Assessment of chance should be removed from protocols for investigating cancer clusters

Michael D Coory1,2* and Susan Jordan3

1Murdoch Children’s Research Institute, Parkville, Victoria, Australia, 2Department of Paediatrics, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia and 3The Population Health Department, The Queensland Institute of Medical Research, Herston, Queensland, Australia

*Corresponding author. E-mail: michael.coory@mcri.edu.au

Assessment of chance, using P-values and confidence intervals (CIs), is still considered an important component of cluster investigations, as evidenced by its continued inclusion in published protocols and reports of clusters. However, true randomness or chance does not exist in epidemiological data. Instead, the concept of chance is introduced to analysis as a pragmatic device to describe the accumulation of many small variations, for which it is not economically feasible or worthwhile to uncover the causes.

The theoretical construct of chance (applied to the data) is of no pragmatic value in cluster investigations because P-values and confidence intervals from cluster investigations suffer from an extreme form of the ubiquitous statistical problem of silent multiple comparisons. As a result, they are impossible to interpret, and confuse rather than assist decision-making.

We therefore recommend that assessment of chance be removed from protocols for investigating clusters, and that there be an associated shift in emphasis to exposure assessment. We also recommend the framing of investigations of cancer clusters as case-series, rather than as retrospective cohort studies that use routine data from population-based cancer registries.

Background

Spontaneous reports of cancer clusters are common.1–3 They reflect community concern that there is a cause of such clusters related to the community’s neighbourhood, school, or workplace, and that if nothing is done there will be more cases.3–5 Not surprisingly, investigations of reported clusters can become highly visible and politically charged.2,6 In response, many government agencies follow published protocols so that all clusters are investigated according to standard procedures.3,7–9 This can help allay accusations from the community or media of cover-ups, or that authorities are not giving serious consideration to a cancer cluster.6 In this paper we concentrate on spontaneous reports of cancer clusters; however, the principles are the same for other spontaneous reports of non-communicable disease clusters, such as birth defects, multiple sclerosis, autism, and asthma.

Different protocols give slightly different emphases to different aspects of the investigation of cancer clusters. All include initial triage, careful case definition, and a detailed description of the cases in the cluster, with verification of the diagnosis, assessment of exposure, and empathetic communication of risk to the public (Table 1). It is not the intention of this paper to review these well-established steps. Instead, the paper examines the particular question of whether or not a cluster can be attributed to chance.7

The assessment of chance is typically based on a retrospective cohort analysis of routine data (e.g., data from a population-based cancer registry).2,3,7 In such an analysis, the exposed group (often based on a neighbourhood, workplace, or school) is conceptualised as being followed forward in time (although the data have already been collected) and the (age-adjusted) rates of cancer in the cohort are compared with those of an unexposed group, usually consisting of the non-cohort population covered by the cancer registry. The product of this analysis is a point estimate of the (age) standardised incidence ratio (SIR), and an associated P-value or confidence interval (CI), to assess the role of chance in the occurrence of the relevant cancer.

Many papers have pointed out that P-values and CIs from clusters should be interpreted cautiously because of the problem of multiple comparisons.5,6,8,10–15 This problem also afflicts many other applications of statistics16; however, it is particularly important for clusters because the number of comparisons in a cluster
Table 1 Major steps in cancer-cluster investigation, following initial triage

- Development of a case definition
- Verification of cases
- Description of cases in time, place, and person (e.g. age, sex, occupation, date-of-diagnosis)
- Exposure assessment, both in relation to concerns of the community and biological plausibility
- Calculation of the ratio of observed to expected cases, with associated P-values and confidence intervals (e.g. based on routine data from a population-based cancer registry)
- Assessment of whether the aggregation of cases meets the definition of a cluster (e.g. whether or not it is a chance event, based on P-values and confidence intervals)
- Judgement about whether to take public-health action (e.g. close the workplace in which the cancers occurred)
- Judgement about whether to proceed to more detailed case-control or cohort studies

analysis is unknown (i.e. silent, as opposed to being known or visible) and almost certainly very large (e.g. >10,000).17

Visible multiplicities, such as occur with pre-specified subgroup analyses, sequential monitoring of randomised controlled trials (RCTs), or genetic association studies, are difficult enough as matters that must be addressed in statistical analyses, but at least in these circumstances it can be estimated how many multiple comparisons are in play. Knowing the number of multiple comparisons allows adjustment of P-values and CIs using standard statistical methods, such as the Bonferroni correction or Dunn-Sidak multiple comparison test.

Much more difficult are silent multiplicities, such as occur with publication bias,18 reporting bias,19 or post-hoc subgroup analyses, in which it is not known how many multiple comparisons are in play, although the number is unlikely to be large. Thus, for example, there are unlikely to be more than e.g. 10–20 unpublished studies (in an investigation for publication bias), or more than e.g. 10–20 unreported outcomes (in an analysis for reporting bias), or more than e.g. 10–20 post-hoc subgroup comparisons in a study of the cause of a particular condition. Clusters are an extreme example of the problem encountered in these examples because the number of silent multiple comparisons in their investigation could be extremely large.

In the context of cluster investigations, the problem of silent multiple comparisons has been described as the ‘Texas sharp-shooter problem’ to emphasise that the boundaries of the cluster in time, place, and person are defined after the event. The statement means that a person fires at the side of a barn and then draws a bullseye around the bullet hole.17 This description helps to connect silent multiple comparisons with post-hoc analyses. Heuristically, silent multiplicities, in which the number of comparisons in play is unknown, occur with post-hoc analyses, whereas visible multiplicities, in which the number of comparisons is known, are associated with pre-specified analyses.

Despite this serious problem with the interpretation of P-values and CIs, their calculation is still considered to be an important component of cluster investigations, as evidenced by their continued inclusion in protocols7 and the emphasis that is often given to them in published reports.20–24 Indeed, even papers warning of the problems of interpreting P-values and CIs in analyses of clusters still recommend their use, with the caveat that they be interpreted cautiously.5,6,8,10–14

This paper goes further than that and argues that because cluster investigations suffer from a severe form of the problem of silent multiple comparisons, the assessment of chance should be removed entirely from cluster investigation protocols. Some readers might be perplexed by this argument. Assessment of chance (e.g. through the use of P-values or CIs or both) is considered a part of good epidemiological practice, and this is true for most areas of epidemiology (e.g. clinical trials, surveys, cohort studies, case-control studies). However, as this paper explains, assessment of chance does not assist with decision-making for cluster investigations.

The argument in this paper differs from the well-rehearsed argument against the use of P-values because they confound effect size with sample size.25 It also differs from arguments around the protracted battle between frequentists and Bayesians.26 It is possible that Bayesian methods, by incorporating subjective judgements into numerical calculations, provide another perspective on the investigation of clusters.27 However, if only to show that the available data ‘cannot lead to identical beliefs within the scientific community,’28 Nevertheless, cluster-investigation protocols currently use frequentist methods (i.e. P-values, CIs), and this paper concentrates on the assessment of chance as measured by these metrics and demonstrates their utility when deciding upon the actions to take in response to a reported cluster.

It is important to note that this paper concentrates on spontaneous reports of clusters in the general population. Another and different body of work that is labelled with the term ‘cluster’ has to do with the pre-specified analysis of a database (e.g. a population-based cancer registry) that aims to identify temporal and spatial patterns of disease clustering and their relationship to putative exposures. Frequentist P-values and CIs or Bayesian probabilities and credible intervals are useful for these pre-specified analyses, in which the number of
multiplicities can be estimated (i.e. they are visible) and for which appropriate adjustment can be made.29

The meaning of ‘chance’

True randomness probably does exist in the sub-atomic world of quantum physics; however, for the problems that epidemiologists address, chance is more a reflection of the limits of human knowledge. Thus, for example, in epidemiology (and other areas of science, besides quantum physics), chance is a pragmatic way of describing the accumulation of many small variations for which it is either economically feasible nor worthwhile to uncover the causes.30,31

More specifically, chance is a convenient to describe the theoretical construct of sampling variation.32 We imagine drawing repeated random samples from a hypothetical (super) population and applying a theoretical probability distribution (e.g. a Poisson distribution for cancer incidence in cluster investigations) to describe the likely variation (imprecision) of a sample statistic under these theoretical conditions. This hypothetical model is used as a matter of convenience; there are reasons for the variation (i.e. it is not truly random and could be explained with better information).33

This pragmatic approach to chance is easiest to conceptualise for surveys based on random sampling and for RCTs. Chance is more difficult to conceptualise, but is still useful, for observational cohort and case–control studies, in which no random sampling or randomisation is done, but in which the sampling could be (arguably) conceptualised as random, after adjustment for (known) confounders.34 The key point is that chance is a theoretical model that is applied to the data rather than being an intrinsic property of the data. As with any model, use of the concept of chance must be justified on pragmatic grounds, such as that it helps with interpretation of the results or assists with decision-making based on the results.

To date, and in terms of current protocols, the theoretical construct of chance has been applied to the data available for cluster investigations. The present paper argues that this practice is of no pragmatic value because clusters suffer from an extreme form of the problem of multiple comparisons in that the number of comparisons that they involve is unknown (silent) and large. As will be discussed in the next section of the paper, this means that P-values and CIs from cluster investigations are always inconclusive and therefore do not help with decision-making about what to do about a reported cluster.

Example of the problem of silent multiple comparisons

Ten cases of invasive breast cancer were reported by women at the Australian Broadcasting Corporation (ABC) studios in Brisbane, Australia, from 1994–2006, when only 1.6 cases would have been expected had the workforce at the studios had the same age-specific rates of breast cancer as all women in the Australian population, giving a standardised incidence ratio (SIR) of 6.25. There was confusion over the improbability of obtaining such a large SIR. Initial investigations reported a P-value of 0.000001 for its probability, prompting a television documentary program on the cluster to use the title ‘One in a Million.’ However, the scientific investigation panel that was formed to investigate the cluster reported a final P-value, adjusted for multiple comparisons, of 0.04 (1 in 25).20

This statistical adjustment was based on 40 000 silent multiple comparisons, which was an estimate of the number of groups of 150 women (the size of the female workforce at the ABC studios) that could be formed from the Australian female population aged 15–64 years.20 The formula (Dunn–Sidak) for calculating this is: 1 – (1 – 0.000001)^40000 This is not necessarily wrong, but it highlights the arbitrary nature of the analysis.

More specifically, there is no compelling reason to frame the problem of multiple comparisons with reference to the Australian population. If it were framed with reference the city of Brisbane, the location of the workplace in which the cases of cancer occurred, the P-value would be ~0.006 (1 – (1 – 0.000001)^6500), which would still be considered statistically significant according to the accepted, although arbitrary, P-value threshold of 0.05. Alternatively, if the problem were framed with reference to all other countries with a high income similar to that of Australia, then the P-value would be ~0.99 (1 – (1 – 0.000001)^15000000), which suggests that there is almost certain to be a report of a cluster of breast cancer cases from somewhere.

Besides the number of workplaces with a population of 150 workers, there are other sources of multiplicity in cluster analyses. Women at the ABC studios in Brisbane could have reported a cluster of ovarian, bowel, brain, or any other type of cancer.8 Also, the workplace had been open since the 1970s so that, setting aside latency (which could narrow the temporal window somewhat), the time period within which the cases of breast cancer occurred is also a source of multiplicity.8 Figure 1 shows that the P-value can vary greatly, depending on arbitrary assumptions about the number of multiple comparisons.

An alternative to using P-values for assessing statistical significance is the use of CIs, which have the advantage of focussing attention on the point estimate (e.g. SIR=6.25). However, CIs are based on the same frequentist mathematics as are P-values and are often used as surrogates for tests of hypotheses. In the case of cluster investigations, interest usually centres on whether the CI includes the null value (i.e. SIR=1.0), which is equivalent to conducting a test of a hypothesis. Confidence intervals can be adjusted (widened) to account for multiple comparisons, but suffer the same problem as P-values in that the
number of comparisons is unknown and therefore any adjustment is arbitrary.

Other problems with current protocols
Besides the problem of silent multiple comparisons, there are several other problems with cluster investigations that obtain \( P \)-values or CIs based on routine data. For example, data are usually not available to account for potential confounders (other than age); people can move in and out of the neighbourhood, school, or workplace of interest; and many exposure–cancer pairs have long induction and latency periods.\(^3\)

Also, if the only investigation of a cluster is a standard analysis of routine data, then there is a small probability that something important might be missed, because of overemphasis of the results of statistical analyses, rather than active consideration of a possible common environmental cause (exposure) of a disease or other event. For example, in Woburn, Massachusetts, 12 cases of childhood leukaemia were reported between 1969 and 1979, when 3 cases were expected.\(^3\) Under current protocols, the SIR of 4.0 for this result could easily be explained as sampling variation (chance), after adjustment for multiple comparisons. However, there was concern about industrial contamination of two water-supply wells, and it was only when households that received water from these particular wells were specifically studied that a causal link became a distinct possibility.\(^6,35,36\)

Change to improve protocols
Assessing whether a reported cluster could be to the result of chance is meaningless in the way in which this is typically done, and such assessment should be removed from cluster-investigation protocols. This would require an alternative definition of a cluster. The current definition is based on chance (i.e. ‘an aggregation of cases that is not explainable as a chance event’),\(^7,9\) and implies that clusters should be assessed with \( P \)-values and CIs.

A more useful definition was suggested by Rothman more than 20 years ago.\(^17\) He pointed out that the question of interest in a cluster investigation is whether the cases have causal mechanisms that include a common component. Exposure assessment is the key here, not \( P \)-values or CIs based on routine data. The history of cluster investigations shows that for (rare) clusters, in which a common component cause was identified (and required public health action), exposure levels were high and well-documented. Some examples are given in Table 2.\(^37\) In contrast, virtually all reports of clusters are false-positives and have as a defining feature community concern about exposures (related to a neighbourhood, school, or workplace) that are poorly characterised, varied, and of low concentration.\(^17\) Figure 2 is an illustrative flow chart for cluster investigations that focuses attention on exposure assessment.

**Figure 1** ABC cluster, showing adjusted \( P \)-values for different numbers of multiple comparisons. In cluster investigations the number of multiple comparisons is large (>10000) and unknown. In such circumstances chance is not a useful theoretical construct.

**Table 2** Examples of discoveries of the causes of cancer from clusters\(^a\)

<table>
<thead>
<tr>
<th>Year</th>
<th>Cancer, setting, cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>1775</td>
<td>Scrotal cancer in chimney sweeps exposed to soot from coal</td>
</tr>
<tr>
<td>1926</td>
<td>Lung cancer in coal miners exposed to radioactive dust</td>
</tr>
<tr>
<td>1929</td>
<td>Osteosarcoma in watch dial painters exposed to radium</td>
</tr>
<tr>
<td>1952</td>
<td>Bladder cancer in workers in a synthetic dye factory exposed to aromatic amines</td>
</tr>
<tr>
<td>1965</td>
<td>Mesothelioma and lung cancer in workers exposed to asbestos</td>
</tr>
<tr>
<td>1971</td>
<td>Vaginal cancer in daughters of women given diethylstilbestrol during pregnancy</td>
</tr>
<tr>
<td>1974</td>
<td>Angiosarcoma of the liver in chemical workers exposed to vinyl chloride monomer</td>
</tr>
<tr>
<td>1981</td>
<td>Kaposi sarcoma in homosexual men with AIDS</td>
</tr>
</tbody>
</table>

Completing this change in emphasis towards exposure assessment, we propose changing the framing of the epidemiological component of a cluster investigation to a case series. One benefit of this framing is that it discourages the use of $P$-values and CIs, in that few people would add $P$-values or CIs to a report of a case series. Traditionally, cluster-investigation protocols have conceptualised the epidemiological component of the investigation as a retrospective cohort study (based on the available, routine data). This has not proved helpful because it carries the expectation that the associated $P$-value and CI have a standard interpretation, as they would in a retrospective cohort study that was specifically designed to answer a pre-specified research question. As discussed in this paper, for spontaneous reports of clusters, $P$-values and CIs do not have a standard interpretation.

Besides discouraging the use of $P$-values and CIs, framing a cluster investigation as a case-series has the additional benefit of emphasising that most reported clusters are false-positives, just as are most case series. Thinking about study design in terms of whether the results might be false-positives (or false-negatives) is associated with the recent re-evaluation of hierarchies for study design, and in particular the recognition that there are different hierarchies for evaluation than for discovery.38,39

For evaluation, systematic reviews of RCTs are at the top of hierarchies of study design, whereas case series are at the bottom. Well-conducted RCTs of adequate sample size have high sensitivity (low false-negative rate) and high specificity (low false-positive rate). For discovery, the ranking of study designs is reversed, but it is not possible to identify a study design that combines high sensitivity with high specificity (as for

---

**Figure 2** Illustrative flow chart for cluster investigation that emphasises exposure assessment

- **Dialogue with community**
  - Concern about one specific exposure?
    - yes: Environmental assessment with immediate clean-up, if required, regardless of whether linked to cancer.
    - no: Concern about poorly characterised, heterogeneous exposures associated with neighbourhood, school or workplace from which the cluster arose.
- **Are the exposure levels high enough to cause cancer?**
  - yes
    - Media or political pressure
      - yes: Take detailed history from cases
      - no: Suspicion of common component cause?
        - yes: Take action to remove risk
        - no: Stop investigation
          - Explain reasoning to community
    - no: Stop investigation
      - Explain reasoning to community

---

Note: in every (rare) true-positive cluster, identified to date, exposure levels were high and well-documented.
However, whatever the context, the pattern is somewhat smaller in other contexts (e.g. adverse drug context. It is large for reported clusters, but could be false-positive problem for a case series depends on the context.38,39 which would (appropriately) involve investigations would need to be subsequently investigated for evaluation,38 which would (appropriately) involve RCTs for evaluation). Instead, case series are placed at the top of the hierarchy for discovery because they have a high sensitivity for discovery (few false-negatives); however, they have the trade-off of having extremely poor specificity (most of what are called cases are false-positives).

Table 3: Cross-classification of spontaneous reports of clusters versus clusters with a common component cause

<table>
<thead>
<tr>
<th>Spontaneous community report of cluster?</th>
<th>Yes</th>
<th>No*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common component cause?</td>
<td>a</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>d</td>
</tr>
</tbody>
</table>

False-negative proportion = \( \frac{a}{a+b} \); \( b \ll a \); therefore, the false-negative proportion is close to 0.0.

False-positive proportion = \( \frac{c}{c+d} \); \( c \gg d \); therefore, the false-positive proportion is close to 1.0.

Positive predictive value = \( \frac{a}{a+b+c} \); \( c \gg a \); therefore, the positive-predictive value is close to 0.0.

Reported clusters are similar to case-series: there is a high sensitivity for new discoveries (few false-negatives; a is large relative to b), poor specificity (many false-positives; c is large relative to d), and low positive-predictive value (c is much larger than a).

*Not measurable; inferred from the history of cluster investigations.

RCTs for evaluation). Instead, case series are placed at the top of the hierarchy for discovery because they have a high sensitivity for discovery (few false-negatives); however, they have the trade-off of having extremely poor specificity (most of what are called cases are false-positives).

History shows that this accurately reflects the situation with clusters. True-positive clusters, in which the cases have a common component cause (and where public health action is needed) are extremely rare. Virtually all reported clusters are false-positives (see Table 3). Of course, the actual size of the false-positive problem for a case series depends on the context. It is large for reported clusters, but could be somewhat smaller in other contexts (e.g. adverse drug events).40 However, whatever the context, the pattern is the same: the trade-off for high sensitivity for discovery is low specificity. As a corollary, the probability of finding a common component cause among the cases in a particular cluster (positive predictive value) is low (Table 3)

A final benefit of framing a cluster as a case-series is that it clarifies that apparent discoveries from cluster investigations would need to be subsequently investigated and refined through study designs more suited to evaluation, which would (appropriately) involve P-values and CIs. An example is the cluster of mesothelioma cases caused by environmental exposure to asbestos at Wittenoom, Australia. A case-series established a common environmental cause for these cases (i.e. exposure to asbestos) and a subsequent cohort study was later used to answer questions about whether the risk was higher with childhood exposure and to assess the risk associated with low levels of exposure.41

Discussion
Investigations of one-off, spontaneous reports of clusters are sometimes highly visible and politically charged. In such circumstances, it would be desirable to have an objective, externally verifiable, and data-driven process for deciding whether the cluster raises cause for concern and whether public-health action is required to address it. The problem is that P-values or CIs from a cluster investigation are not objective. They rely on a series of subjective judgements about the population at risk (e.g. time frame, geographic boundaries) and, in connection with this, subjective judgements about the number of silent comparisons involved in characterizing the cluster.

We have proposed removing the assessment of chance (P-values and CIs) from cluster-investigation protocols. Associated changes include the adoption of Rothman’s definition of a cluster (based on common component causes), a framework for investigating clusters that concentrates on exposure assessment, and recognition that a reported cluster is a case series (and should not be used to define the exposed group in a retrospective cohort study based on routine data). This is standard epidemiology, applied in an appropriate way to cluster investigations. Adoption of these changes might be facilitated by avoiding use of the word ‘cluster’ altogether. A better label might be, e.g. ‘cancer series’.

Some readers might be concerned that the proposed framework will close off the possibility of using reported clusters (i.e. cancer series) to identify (previously undiscovered) causes of cancer. History shows that such concern is not warranted.5,15,37,41,42 Some readers might be concerned that the proposed framework will close off the possibility of using reported clusters (i.e. cancer series) to identify (previously undiscovered) causes of cancer. History shows that such concern is not warranted.5,15,37,41,42 Spontaneous reports of clusters of cancer, for example, have high sensitivity for identifying new causes of cancer. This was emphasised at the Cancer Clusters Hearing before the Senate Cancer Coalition (in the United States), which also heard that the rare instances of new causes of cancer occurring in clusters were identified by observant clinicians or public-health officials, who identified a common exposure to a causative factor.37 P-values or CIs were not required and would not have been helpful.5

This framework gives higher priority to judgements of experts about exposures than do the available statistical data. Although a long list of potential causative agents might be drawn up, many of them can/will be quickly dismissed by expert assessment of measured exposures, biological plausibility and pre-existing evidence. Giving priority to expert judgement (even when based on explicit criteria) will be counter-intuitive to some public-health professionals, who
are told in their training that human observers are frail meters of truth, too prone to see what they expect to see and too trusting in anecdotal evidence. Only good data and strong statistical methods can protect the human mind from its own biases.

Although such thinking is helpful in most areas of epidemiology, the problem with applying it to cluster investigations is that, by definition, it is not possible to apply strong statistical methods in such situations. Regardless of which theoretical probability distribution is used (e.g. normal, Poisson, binomial, extra-Poisson, negative binomial, etc.), P-values and CIs obtained from clusters test a hypothesis on the same data that generated the hypothesis, and any external calibration of the P-value or CI for multiplicities in the cluster is so arbitrary as to be useless because the number of multiplicities is unknown (silent) and large (in the tens of thousands). There is no statistical solution to this problem.

One issue we have not addressed in this paper is whether reporting the point estimate of the SIR is useful in cluster investigations. For example, Neutra suggested that ‘a cluster is not a chance event if there are at least 5 cases and the relative risk is at least 20.’ He supported this suggestion with two examples of true-positive clusters: 11 (homeless) men with methaemoglobinemia and 5 young women (post-menarchial adolescents < 20 years of age) with clear-cell carcinoma of the vagina. In both of these examples the exposure levels were high and well documented: sodium nitrite for the men and diethylstilbestrol (in mothers) for the young women. This is consistent with the argument in this paper that in all previous circumstances in which a common cause for the cases in a cluster was identified, the exposure levels were high and well documented. P-values or CIs were not required, although a heuristic approach, based on an SIR of at least 20, could be helpful. At the other extreme, an SIR close to 1.0 might be an effective way of allaying community concerns about a reported cluster that is clearly a false-positive occurrence. However, the SIR might be close to 1.0 because the analysis was of a non-specific, small area and did not concentrate on the exposure of concern (e.g. the Woburn cluster of cases of childhood leukaemia). On balance, we think the point estimate of the SIR (based on routine data) is useful secondary information, but that the investigation should concentrate on exposure assessment.

Proponents of continuing to use P-values for cluster investigations might argue that the concerned community (and the media) want to know whether an apparent excess of cases (as measured by the SIR) is to the result of chance. Little research has been done on this topic; the evidence that is available suggests that the public are not convinced or reassured by statistical reasoning.

More research is needed on the type of information the public wants from cluster investigations, as it is needed in other areas of communicating risk to the public. In the meantime, Rothman’s comment of more than 20 years ago still provides a useful reference point: ‘The community wants to know one of two things. If it is convinced that a certain exposure is responsible, it wants to know how fast and how completely the exposure can be removed or lessened. In this instance, an epidemiologic study can delay the cleanup and divert precious resources from it. On the other hand, the community may want to know what explains the cluster. In this instance, it would virtually always be more worthwhile for the community to ask for exposure assessment or to contribute in some way toward large potentially explanatory studies in other populations than to push for a small, inadequate study of the cluster itself.’

**KEY MESSAGES**

- Assessment of chance, using p-values and confidence intervals, is still considered an important component of cluster investigations; as evidenced by its continued inclusion in published protocols and reports of clusters.

- True randomness or chance does not exist in epidemiological data. Instead, the concept of chance is introduced to analysis as a pragmatic device to describe the accumulation of many small variations, for which it is not economically feasible or worthwhile to uncover the causes.

- The theoretical construct of chance (applied to the data) is of no pragmatic value in cluster investigations because p-values and confidence intervals from cluster investigations suffer from an extreme form of the ubiquitous statistical problem of silent multiple comparisons.

- As a result, they are impossible to interpret and confuse, rather than assist, decision-making.

- We therefore recommend that assessment of chance be removed from protocols for investigating clusters and that there is instead an associated shift in emphasis to exposure assessment.

- We also recommend framing the cluster investigation as case-series, rather than a retrospective cohort study that uses routine data from a population-based cancer registry.
Conflict of interest: None declared.

References