 (or \(\delta\)).

According to Taubes (1995) the use of a relevance efficacy threshold \(\delta > 1\) is
recommended, because with the commonly used \(\delta = 1\) statistically significant but
irrelevant findings frequently occur. The direct inclusion of \(\delta\) in the graphical pre-
sentation is helpful for a decision of causality. For simplicity in Figure 1 a relevance threshold of $\delta = 1$ is used.

It should be noted that this procedure is not robust against downturns at high exposure(s). In this case an approach without order restriction is appropriate, which cannot easily be implemented in the simple approach illustrated in Figure 1.

3.2 Identifying no-observed-adverse-effect-level (NOAEL)

Start with the comparison $E_1$ vs. $E_0$. If the upper limit of its one-sided $(1 - \alpha)$ confidence interval is smaller than the a-priori defined safety threshold $\varepsilon$, then go to the second step. Otherwise, stop with the conclusion that no NOAEL exists. In the second step, compare $E_2$ vs. $E_0$. If the upper limit of its one-sided $(1 - \alpha)$ confidence interval is smaller than the safety threshold $\varepsilon$, then go to the third step. Otherwise, stop with the conclusion: that NOAEL $= E_1$. Continue up to the maximum exposure level possible. The a priori definition of the safety threshold $\varepsilon$ is necessary. Again, the inclusion of the safety threshold $\varepsilon$ directly in the graph is helpful for making a decision concerning safety. In Figure 2 the safety threshold $\varepsilon = 1.5$ is used as an example.

It should be noted that only pair-wise contrasts (or confidence intervals) with a priori ordered hypotheses are appropriate for testing on NOAEL (see details in Hothorn and Hauschke, 2000).

Both approaches test stepwise at level $\alpha$ (that is no multiplicity adjustment is needed), assuming an order restriction based on the inter-section union principle. Notice, that either the proof of safety, i.e. identification of NOAEL or the proof of hazard, i.e. identification of LOAEL should be used, but never both approaches simultaneously.

3.3 Identification of incremental trend(s)

Use point estimators and their $(1 - \alpha)$ one-sided lower confidence intervals for the increments $E_i$ vs. $E_{i-1}$, $i \in (0, \ldots, k)$ that represent incremental trend tests. The strict definition of an incremental trend test is equivalent to ‘all lower confidence intervals are $> 1$’ (see Figure 3).

In Figure 1 the example on the risk of bladder cancer for males with increasing duration of exposure to all chlorinated water sources is used for demonstrating identification of NOAEL. For simplicity mid-exposure scores were used in the graphical presentation. Only crude estimators are used, because not all the data for the logistic regression model are available. Asymptotic confidence intervals for the odds ratios are used due to the rather large sample sizes (see Table 1).

For small sample size studies, the mid-p version which is available in Statxact (Mehta and Patel, 1995) can be recommended.
From Figure 1 we can see the shape of the exposure-response relationship with increasing confidence intervals (reflecting decreasing sample size with increasing exposure) — the frequently observed ‘dilution effect’ in exposure studies. The commonly defined LOAEL (that is, based on $\delta = 1$) is 20–39 years because that is the lowest exposure with the lower limit of the confidence interval is being larger than 1. However, causality seems questionable, because neither the point estimate nor the lower confidence intervals are above a relevance limit of efficacy $\delta = 2.5$, that is, no LOAEL and hence no global trend exists for the strictly relevance-shifted definition according to Taubes (1995).

Figure 2 shows the point estimate of the odds ratio and the upper one-sided confidence limits for identification of the non-observed-adverse-effect level. The NOAEL is 0–19 years because this is the lowest exposure level with an upper confidence interval smaller than the a-priori defined safety threshold $\varepsilon = 1.5$.

<table>
<thead>
<tr>
<th>$E_\ell$</th>
<th>Cases</th>
<th>Controls</th>
<th>Lower limit CI $E_\ell$ vs. $E_0$</th>
<th>Upper limit CI $E_\ell$ vs. $E_0$</th>
<th>Lower limit CI $E_\ell$ vs. $E_{\ell-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never used</td>
<td>174</td>
<td>337</td>
<td>0.924</td>
<td>1.441</td>
<td>0.924</td>
</tr>
<tr>
<td>0–19 years</td>
<td>168</td>
<td>282</td>
<td>1.168</td>
<td>1.773</td>
<td>1.007</td>
</tr>
<tr>
<td>20–39 years</td>
<td>237</td>
<td>319</td>
<td>1.171</td>
<td>1.789</td>
<td>0.823</td>
</tr>
<tr>
<td>40–59 years</td>
<td>222</td>
<td>297</td>
<td>1.437</td>
<td>2.682</td>
<td>0.997</td>
</tr>
<tr>
<td>$\geq$60 years</td>
<td>74</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1
Raw Data and estimated confidence intervals for the risk of bladder cancer with increasing duration of exposure to all chlorinated water sources.

Fig. 1. Point estimators of the odds ratio (OR) and lower one-sided confidence intervals ‘exposed versus unexposed’ for identification of LOAEL in bladder cancer data with increasing duration of exposure to all chlorinated water sources.
From Figure 3 we can see that only the increment ‘\(E_2\) vs. \(E_1\)’ is significant (that is the lower confidence interval is larger than 1), and therefore no epidemiological trend according to the strict definition can be derived.

4. Conclusions

The proposed approach follows the intention of Lang et al. (1998), but avoids their hidden model assumption (‘slope of a trend line’). One-sided upper confidence intervals can be used for identifying the lowest-observed-adverse-effect-level (LOAEL) (comparing with an ‘efficacy’ threshold \(\delta\)) or lower confidence intervals
for identification the non-observed-adverse-effect-level (NOAEL) (comparing the safety threshold ε).

Depending on the reader’s preference, the approach can be used in a more descriptive or in a more confirmatory manner, as proposed here. All necessary information is available in Figure 1.

This approach can be used for risk ratios and differences of proportions analogously and is recommended for other non-designed dose-response studies outside epidemiology, for example, using parametric confidence intervals for the endpoint litter weight in a reproductive animal study.

Finally, this approach is simple, an important argument for using it to explain the results to non-statisticians.

Acknowledgements

This work was done during a sabbatical at the University of California at San Francisco, Department of Epidemiology and Biostatistics. Partly supported by a grant from Volkswagen-Stiftung, Hannover/Germany.

The author thanks the editor, Dr. Dieter Hauschke, Konstanz, and the referee for their valuable comments and suggestions.

References


NEUHAUSER, M., 1996: *Trend tests for a priori unknown expected value profile* [dissertation]. University of Dortmund [in German].


Prof. Dr. LUDWIG A. HOTHORN
Universität Hannover
LG Bioinformatik
Herrenhäuser Straße 2
D-30449 Hannover
Germany
E-mail: hothorn@ifgb.uni-hannover.de

Received, November 1998
Revised, July 1999
Accepted, July 1999