Risk Analysis, Vol. 28, No. 1, 2008 DOI: 10.1111/j.1539-6924.2008.01009.x

Aggregating Disparate Epidemiological Evidence: Comparing Two Seminal EMF Reviews

Michael J. O'Carroll[1] and Denis L. Henshaw[2]*

Two seminal reviews (IARC, 2002; CDHS, 2002) of possible health effects from power-frequency EMFs reached partly different conclusions from similar epidemiological evidence. These differences have an impact on precautionary policy. We examine the statistical aggregation of results from individual disparate studies. Without consistent exposure metrics, the advantage of meta-analysis to estimate magnitude of effect is lost. However, counting positive and statistically significant results yields important information. This is not a substitute for meta-analysis, but a fall-back when meaningful meta-analysis is not available. Representative results from 33 independent adult leukemia studies tabled by IARC yielded 23.5 positives ($p \approx 0.01$) and 9 significant-positives ($p < 10^{-7}$). From 43 representative results from CDHS, there were 32 positive ($p < 0.001$) and 14 significant-positives ($p < 10^{-12}$). There were no significant-negative results in either list. Results for adult brain cancer gave a similar, but less clear, message. Childhood leukemia EMF studies have been sufficiently comparable to allow selective pooled analysis, which was important in classifying carcinogenicity. Aggregating all the studies suggests that results for childhood leukemia are not stronger, numerically, than those for adult leukemia. CDHS did not note the number of significant-positives, but noted the meta-analytic summary and the number of positives, forming a view about the strength of these findings. IARC shows no evidence of considering the aggregation of results other than subjectively. It considered individual studies but this led to a tendency to fragment and dismiss evidence that is intrinsically highly significant. We make recommendations for future reviews.

KEY WORDS: Adult leukemia; aggregating evidence; brain cancer; childhood leukemia; electric and magnetic fields; EMFs; health effects; risk; statistical significance

1. INTRODUCTION

Our motive for this article has been to try to understand how two seminal reports from major health bodies, reviewing the possible health effects of exposure to power frequency electric and magnetic fields (EMFs), reached different conclusions from what was largely the same body of evidence. While there are constitutional and procedural differences between the review bodies, we have focused on a striking difference in how they went from critical review of the many individual studies of EMF health effects to a summative assessment of the overall weight of evidence.

The review bodies were the International Agency for Research on Cancer (IARC), an agency of the World Health Organization, and the California

[1] Vice-Chancellor's Office, University of Sunderland, Chester Road, Sunderland SR1 3SD, UK.
[2] H H Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol, BS8 1TL, UK.
* Address correspondence to Denis L. Henshaw, H H Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol, BS8 1TL, UK; tel: +44 1179 260353; fax: +44 1179 251723; d.l.henshaw@bris.ac.uk.

EMF Program Team of the California Department of Health Services (CDHS). Both published the reports of their reviews of EMFs and health in 2002. Both rated power-frequency EMFs as "*possibly carcinogenic to humans*" (the IARC Class 2B), on the basis of epidemiological evidence relating to childhood leukemia. In respect of all other cancers, IARC concluded the epidemiological evidence was "*inadequate,*" whereas CDHS concluded it was "*limited*" for four other health outcomes, including two cancers. The "*limited*" assessment supports Class 2B for the agent.

There have been other reviews, before and since 2002. For example, the NRPB reviews in the United Kingdom have consistently recognized the possibility of cause of cancer, but did not use a formal classification system for assessment. Some reviews never reached publication, for example, in the United States the NCRP review in 1995, though its conclusions were leaked. The subject of power-frequency EMFs has been controversial. In the 1990s there were calls to halt research funding on the basis that any potential risk had been dismissed. However, the evidence of adverse health effects has persisted to the point that precaution against exposure to EMFs is now being considered.

The U.S. National Institute of Environmental Health Sciences (NIEHS) EMF-Rapid Program concluded in 1998 that the evidence for both childhood and adult leukemia supported a 2B classification, the latter being "somewhat weaker" and specifically for "chronic lymphocytic leukemia in occupationally exposed adults." Two key pooled analyses, Ahlbom *et al.* (2000) and Greenland *et al.* (2000), reinforced concerns, showing, by statistical aggregation, that the fragmented findings for childhood leukemia became stronger when pooled, the former revealing a twofold increase in risk associated with time-weighted average magnetic field exposures above 0.4 μT, the latter a 1.7-fold increase above 0.3 μT.

When IARC made its formal 2B classification in 2002 against this background, a basis was set for precautionary policy that is under development in several countries and in the WHO. Policy based on the risk of childhood leukemia alone tends to be limited because the normal incidence is comparatively rare and the attributable risk very small. The question of additional risk of other diseases then becomes important. Hence the California review, by recognizing five health outcomes corresponding to the 2B classification, challenges the limitation of proportionate precautionary measures to those of very low cost.

An understanding of the differences between these seminal reviews is therefore important to the present development of precautionary policy for EMFs.

2. THE TWO SEMINAL EMF HEALTH REVIEWS OF IARC AND CDHS

Both IARC (2002) and CDHS (2002) evaluated the possible risks to public health from EMFs at supply frequency. CDHS considered only power-frequency EMFs, whereas IARC considered other frequencies but specifically assessed power frequency (or ELF). IARC assessed carcinogenic risk whereas CDHS assessed both carcinogenic and other health outcomes. Nevertheless, the two reviews had a large area of common ground in the body of evidence relating to power-frequency fields and various cancer outcomes.

The formative work for IARC (2002) was carried out at a Working Group meeting in June 2001. The CDHS program extended over several years with *Consultation Draft 3* published in April 2001. IARC (2002) did not refer to recent CDHS drafts but did refer to an earlier progress review (Neutra *et al*., 1996). The final report of CDHS (2002) referred to IARC (2001 in press) and particularly addressed the question of their differing conclusions.

IARC (2002) listed some 800 references, covering both ELF (mainly power-frequency) and static fields. CDHS listed some 400 references (ELF only).

The IARC classification system formally combines assessments of evidence in humans (essentially epidemiology) and evidence in animals. Both reviews were agreed in assessing the evidence in animals as "*inadequate.*" A particular difficulty in relation to childhood leukemia is that animal studies could be considered inappropriate because there is no animal model for acute lymphoblastic leukemia, the common leukemia type in children. Both reviews were agreed in assessing the evidence in humans in relation to childhood leukemia as "*limited,*" and hence were led by the formal classification system to the overall IARC 2B assessment.

CDHS (2002) uses another formal system of assessment, a "*qualitative Bayes*" approach, which is a central feature of the review and an interesting innovation. However, for the purpose of comparison, they also provide assessments on the IARC classification system. The five health outcomes identified by CDHS as each warranting IARC 2B classification of EMFs were childhood leukemia, adult leukemia,

adult brain cancer, miscarriage, and amyotrophic lateral sclerosis (ALS), a form of motor neurone disease.

This article compares the bases of epidemiological evidence in the two reviews, specifically for adult leukemia and adult brain cancer. Both reviews, directly or indirectly, consider selection and quality of studies, and the epidemiological holy trinity of chance, bias, and confounding at some length. We find that the more material differences lie in their approach to aggregation rather than in the body of evidence.

3. STATISTICAL AGGREGATION OF DISPARATE EVIDENCE FROM EPIDEMIOLOGICAL STUDIES

It is not unusual to find a range of reasonably independent epidemiological studies, each with its limitations and statistically weak findings, but nevertheless with an overall tendency to indicate a possible effect. One way of aggregating the evidence from such studies is by meta-analysis or pooling, which may be defined in slightly different ways.

This has the advantage of estimating the magnitude of an effect and providing confidence limits from the aggregate evidence. Such estimates are most meaningful when aggregating on a like-for-like basis with regard to exposure metric, specificity of cases, relevant subsets of population, and study methods.

Sometimes, the evidence is more disparate, so that only limited numbers of similar studies can be pooled to give very meaningful estimates of parameters such as risk estimates representing magnitude of a possible effect. For example, in the context of EMF, the majority of studies have been concerned with the effects of exposure to magnetic fields. Studies may vary according to type of exposure (residential, occupational), subsets of population (gender, race, age, susceptibility), exposure metric or proxy (measurement, proximity, job title, average, peak), or risk measure (odds ratio or standardized incidence rate), and so on.

While taking account of the caveats and qualifications relating to significance and hypothesis testing, as discussed, for example, in Rothman and Greenland, 1998, ch. 12), it is nevertheless possible to make some assessment of the strength of aggregate disparate evidence. This may be useful in supporting formal assessment of evidence, in comparing different aggregate sets of studies, and in comparing different conclusions reached by review bodies.

By "disparate" we mean evidence that does not readily support meaningful meta-analysis. By implication this may relate to a broad underlying hypothesis, such as a class of exposures through varying metrics affecting biological systems in different ways among differently susceptible populations manifesting in a range of health outcomes showing only weak associations in the general population. In this broad sense "disparity" is not necessarily the same as "heterogeneity" as sometimes evaluated within meta-analysis. The present situation is not so broad, the main disparate feature being lack of a well-defined common exposure metric, especially for occupational exposure.

The two reviews each address a range of different health outcomes that might lead to compound hypotheses such as causation of both childhood and adult leukemia (or both acute and chronic), or alternatively of one and not the other, but we shall consider them more specifically, as did the reviews. CDHS did briefly address the implications for one outcome of findings for another, and the IARC evaluation structure addresses the carcinogenicity of an agent rather than hypotheses for specific outcomes, but neither review formulated or examined compound hypotheses *per se.*

This article illustrates two simple methods of aggregation: counting numbers of positive findings and counting numbers of statistically significant-positive findings. The more disparate the studies and findings considered, the blunter the implied hypothesis, whose negation is the null hypothesis under examination. For example, aggregating both residential and occupational studies implies a hypothesis that both "exposures" are causal risk factors for the specified disease. That is more demanding than a choice of either sharper hypothesis with a more consistent exposure. It is not the purpose of this article to provide a formal analysis of sharp, blunt, and compound hypotheses.

Some epidemiologists might feel that such a simple method of aggregation is too simplistic to consider and that epidemiology has long progressed to more sophisticated analyses. However, such simplistic aggregation is fundamental statistically and provides the sort of elementary test that should always be considered prior to more sophisticated analysis, especially if it yields an unexpected result.

Thus, these counting methods are not a substitute for meta-analysis or pooling, when available, but

can be a fall-back for when they are not available. These are indicative, rather than conclusive, methods. Not all of the statistical information is used. For example, the varying size of studies is lost in counting positive results. On the other hand, there is some importance of different studies when they are independent. Counting significant results does reflect the statistical strength of the findings, though not the statistical power of the studies, so it partly overcomes the problem of failing to discriminate between studies of different size and power. We do not advocate these methods as a panacea, but we do suggest that in the absence of anything better, they should not be overlooked. What is surprising in this instance is that there is a similar underlying statistical strength of data in both reviews, partly observed in one but seemingly overlooked in the other.

4. GENERAL COMPARISON OF EVIDENCE BASES FOR ADULT LEUKEMIA

IARC selected and tabled results including odds ratio (OR) or standardized incidence ratio (SIR) with confidence interval (CI) data from 37 (33 independent) human epidemiology studies for adult leukemia, and CDHS did so for 43. However, despite the reviews' publication in the same year and despite the common reference to previous reviews, these sets of studies had surprising differences. Both reviews identify residential and occupational studies specific to adult leukemia. IARC's 37 included 6 residential whereas CDHS's 43 only included 2 residential.

The 43 studies listed by CDHS are derived principally from the same reference source (Kheifets *et al.*, 1997a). Of the 41 occupational studies, 17 are included in the IARC tables for adult leukemia, 18 are not (of which 5 are, however, listed in IARC's references), and the other 6 refer to similar studies by the same authors (e.g., with different dates), so may overlap.

Of the 32 occupational studies considered by IARC, after deducting 17 common and 6 similar studies, there remain 9 that are not in CDHS. Of the 6 residential studies listed by IARC (Table 25), only 1 (Severson 1998) is listed in the CDHS table for adult leukemia. The second residential study listed by CDHS is of Wertheimer and Leeper, (1982), which is not in IARC's Table 25.

Of the 41 occupational studies listed by CDHS, there are two sets of multiple studies from the same source (three from Theriault *et al.* (1994) and two from Tynes *et al.* (1994); the bibliography lists two studies by Theriault *et al.* (1994) and two by Tynes *et al.* (1994)). IARC lists three studies by Theriault *et al.* (1994), in Quebec, France, and Ontario, and a fourth updating paper by Miller *et al.* (1996).

Both reviews take account of the caveats and qualifications in the various studies, comment on their shortcomings, and draw on previous reviews in that respect. Neither review body is unaware of these qualitative considerations. Both are aware of the potential for bias and confounding and both address this specifically. Their conclusions, however, do give different weight to the human epidemiology studies in aggregate; this was the main observation in the CDHS comparison of the two reviews and reasons for their differences.

This article examines the statistical aggregation of evidence for the two reviews' sets of studies. Given the different conclusions from the two reviews, it might be expected that the content of their respective sets of studies, albeit overlapping, might differ in the strength of evidence for association. The CDHS conclusions drew on some aggregate statistics to support association, whereas IARC found limitations in the separate studies and did not support association. It was surprising, therefore, to find that the aggregate statistics for the IARC set of studies showed similar support for association as did the CDHS set.

5. STATISTICAL AGGREGATION OF THE CDHS SET OF ADULT LEUKEMIA STUDIES

In the CDHS set of adult leukemia studies, there are similar relative risks or odds ratios for the residential and the occupational studies. While the exposures are disparate between residential and occupational studies, the strengths of association are similar. The summary table (fig. 8.1.1, p. 121) combined one odds ratio (OR) result, with its 95% confidence interval (CI), from each of the 43 studies. Taking all 43 studies together, the meta-analytic summary was OR = 1.2 with CI = 1.12–1.24. (The CI was given in the draft 3 CDHS report but not in the final version.) The summary notes that 29 had OR > 1 with $p \leq 0.01$. That is, in aggregate, the occurrence of positive results is statistically significant at a 99% confidence level. The selection of studies and of results from each study was derived principally from a previous reviewer (Kheifets *et al.*) and adopted by CDHS in preference to introducing its own selection. Hence the 43 results were taken as reasonably independent

Table I. Aggregation of the Adult Leukemia (AL) Studies Considered by CDHS (2002)

CDHS/AL (1 per Study)	No. of ORs	Positives	*P*-Value for Positives	Significant-Positives	*P*-Value for Sig-Pos*
Residential	2	2	0.25	1	0.049
Occupational	41	30	0.002	13	1×10^{-11}
Total	43	32	0.001	14	1×10^{-12}

*One-sided, $p < 0.025$.

and representative of a random sample of the population of all possible relevant studies.

There were six results with OR = 1.00 within the truncation of the report. It is more appropriate to count such results as half negative and half positive, as that would give an unbiased estimate of the true value (50%) under the null hypothesis. Some studies in other sets have a coarser truncation to only one decimal place, with a more substantial truncation bias. While CDHS deploy what it calls the sign test, it uses a biased version of it that substantially understates the strength of evidence. In the above, including results with OR = 1.00 as half positive and half negative gives 32 positive results with $p < 0.001$, which is highly significant.

A much stronger statistical observation, not made by CDHS, is the number of significant-positive results. They are results with 95% confidence intervals wholly above 1. Although the intervals may be based on two-sided *p*-values of 0.05, they invariably correspond to one-sided values of 0.025 for positive results. There are no significant-negative results in the reviewers' lists for adult leukemia. Whether the confidence limits have been calculated by a fully frequentist approach or by inference from sample to whole population, each occurrence of a significant-positive result will have, by the same statistical model as used in the calculation, a probability $p < 0.025$.

There are nine such occurrences, that is, strictly significant-positive results from the 43 listed results, with lower confidence limit (CL) strictly > 1, plus five results with lower CL = 1.00, and no significant negatives. The significance boundary is different from the 50-50 split for simple positives, so that a marginal occurrence with lower confidence limit equal to 1 might now be counted as with $p = 0.025$ for the occurrence. Although the truncation may slightly bias an estimate of a true value under a null hypothesis, it will be a good approximation as long as the truncation error (here, 0.005) is small compared with OR –1, which it is.

Therefore, such marginal occurrences of significant-positives should reasonably be fully counted as instances with $p = 0.025$, giving 14 in all. As long as these results are independent and represent a random sample, and considering only random error and not bias or confounding (which have been addressed in the reviews), the probability of 14 such results out of 43 can be calculated by the cumulative binomial distribution as about 10^{-12}, which is extremely significant. Even that is conservative, for most of the separate *p*-values will be strictly less than 0.025. If the five marginal occurrences were only counted as halves there would still be 11.5 occurrences with aggregate *p*-value approximately 10^{-8}, which is still extremely significant, although that would not be the appropriate form of counting.

Although CDHS did not note the number of significant positives, it did note the meta-analytic summary and the number of positives, and formed a view about the strength of these findings that led it to give them greater weight than, seemingly, did IARC. The aggregation of the studies considered by CDHS is summarized in Table I.

6. STATISTICAL AGGREGATION OF THE ADULT LEUKEMIA RESULTS CHOSEN BY IARC

IARC discusses a range of adult leukemia studies and selects 37 studies with ORs or SIRs with CI data for summary description in Tables 25, 29, and 30 in their work. The tables list some 176 results, including multiple results from single studies, and including both high- and low-exposure categories. These are not independent, for example, some are totals of other results for subtypes of leukemia, so aggregating them by cumulative binomial distribution would not be valid.

It is surprising however, not least since low-exposure categories may dilute the overall apparent significance that simply lumping all the IARC-reported results together (omitting only base or

Table II. Aggregation of Studies of Adult Leukemia in IARC (2002) on the Basis of Selection Criteria to Identify One Representative Result per Study

IARC/AL (1 per Study)	No. of ORs	Positives	*P*-Value for Positives	Significant-Positives	*P*-Value for Sig-Pos
Residential	5	3.5	0.19–0.5	2	0.0059
Cohort occupational	17	11.5	0.07–0.17	4	0.0007
Case-control occupational	11	8.5	0.03–0.11	3	0.002
Total	33	23.5	0.007–0.018	9	1×10^{-7}

reference levels) would show an apparent strong aggregation, with 111.5 positive results and 31 significant-positives out of the 176. If the 176 results were independent and a random sample, those counts would have *p*-values of 0.0003 and 3×10^{-15}, respectively, which we note for reference when considering the effect of selecting more independent subsets of results. Truncated marginal ORs or lower CLs are again counted as explained above for CDHS.

In order to obtain a more independent set of results for aggregation, select at most one representative result from each study, using a common set of selection criteria:

- Omit studies and results that do not record either the OR or the CI.
- Where there are multiple results for subtypes of leukemia, select only the total or "all leukemias" results, if available, so that subtype results are not repeated. While this loses specificity, and so may dilute findings, the alternative would be to apply the same specificity throughout all selected studies. Similarly, take Theriault *et al.* (1994) combined cohort results, not the separate ones for France, Quebec, and Ontario.
- Where there are separate results tabled for different exposure bands from the same study, select only the highest band, so that the most relevant test to detect an effect (positive or negative) is used. That will typically be with a cutpoint at 0.2 μT, which is lower than the principal categories of Ahlbom *et al.* and Greenland *et al.* for childhood leukemia.
- Where there are separate results for different occupations, select the results for the occupation likely to be most exposed, and if that is not known, select the most populous result.
- Where there are different results for males and females, and no combined gender results, select the most populous results (usually males). While this loses specificity, the alternative would be to seek separate results for males (or females) throughout all selected studies.
- Where the choice remains ambiguous on the above criteria, and yet would make a difference, select an appropriate balance, e.g., half positive and half negative.
- Where two articles from the same source draw on the same data set but analyze it in different ways, select only one result using the above criteria.

We emphasize that these are our selection criteria. They were not applied by either of the review bodies.

Such a selection leads to the summary in Table II. While still showing highly significant results, selection has moderated, not exaggerated, the strength of the crudely aggregated original data. For example, the results would have been slightly stronger if the significant-positive finding by Alfredson *et al.* (1996) for 10 lymphocytic leukemia cases for ages 20–64 years were included; while some significant information was lost, the objective selection criteria chose 20 all-leukemia all-ages cases instead.

Further selection may be made according to the additional criteria:

- Omit results that give low cumulative exposures in μT-years, typically below average 0.2 μT.
- Omit occupational studies that give no estimate of exposure.

This gives the results in Table III.

As would be expected, the effect of our selection is to reduce numbers of results admitted, and to reduce *p*-values, while increasing the percentage both of positive results and of significant-positive results.

IARC also summarizes four studies of electric fields (EF) and adult leukemia (Table 31). Leaving out the baseline (reference) exposure bands, there are 23 ORs, of which 13 are positive, with 2

Table III. Aggregation of Studies of Adult Leukemia in IARC (2002) on the Basis of Additional Selection Criteria to Identify Results with Comparable Exposures

IARC/AL (Select Results)	No. of ORs	Positives	*P*-Value for Positives	Significant-Positives	*P*-Value for Sig-Pos
Residential	4	3.5	0.06–0.31	2	0.0036
Cohort occupational	4	4	0.0625	2	0.0036
Case-control occupational	5	4	0.1875	2	0.0059
Total	13	11.5	0.002–0.01	6	3.6×10^{-7}

significant-positives and no significant-negatives. The *p*-values are 0.34 for the positives and 0.11 for the significant-positives. Selection of high-exposure results does not substantially change the picture. These studies do not give the same kind of message as the magnetic field results.

7. COMPARISON OF REVIEWS FOR ADULT BRAIN CANCER

CDHS again addresses the question of aggregation, citing 32 studies for adult brain cancer in Table 9.1.1 of their study, comprising 29 for occupational and 3 for residential exposures, and listing one representative result for each study (OR or other risk measure, with confidence limits). CDHS refers to a meta-analysis by Kheifets of the 29 occupational studies with overall OR of 1.2 (95% CI: 1.1–1.3) and to numbers of positive results and numbers with OR above 1.2. CDHS did not count numbers of significant results.

In Table 9.2.2 CDHS includes 7 additional studies to the 32 in Table 9.1.1 but gives confidence intervals for only 5 of them. One study combines residential and occupational exposure, and one is for electric fields (with a significant positive result). While CDHS discusses these additional seven studies, they are not included in its aggregation (and make little overall difference to it). Our summary for the 32 cited studies is given in Table IV.

While these aggregations are not as strong as those for adult leukemia, they are highly significant.

IARC selects 38 studies with brain cancer results for setting out in its main tables, of which 5 are residential (Table 26), 15 are occupational cohort studies (Table 29), and 18 are occupational case-control studies (Table 30). The respective numbers of results with risk measures and confidence intervals are 24, 32, and 53, that is, 109 in all, but these include repetition of subtype results in totals and include low-exposure as well as higher-exposure results from the same studies.

These include two studies (Spinelli, 1995; Ronneberg *et al.*, 1999) that are for exposures to static magnetic fields. They would have been better excluded when assessing results for ELF (principally power-frequency) fields, as the two exposures are quite different. However, the CDHS lists also included one of these studies, Spinelli (1995), with results for both brain cancer (positive) and for leukemia (negative). IARC includes these plus the negative results from Ronneberg *et al.* (1999). In treating CDHS and IARC comparably, these inappropriate results are here left in. The effect is slight, by way of diluting any overall findings.

Of the 109 crude results, 71 are positive and 16 are significant-positive. That would be highly significant under a null hypothesis for a random sample of independent results. There are also three significant-negative results, each with the upper confidence limit just on 1.0. That would not be remarkable under a null hypothesis for 109 independent results ($p = 0.51$), but could be under a stronger alternative test hypothesis with a positive association.

One significant-negative result is for a low-exposure category residential study in Table 26; that study is declared for "nervous system" cancer rather than brain cancer *per se* but it is included in Table 26,

Table IV. Aggregation of Studies of Adult Brain Cancer in CDHS (2002)

CDHS/Brain (1 per Study)	No. of ORs or Risk Measures	Positives	*P*-Value for Positives	Significant-Positives	*P*-Value for Sig-Pos
Residential	3	2	0.5	0	1.0
Occupational	29	23	0.001	6	7×10^{-5}
Total	32	25	0.001	6	0.0001

Table V. Aggregation of Studies of Adult Brain Cancer in IARC (2002) on the Basis of Selection Criteria to Identify One Representative Result per Study

IARC/Brain (1 per Study)	No. of ORs	Positives	*P*-Value for Positives	Significant Positives	*P*-Value for Sig-Pos
Residential	5	3	0.5	0	1.0
Cohort occupational	15	9.5	0.15–0.3	3	0.0057
Case-control occupational	15	10	0.15	3	0.0057
Total	35	22.5	0.04–0.09	6	0.0002

which is for brain cancer. The other two significant-negatives are in Table 29 and are for males, while they are accompanied by nonsignificant-positives for females; the selection criteria chose the more populous males, though if males and females were combined the significance would be lost. One significant-positive result was similar but the other way round, being just significant-positive for males alongside a nonsignificant-negative for the less populous females. Some significant-positives listed for Cocco *et al.* (1999) in Table 30 may look suspicious at first sight, as three are reported as having OR as 1.2 (95% CI = 1.1–1.2), which seems odd but could be accounted for by round-off from, for example, 1.17 (1.11–1.24) consistent with the usual log-normal model.

Applying our selection criteria obtains a more independent set of results, at most one per study, although as noted above it selects more populous male studies that would be partly countered by female studies. One study remained ambiguous and offered alternative opposite results of fairly equal weight (in Table 29, the Floderus *et al.* (1994) study of engine drivers or railway workers from the 1960s or 1970s); it was represented here as half positive and half negative. The result of selecting one result per study is summarized in Table V.

At this point the selection process has greatly weakened the aggregate evidence, largely because so many stronger results were in subsets. The selected evidence remains significant, if marginally so, and should not be dismissed, although it would not have the same statistical weight in assessment as that for adult leukemia. In addition, the significance of the number of significant-positive results is tempered by the existence of three marginal significant-negatives in the crude data set, two of which survived selection of one result per study.

Applying the extra set of selection criteria loses even more strength of evidence, as so many of the brain cancer studies do not have an exposure assessment in terms of field strength. The result, in Table VI, has now lost significance, bearing in mind that one significant-negative result was also selected.

8. COMPARISON WITH EVIDENCE ON CHILDHOOD LEUKEMIA

The same approach to aggregation may be applied to the CDHS set of 19 studies for childhood leukemia listed in Table 8.1.2. As with other health outcomes, CDHS applies its "sign test" and observes the 16 positive results out of 19, citing $p = 0.0004$ although our cumulative binomial calculation for 16 or more out of 19 gives $p = 0.002$. CDHS does not consider the number of significant results (3 out of 19; $p = 0.01$). There were no significant negative results and no results for OR or CLs truncated to 1.00.

IARC tables detailed results for childhood leukemia for 14 childhood leukemia studies in Tables 18, 19, and 23, comprising 10 residential exposure studies and 4 relating to use of domestic appliances. Applying the same processes as for adult leukemia, we find out of 14 results (one per study) there are 13 positive and 3 significant-positive results, with *p*-values of 0.0009 and 0.005, respectively.

On this assessment of the value of the listed sets of studies, the evidence for adult leukemia appears

Table VI. Aggregation of Studies of Adult Brain Cancer in IARC (2002) on the Basis of Selection Criteria to Identify Results with Comparable Exposures

IARC / Brain (Select Results)	No. of ORs	Positives	*P*-Value for Positives	Significant Positives	*P*-Value for Sig-Pos
Residential	5	3	0.5	0	1.0
Cohort occupational	4	3	0.3	2	0.0036
Case-control occupational	5	3	0.5	0	1.0
Total	14	9	0.2	2	0.047

more significant than that for childhood leukemia. However, we have not taken into account consistency of exposure type or measurement, or magnitude of apparent effect such as represented by ORs. CDHS refers to Wartenberg (2001), with a meta-analytic summary OR of 1.3 (1.0–1.7) for childhood leukemia. This might reasonably be compared with the meta-analytic summaries cited for adult leukemia of 1.2 (1.12–1.24) and for brain cancer of 1.2 (1.1–1.3) with reference to Kheifets *et al.* (1997a). There is not much difference between all three cancer groups at this level of meta-analysis.

The two-pooled analyses for childhood leukemia, with ORs of 1.69 (1.25–2.29) for exposures above 0.3 μT by Greenland *et al.* (2000) and 2.00 (1.27–3.13) for exposures above 0.4 μT by Ahlbom *et al.* (2000), provide stronger results by focusing on fewer more coherent and comparable studies. Although the adult leukemia and brain cancer studies may be more disparate than those entirely residential studies pooled for childhood leukemia, it would seem plausible that if they had better exposures measurements that could be used for selection, the result would also be to strengthen the overall finding.

9. CONCLUSIONS

There is a risk that review bodies, however august, may overlook the statistical weight of aggregate evidence in a collection of disparate studies that are individually inconclusive. It would be helpful in improving confidence in their reviews and assessment decisions if the issue of aggregation of disparate evidence could be seen to be addressed explicitly, preferably by a formal pooled analysis or meta-analysis to give an overall risk estimate, or if that is not available, at least by the sort of significance analysis that we have demonstrated in this article.

In aggregating evidence by the simple significance analysis we illustrate, when using the "sign test" (counting numbers of positive results), odds ratios reported as truncated at 1.0 or 1.00 etc. should be counted as half positive and not discounted. Counting numbers of significant results in this case gives stronger information than the simple sign test. If using the cumulative binomial distribution to assess the significance of numbers of positive results or of significant-positive results, it is important that the individual results are independent. We have suggested a set of selection criteria to produce at most one result per study for this purpose. However, one disadvantage of this approach is that a genuinely raised risk in a particular cancer subtype could become lost in considering all subtypes in one group, for example *all leukemia*, or *all brain cancer*.

The CDHS has addressed the aggregation of results, using the sign test and referring to external meta-analytic summaries, but it has not considered counts of significant results. The IARC review shows no evidence of having considered the aggregation of results other than subjectively. It has considered individual studies in detail and identified their shortcomings, but this has led to a tendency to fragment and dismiss evidence that is intrinsically highly significant.

Review bodies have a right to dismiss evidence on rational grounds, taking into account potential bias, confounding, and methodological limitations, as well as statistical strength, but should not do so without being seen also to take statistical aggregation into account.

The CDHS review offers a useful complementary insight into the weight of epidemiological evidence in human studies. It adds a perspective that the mainstream EMF international review bodies seem to have overlooked.

The differences in the conclusions of the IARC and CDHS reviews are not explained by differences in the sets of studies they considered. Their overlapping data sets on adult leukemia, while surprisingly different in the studies included, both represent a highly significant body of aggregated evidence. In the case of brain cancer, the crude sets of data both appear highly significant in aggregate, though our selection criteria applied to the IARC data produced only a marginally significant aggregate result.

It is debatable whether the IARC classification system should be used to distinguish between specific diseases, since it seems to be designed to classify agents. It would be reasonable for the IARC classification to refer to evidence on childhood leukemia in reaching a 2B classification. The additional evidence on adult leukemia and brain cancer might then add further support, when taken in addition to childhood leukemia.

By separating the evidence in humans for "all other cancers" (besides childhood leukemia) and summarily classifying it as "inadequate" (Section 5.5, p. 338) IARC may be seen as effectively promoting a hypothesis that EMFs may be a cause of childhood leukemia alone and of no other cancers. That is how we see policymakers interpreting it. We do not think this is rational for complex multicausal diseases, especially bearing in mind evidence for possible systemic effects that could affect causation of several

diseases. IARC does not seem to have addressed the question of compound hypotheses.

This exclusive attribution of the IARC 2B classification to childhood leukemia has repercussions in precautionary policy, as manifest in the draft WHO Precautionary Framework (2006). Owing to its rarity, childhood leukemia has relatively little impact on society and its avoidance therefore has relatively little benefit, compared with the substantially more prevalent adult leukemia and brain cancer, as well as the other outcomes rated as 2B by CDHS.

An earlier review by the NIEHS (1999) had associated both adult and childhood leukemia with a 2B classification, and both the IARC and CDHS reviews were informed by this. Given the extent and aggregate strength of the evidence for adult leukemia, both in itself and in comparison with that for childhood leukemia, it is difficult to see a clear division that would support an exclusive hypothesis of carcinogenicity of EMFs for childhood leukemia but not for adult leukemia.

Postscript: Since our first writing of this article, the study by Lowenthal *et al.* (2007) has appeared, as has our commentary in the same journal (O'Carroll and Henshaw, 2007). These results reinforce our conclusions in respect of adult leukemia, though our argument is principally about methodology.

10. RECOMMENDATIONS

The following recommendations are made for future reviews of EMF health effects.

(i) IARC and other review bodies should incorporate expressly into their methodology some assessment of aggregate value of disparate evidence. Such assessment should not itself determine the overall assessment decision, but it is better to be aware of the nature of the aggregated data.

(ii) A focused pooled analysis should be undertaken for adult leukemia to parallel, as far as possible, those of Ahlbom *et al.* and Greenland *et al.* for childhood leukemia.

(iii) Advisory bodies considering precautionary policy relating to EMFs should take into account both the IARC and CDHS reviews, including the failure of IARC to demonstrate any assessment of aggregate value of evidence.

(iv) The WHO EMF team, in forming its Precautionary Framework, should expressly address the impact of possible health outcomes other than childhood leukemia, noting especially their relatively high incidence compared with childhood leukemia, and giving particular attention to the five outcomes classified by CDHS as corresponding to IARC Class 2B.

ACKNOWLEDGMENTS

We thank Children with Leukemia (UK Registered Charity No. 298405) for partial support for this work. The authors acted as expert witnesses at the Beauly-Denny Powerline enquiry in Scotland in March 2007 in support of those objecting to the proposed powerline passing close to houses. The funding source waived any rights to review or approve the article.

REFERENCES

Note: Full references are not reproduced here for every study mentioned in the reviews, since references appear in the reviews themselves and their mention in this article is concerned only with how they are counted in the reviews.

Ahlbom, A., Day, N., Feychting, M., Roman, E., Skinner, J., Dockerty, J., McBride, M., Michaelis, J., Olsen, J. H., Tynes, T., & Verkasalo, P. K. (2000). A pooled analysis of magnetic fields and childhood leukaemia. *British Journal of Cancer*, *83*(5), 692–698.

CDHS. (2002). An evaluation of the possible risks from electric and magnetic fields (EMFs) from power lines, internal wiring, electrical occupations, and appliances, California EMF Program, Final Report June 2002.

Greenland, S., Sheppard, A. R., Kaune, W. T., Poole, C., & Kelsh, M. A. (2000). A pooled analysis of magnetic fields, wire codes and childhood leukaemia. *Epidemiology*, *11*, 624–634.

IARC. (2002). IARC monographs on the evaluation of carcinogenic risks to humans, vol. 80, Non-ionising radiation, part 1: Static and extremely low frequency (ELF) electric and magnetic fields, IARC Press, 2002.

Lowenthal, R. M., Tuck, D. M., & Bray, I. C. (2007). Residential exposure to electric power transmission lines and risk of lymphoproliferative and myeloproliferative disorders: A case-control study. *Internal Medicine Journal*, *37*, 614–619.

National Institute of Environmental Health Sciences (NIEHS). (1999). NIEHS report on health effects from exposure to power-line frequency electric and magnetic fields. NIH Publication No. 99-4493, P. O. Box 12233, Research Triangle Park, NC..

O'Carroll, M. J., & Henshaw, D. L. (2007). Letter to the editor: Adult leukaemia near powerlines. *Internal Medicine Journal*, *37*, 841.

Rothman & Greenland. (1998). *Modern Epidemiology*, 2nd ed. Philadelphia, PA: Lippincott, Williams & Wilkins.

WHO Framework Guiding Public Health Policy Options in Areas of Scientific Uncertainty with Particular Reference to EMFs. Draft for Consultation, May 2006. The International EMF Project Radiation and Environmental Health Unit World Health Organization.