evaluation. In fact it reinforces the conclusion that the resulting evidence is truly equivocal, and this conclusion could not have been reached without statistical evaluation.

The example cited by the authors of a 'cluster' of occupational breast cancer does indeed have an issue with a reference population. However, this is not a spatial cluster example and is of limited relevance to assessment of spatial clustering as an exploratory tool in such studies. Perhaps this points to the need for sensitivity analysis to assumptions about study design and methods employed, rather than the rejection of statistical cluster testing.

Fixed (putative) sources of risk

The special case of a fixed source of risk (such as an incinerator, mobile phone mast, chimney, waste dump site, river or road) is in some ways easier to assess. Often an exposure surrogate is used to provide added evidence for a 'cluster' of unusual risk. Hence distance and/or direction from a source can be used in tests or models to assess linkage between incidence and the location. These will lead to probability statements about the linkage. The problem of competing risks is also apparent here as is the issue of background clustering. Are there other sources of risk in the vicinity and does the disease naturally cluster? These issues can be taken care of with careful generalized designs and models for background heterogeneity. Often these are Bayesian in flavour (see for example Wakefield and Morris,[5] and for reviews in chapter 9 of Elliott *et al*.[6] and chapter 8 of Lawson[3]).

Funding

This work was supported by NIH R01 ES012895-04A2.

Conflict of interest: None declared.

References

1 Elliott P, Wakefield J. Disease clusters: should they be investigated, and, if so, when and how? *JR Stat Soc A* 2001;**164:**3–12.
2 Lenihan J. *Bonnybridge/Denny Morbidity Reviews: Report of the Independent Review Group*. Edinburgh: Scottish Home and Health Department, 1985.
3 Lawson AB. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. 2nd edn. Boca Raton FL: CRC Press, 2013.
4 Adelman AS, Groves FD, O'Rourke K *et al*. Residential mobility and risk of childhood acute lymphoblastic leukemia: an ecological study. *Br J Cancer* 2007;**97:**140–44.
5 Wakefield J, Morris S. The Bayesian modeling of disease risk in relation to a point source. *J Am Stat Assoc* 2001;**96:** 77–91.
6 Elliott P, Wakefield J, Best N, Briggs D (eds). *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press, 2000.

Published by Oxford University Press on behalf of the International Epidemiological Association *International Journal of Epidemiology* 2013;**42**:449–452
© The Author 2013; all rights reserved.
doi:10.1093/ije/dys238

Commentary: Regarding assessments of chance in investigations of 'cluster series'

Lance A Waller

Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA. E-mail: lwaller@emory.edu

Accepted 13 December 2012

In this issue of *IJE*, Coory and Jordan[1] raise several important issues regarding accurate epidemiological views of, and public health responses to, reported clusters of disease. Some issues raised relate specifically to the role of assessments of chance (typically estimated probabilities, *P*-values or confidence intervals) in standard protocols for responding to reported clusters, whereas other issues raise foundational questions relating to the role (if any) of assessments of chance in epidemiology in general.

Whereas I agree with many of the authors' points and note these below, I offer additional perspectives

Downloaded from http://ije.oxfordjournals.org/ at University of North Texas on January 1, 2017

on three elements of the paper, namely:

- the ill-defined role of chance in typical assessments of reported clusters, focusing on context, conditioning and 'silent multiplicities';
- the proposed treatment of reports as 'cluster series', the expanded role of exposure assessment and the role of risk communication in responses to reported clusters; and
- the importance of context and conditioning in the interpretation of probability statements within epidemiology in general.

Chance and response to reported clusters

Authors Michael and Jordan point out that many protocols for response to reported clusters of disease include two related but separate issues: (i) the magnitude of the excess (typically relative) risk in the reported cluster and (ii) an assessment of how unusual an excess of this magnitude is under a null hypothesis of no local excess risk. We consider each issue in turn.

How large is the excess?

As noted by the authors, protocols typically suggest the use of a standardized relative risk estimate (e.g. standardized mortality/morbidity ratio) to compare the number of cases observed in a local area to the number one would expect to observe in the absence of a local risk increase. In the presence of established incidence rates, this calculation is fairly straightforward and noncontroversial (with possible discussion of appropriate standardization to the local age/race/sex distribution or adjustment for other local known risk factors). Typically, a reported cluster will have a higher than expected number of cases (otherwise it is unlikely to have been reported) and the estimated relative risk will be among the highest in the area, raising the question of whether it is 'too high'.

How likely (or unlikely) is it to observe such an excess in the absence of a true local increase in risk?

Since some area will have the highest local observed excess, the magnitude alone rarely serves as a completely satisfactory summary of the cluster, raising the question of whether such an increase is unusual in some way. It is here that many protocols try to assess the probability that the observed excess could have occurred 'by chance'. The phrase 'by chance' and the typical associated reports of a *P*-value suggest the existence of a standard and robust statistical approach for evaluating the likelihood of the reported cluster, but, as noted by the authors, this is not the case. The authors illustrate that this approach as typically applied and reported is difficult to interpret, and is not robust against the variety of uncertainties involved.

Chief among these uncertainties is the issue of 'silent multiple comparisons' resulting from a lack of a clear reference group for the observed excess. In other words, the original cluster report often provides a precise number of cases (e.g. from a registry) and an associated at-risk population within which we observe the cases (e.g. residents of a small census region), leading to the original estimate of the excess. However, as noted by the authors, the universe of possibilities is rarely defined, that is there is no clear standard of comparison in order to determine if the observed excess is unusually large. As a result, the authors suggest 'the theoretical construct of chance (as applied to the data) is of no pragmatic value in cluster investigations because the *P*-values and confidence intervals suffer from an extreme form of the ubiquitous statistical problem of silent multiple comparisons'.

Rather than dismiss statements of chance entirely, I feel the authors' point identifies a key need for clarity in calculation and reporting of probabilities or 'chance', specifically the recognition that any probability calculation involves a context for interpretation. More specifically, 'chance' is not something that is calculated without general assumptions; rather, I would go so far as to say any probability is a conditional probability, and the conditioning must be defined for proper interpretation. For instance, in hypothesis testing, a *P*-value is the probability of observing a more extreme value of a test statistic than observed in the data given the null hypothesis is true. Statistical power for a particular alternative hypothesis is the probability of rejecting the null hypothesis given that the alternative hypothesis is true, and the size of a hypothesis test is the probability of rejecting the null hypothesis given the null hypothesis is true. Specifying this conditioning makes clear that size and power cannot be directly compared since they are not based on the same conditions. More generally, probabilities are not simply numbers between zero and one without context, and they cannot be compared, ranked or interpreted accurately unless they are based on the same conditions.

With the idea of conditioning in hand, it becomes apparent that the issue of 'silent multiplicities' involves two issues relating to lack of clearly specified context. The first involves the multiplicity of reference groups that could be applied to the reported excess, as illustrated by the authors in their example of cancer cases observed in Brisbane, Australia. Specifically, is the observed rate high for a city, district, country or continent? In short, context matters, and reported evaluations cannot be interpreted without clear specification of this context.

The second, and more traditional, multiple comparisons problem involves the multiplicity of local standardized rates that could be compared with the null

Downloaded from http://ije.oxfordjournals.org/ at University of North Texas on January 1, 2017

hypothesis, i.e. what if we report a *P*-value for every local value? This issue arises in many 'big data' settings such as gene expression arrays, bioinformatics and -omics technologies where technology allows automated simultaneous testing of thousands of genes, proteins, etc. Statistical researchers have recognized that approaches that simply repeat individual tests are not satisfactory, nor are traditional multiple comparisons adjustments such as Bonferroni. Instead, statisticians have generated clever approaches in particular application areas such as Kulldorff's[2] spatial and spatiotemporal scan statistics for clusters (evaluating the question: 'Is the highest reported rate unusually high among the highest rates that could be observed given a constant risk across the at-risk population?' rather than 'Is the highest reported rate unusually high for this location?' [see related discussion in Waller and Gotway[3] (pp. 180–83)] and Efron's[4] evaluation of significance of results in gene expression arrays. These are challenging questions built from simple components and remain a fertile ground for further development.

Proposed alternatives to chance in cluster response

In addition to the broader issue of chance, the authors raise several specific proposals for improving response to reported clusters of disease.

Treat reported clusters as 'cluster series'

This proposal is particularly helpful in that it focuses attention on the available data (the case counts) and specifies that there are no matched controls. Instead, one needs to define a population at risk (allowing clarification of the conditioning mentioned above). Viewing the reported cluster as a 'cluster series' also sets appropriate expectations for interpretation and evaluation. The issue of determining whether the observed magnitude of observed effect is 'unusual' remains, but takes a back seat to the accurate description of the observed cases. In addition, as noted by the authors, the concept of a 'cluster series' keeps attention on the data generating the hypothesis rather than a particular test of that hypothesis on the same data generating the hypothesis.

Increase emphasis on exposure assessment in response to cluster series.

I agree that a thorough assessment of potential exposures provides a critical context for interpreting potential local risks. However, it is worth noting that exposure assessments in response to a reported cluster also have *post hoc* elements and may raise another variant of the 'Texas sharp-shooter' problem. That is, a cluster series report often encourages speculation regarding potential local exposures. For instance, a reported cluster in a population near a nuclear power plant typically raises questions regarding radiation, but the area near the plant may also be subject to industrial chemicals associated with the plant, increased traffic, different socioeconomic levels of the nearby population, etc. Any assessment of exposures in the area associated with the reported cluster series must assess how local exposures differ from exposures to similar compounds more generally, raising very similar context issues to those raised regarding the standardized case rates: how high is too high? And what is our comparison population/area? In short, exposure assessments add information but are not immune from many of the same concerns outlined in the authors' manuscript.

Chance, probability and epidemiology

On a more foundational note, the authors review two reasons for considering chance in epidemiological studies, and argue that neither applies to the specific case of cluster assessment. The first is based on the multiple small variations that occur in measurement but which are not necessary to fully understand in order to accurately measure associations between exposures and health outcomes. Put another way, we generally do not see randomness as playing a role in the cause of disease. Rather, we operate under the assumption that, if we could observe absolutely everything involved, the cause of disease would be determined without resorting to chance as an explanation, or as the authors state: 'True randomness or chance does not exist in epidemiologic data'. I feel it would be more accurate to say true randomness may not exist in disease causation. However, since we rarely observe everything we need without error, some elements of randomness exist in all epidemiological data. That said, if we assume the 'signal' of association between a cause and effect is strong enough, we can expect to observe the association through the 'noise' of unmeasured variations by modelling this noise through a random distribution. For example, knowing the exact force applied to a coin along with surface friction, wind speed, height to the ground, etc. will determine precisely whether it lands with heads or tails showing. Since we cannot observe everything involved, we rely on models of randomness to address the variation of these unmeasured causes, often providing adequate information to draw our desired conclusions (e.g. is it a fair coin?). Similarly, we apply models of randomness to address potential confounding (and other) influences across the at-risk population to infer associations in epidemiology. One typically assumes such variation is not related to the outcome and serves to add noise to the observations; then one applies a model of this noise from a probability distribution relating to the type of observation [e.g. Gaussian (normal) noise

Downloaded from http://ije.oxfordjournals.org/ at University of North Texas on January 1, 2017

for continuous outcomes, binomial noise for observing a number of successes in a fixed number of (independent) tests, Poisson noise for counts of events, etc.]. It is important to note that each distribution is a model of a particular type of noise, under certain conditions, and each probability model provides a context for summarizing unobserved variation and interpreting observed values.

The second reason given for considering chance relates to frequency-based sampling. As noted by the authors, this is easiest to conceptualize in an experimental setting such as a randomized clinical trial (RCT). In an experimental setting, subjects are randomized to treatment to effectively reduce the impact of variations in unmeasured factors. By knowing the randomization mechanism, we know the probability that each subject received a particular treatment and the design defines the distribution of unknown factors across subjects. In some observational settings (e.g. case-control or cohort studies), one cannot define this distribution through randomization but one can make design decisions (e.g. careful selection of controls) aimed to reduce potential influences of such factors on observed effects.

The contrast of experimental and observational studies is not a simple dichotomy, however, and it is useful to consider a ranking of study types by the amount of control we have over the design. As noted, an RCT offers control through randomization that defines noise induced by unmeasured factors. Standard epidemiological designs (e.g. case-control, cohort) replace control via randomization by control of comparisons in order to reduce the possibility of observing spurious associations.

In observational settings further removed from the experimental setting, e.g. case series or (as pointed out by the authors) cluster reports, we simply have a record of certain occurrences but no clearly defined reference group. In this case, one often builds probability models where the disease itself acts as a sampling mechanism selecting individuals from the at-risk population and we attempt to determine whether the selection probability varies by location or population subgroup (in particular, whether it is higher in the reported cluster area). Waller and Jacquez[5] illustrate how different models of disease propagation are implicit within particular tests evaluating disease clusters, again illustrating how context and conditioning are essential elements to understand the probabilities calculated when evaluating disease clusters.

Conclusion

Taken together, I agree with the authors that the role of 'chance' in cluster investigations is not clearly defined, uniformly interpreted or consistently applied in responses to reported disease clusters. Unlike the authors, though, I feel there is a role for probability in epidemiological assessments, but our current approaches are neither complete nor sufficient for unambiguous interpretation. As a result, more work is needed in order to provide helpful analytic tools to minimize the risk of misunderstanding.

Funding

This publication was made possible by USEPA STAR grant RD83479901. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

Conflict of interest: None declared.

References

1 Coory M, Jordan S. Assessment of chance should be removed from protocols for investigating cancer clusters. *Int J Epidemiol* 2013;**42:**440–47.

2 Kulldorff M. A spatial scan statistic. *Communications in Statistics—Theory and Methods* 1997;**26:**1481–96.

3 Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*. Hoboken NJ: John Wiley and Sons, 2004.

4 Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Am Stat Assoc* 2004;**99:** 96–104.

5 Waller LA, Jacquez GM. Disease models implicit in statistical tests of disease clustering. *Epidemiology* 1995;**6:** 584–90.

Downloaded from http://ije.oxfordjournals.org/ at University of North Texas on January 1, 2017