Terminology for Patterns of Evidence

In describing a body of evidence we want to avoid using adjectives that presuppose policy directions. We recommend the following terminology:

- To describe relationships between exposures to EMFs and adverse outcomes in general, we will use the following terms: "increase in occurrence," "no change in occurrence," or "decrease in occurrence." The term "occurrence" can refer to any measured outcome.

  We will look at a body of evidence including individual studies which did not reach conventional statistical significance, since a barely detectable association based on the size and quality of the study may only become apparent in a meta-analysis or less formal equivalent review. We will provide confidence limits for individual studies or calculated p values when these are available. There is controversy about the dependence on statistical tests to evaluate or screen studies. We will look at the evidence both ways and comment on whether this alters the conclusions. Where we describe tests of significance we will prefer two-tailed 95% confidence limits, or when only p values are available we will specify if they are one or two tails, with preference for two-tailed tests.

- To describe outcomes that are observed always or almost always in repeated experiments or studies, we will use the word "consistent."

  We will characterize as "recurrent" those outcomes that, while not always seen, are observed repeatedly in studies and have no clear alternative explanation.

It is not uncommon for official agencies in their summary statements after a risk assessment to characterize the strength of an association, not as a number with confidence limits, but as "strong" or "weak." This is probably done with the desire to express the idea in everyday English and "put it in perspective." Underlying this temptation to talk about an effect as "strong" or "weak" is the idea that if few cases of disease would be eliminated by removing this factor from the environment, it is probably not cost beneficial to remove it. Most of the time this may be true, but if there is a very inexpensive way to remove exposure, then it could be cost beneficial to prevent these few cases. We will leave policy analysis to the policy analysts and use terms which are policy neutral. The terms "strong" and "weak" have several quite different interpretations:

> Worthy or unworthy of societal or policy concern. (The absolute lifetime theoretical added lifetime risk is less than 1 per million, the *de minimis* benchmark in some regulations.) "*De minimis*" is a Latin legal phrase used to denote risks considered to be negligible because they are small.
>
> Easily or barely detectable given the size and quality of the scientific studies used. For example, epidemiologists begin to worry about unknown bias, or confounding when relative risks are less than 2.0. They tend to talk about such associations as "small" or "weak," though they can denote a 100% increase in occurrence. Toxicologists begin to worry when the difference of occurrence is not larger than the historical fluctuation of the occurrence in control groups. When an effect fails this informal evidentiary test they say the effect is not "robust." It is difficult to define what is detectable quantitatively, but this gets at the idea of the relative size of signal and noise.
>
> Large or small compared to some other association. (The association or the added risk is small compared to some other disease like AIDS, which people fear.)

We prefer to use language that does not confuse these different meanings. So in public summary statements we will not use words like " strong," "robust," or weak. Instead we will use phrases like:

> "The magnitude of theoretical attributable lifetime risk (for cancer risk) is (larger/smaller) than the one per 100,000 level that triggers notice under Proposition 65."

> "The difference of occurrence between exposed and unexposed individuals was (easily, barely, not reliably) detectable given the size and quality of the studies available."

> "The added risk or proportion of total cases of disease x attributable to EMFs is (larger, same, smaller) than the added risk or proportion of total cases of disease **x** attributable to agent **y**."

It should be noted that even barely detectable effects, based on the size and quality of epidemiological studies, would tend to be larger than those that would trigger notice under Proposition 65 in California.

The word "robust" is used as a term of art to describe experimental studies without careful defining it. Since "robust" can also have multiple interpretations we will avoid its use and instead say:

The size of the effect was easily detectable given the size and quality of the study, was seen consistently in repeated experiments and was larger than the variation between the various control groups."

We wish to avoid the ambiguity of such statements as "there is no evidence that x causes y" which could mean that there are no studies on this topic, or that there are plenty of studies but all of them fail to show that x causes y. We will therefore talk about the "evidentiary base" to describe the volume of evidence and will characterize it as absent, scant, moderate in size or voluminous. We will talk about the "pattern of evidence" to denote the results in that evidentiary base. So we might say "There is no evidentiary base to address the question whether x causes y" or "There is a voluminous evidentiary base on whether x causes y and consistently the pattern of evidence suggests that x does not cause y".

Dealing with and Describing Study Quality

With regard to quality, we intend to review studies that have been published, or accepted for publication. For studies the California EMF program has sponsored, we will include those studies that have passed the external peer review which we have arranged, even if the study has not yet been submitted for publication.

Epidemiologists tend to think about quality issues differently than experimentalists. Since epidemiologists rarely perform experiments (randomized trials are the exceptions) they rarely can eliminate bias and confounding and measurement error to the degree which is possible in an experiment. The experimentalist tries to control everything and will often discard a study entirely if there was a failure to control any of the desired parameters. The experimentalist tends therefore to think in terms of "good quality studies" and "bad quality studies" and simply ignores the latter category. The epidemiologist does not have this luxury and tends to evaluate the direction of the biases induced by the inevitable lack of perfection in his or her study designs. Although we will acknowledge standard experimental practice and whether an experimental study was carried out under standard regulatory "Good Laboratory Practices" when discussing experimental studies, we will also tend to discuss the expected direction of bias and confounding in both experimental and epidemiological studies.

The terminology for describing the quality of epidemiological evidence will go beyond "good quality" or "bad quality." We will discuss the potential for confounding, bias, measurement error and the direction of the bias they are expected to produce.

The structured questions in Section Two assure that these issues are explicitly dealt with.

How Degree of Confidence and Magnitude of Risk (if Real) Could Be Explained to the Public

This way of talking about the evidence can be illustrated by applying it to the evidence related to the carcinogenicity of benzene, arsenic and the carcinogenicity of ferric oxide.

Benzene: The US EPA and CalEPA have classified benzene as a known human carcinogen on the basis of a voluminous evidentiary base in animals of acceptable quality and a number of occupational studies of acceptable quality in humans that show an easily detectable increase of cancer occurrence given the strength and weaknesses the studies. Scientists at DHS thinks it is somewhere between more than 50% certain but less than virtually certain that benzene in typical urban air could increase the rate of leukemia in the population to some degree. However, the upper bound of theoretical increase in occurrence would be well below the power of the best epidemiological studies of the general population to detect that effect. The upper bound of theoretical risks from a lifetime of exposure would be on the order of 10 per 100,000 and is of regulatory concern since California regulates at the 1 per 100,000 level of theoretical lifetime risk. Individuals want information on individual risk. The chance of escaping leukemia after a lifetime of breathing benzene in urban air would be 99,990 per 100,000, so the individual risk is small. Some people want to know what proportion of the total burden of disease in the population is attributable to a factor like benzene in typical urban air. The total lifetime risk of leukemia from all causes is about 700 per 100,000. Thus benzene in air would not account for much of the total leukemia rate in the population.

Arsenic: The US EPA and Cal EPA have classified arsenic as a human carcinogen based on a voluminous evidentiary base of human occupational and drinking water epidemiology which includes good quality studies showing effects easily detectable given the size and quality of the studies and despite an adequate evidentiary base in animals which until recently failed to experimentally demonstrate cancer in animals. DHS scientists believe that it is highly probable to virtually certain that arsenic in occupational settings and in drinking water can produce some cancer. Epidemiological evidence suggests that in some parts of California with high arsenic content in water the lifetime theoretical risk could reach 1,000 per 100,000, far above the one per 100,000 regulatory level. Even in these areas an individual would have a 99% chance of escaping cancer caused by arsenic. We do not have sufficient exposure information about the general public to estimate the excess of cancer caused by arsenic.

Ferric Oxide: Based on an adequately voluminous evidentiary base in animal studies which have not shown an increased occurrence of tumors in animals and a number of occupational studies in humans which have not shown an increased cancer rate, when other known carcinogens were absent from the work place, IARC has classified this agent as “not classifiable as to human carcinogenicity and with animal evidence suggesting lack of animal carcinogenicity. DHS scientists would estimate that ferric oxide is very unlikely to extremely unlikely to cause cancer in occupational or environmental settings.

DEGREE OF CERTAINTY APPROACH TO EVALUATION

The DHS scientists found the usual process of describing the pattern of evidence in some detail and then expressing an opinion (without explaining the rationale for that opinion) to be insufficiently transparent. Accordingly, they supplement the usual International Agency for Research into Cancer (IARC) procedure with an additional form of presentation and an additional form of judging whether EMFs are a cause of disease. The following table shows the questions that were systematically addressed. For definitions of epidemiological terms in the table see pages 19 and 20.

TABLE 1.1 QUESTIONS RELEVANT TO DEVELOPING A DEGREE OF CERTAINTY ABOUT CAUSALITY

EXPLANATIONS OF A STATISTICAL ASSOCIATION OTHER THAN A CAUSAL ONE
Chance: How likely is it that the combined association from all the studies of EMF and disease is due to chance alone?
*Bias: How convinced are the reviewers that EMFs rather than a study flaw that can be **specified and demonstrated** caused this evidentiary pattern? If no specified and demonstrated bias explains it, how convinced are they that EMFs caused these associations rather than **unspecified** flaws?*
*Confounding: How convinced are the reviewers that these disease associations are due to EMFs rather than to another **specified and demonstrated** risk factor associated with EMF exposure? If not due to a specified risk factor, how convinced are they that they are due to EMFs rather than to **unspecified** risk factors?*
*Combined effect: How convinced are the reviewers that these disease associations are due to EMFs rather than to a combined effect of chance and specified or **unspecified** sources of bias and confounders?*
ATTRIBUTES SIMILAR TO HILL'S (HILL, 1965) THAT ARE SOMETIMES USED BY EPIDEMIOLOGISTS TO EVALUATE THE CREDIBILITY OF A HYPOTHESIS WHEN NO DIRECT EVIDENCE OF CONFOUNDING OR BIAS EXISTS
Strength of association: How likely is it that the meta-analytic association is strong enough to be causal rather than due to unspecified minor study flaws or confounders?
Consistency: Do most of the studies suggest some added risk from EMFs? How likely is it that the proportion of studies with risk ratios above or below a RR of 1.0 arose from chance alone?
Homogeneity: If a large proportion of the studies have risk ratios that are either above or below a RR of 1.0, is their magnitude similar (homogeneous) or is the size of the observed effect quite variable (heterogeneous)?
Dose response: How clear is it that disease risk Increases steadily with dose? What would be expected under causality? Under chance, bias, or confounding?
Coherence/Visibility: How coherent is the story told by the pattern of associations within studies? If a surrogate measure shows an association, does a better measurement strengthen that association? Is the association stronger in groups where it is predicted? What would be expected under causality? Under chance, bias, or confounding? How convinced are the reviewers that the magnitude of epidemiological results is consistent with temporal or geographic trends?
Experimental evidence: How convincing are the experimental pathology studies supporting the epidemiological evidence? What would be expected under causality, bias, chance, or confounding?
Plausibility: How convincing is the mechanistic research on plausible biological mechanisms leading from exposure to this disease? What would be expected under causality, chance, bias, or confounding? How influential are other experimental studies (both in vivo and in vitro) that speak to the ability of EMF to produce effects at low dose?
Analogy: How good an analogy can the reviewers find with similar agents that have been shown to lead to similar diseases? What would be expected under causality, chance, bias, or confounding?
Temporality: How convinced are the reviewers that EMF exposure precedes onset of disease and that disease status did not lead to a change in exposure?
Specificity and other disease associations: How predominantly are EMFs associated with one disease or subtypes of several diseases? What would the reviewers expect under causality, chance, bias, or confounding? How much is their confidence in EMF causality for disease X influenced by their confidence that EMFs cause disease Y?

We recognizes that a reassuring pattern of evidence from a stream of evidence that often misses a harmful effect does not allay one's suspicion much, even though an alarming pattern of evidence from that same stream of evidence might increase suspicion a lot. For example: if birds sometimes survive eating fruits that are lethal to humans, then reassuring evidence from bird experiments would not allay suspicion as much as the death of the birds after eating the fruit would increase our suspicion. In the terminology of probability, the relative likelihood conveyed by a positive or negative result depends on the false positive rate and false negative rate characteristic of that stream of evidence. The mathematical basis for this insight is discussed in the Risk Evaluation Guidelines (www.dhs.ca.gov/ehib/emf). It resulted in realizing that any stream of evidence, judged by the extent to which it usually produced false positive and/or false negative results, could be classified into four possible types: 1) capable of strengthening OR weakening one's certainty, 2) predominantly capable of strengthening certainty (like the bird feeding example given above), 3) predominantly capable of weakening certainty and, 4) uninformative, neither capable of strengthening nor weakening one's confidence. . It should be noted that the Hill's attributes are like the bird feeding example. If they are present they strengthen confidence, but if they are absent, confidence falls only a little.

Expressing a Degree of Certainty that An Association is Causal

the California Guidelines specified that in order to accommodate the probability-based computer models of the program's policy projects each of the DHS reviewers would individually assign a number between 0 and 100 to denote their degree of certainty that epidemiological associations between EMFs and certain diseases were causal in nature. The guidelines, which were modified with advice from public comment and the SAP and the DHS reviewers, attached pre-agreed-upon English language phrases to various ranges of this degree of certainty. These are presented below in Table I.

If all three judges had best judgments above 50 out of 100, but that fell in different categories in Table I, judges were said to be "inclined to believe" that EMFs increased the risk of that disease to some degree.

If they found themselves in different categories below that point, they were said to be "inclined not to believe that EMFs increased the risk of that disease to any degree."

TABLE 1.4 EVERY DAY ENGLISH PHRASES TO DESCRIBE DEGREES OF CERTAINTY OF CAUSALITY (GRAPH ILLUSTRATES THE RANGE OF CERTAINTY NUMBERS TO WHICH THE PHRASES PERTAIN)

ARE THE HIGHEST EMFS AT HOME OR AT WORK SAFE, OR DO HIGH EMFS INCREASE THE RISK OF ..... TO A DEGREE DETECTABLE BY EPIDEMIOLOGY?	DEGREE OF CERTAINTY ON A SCALE OF 1 TO 100
Virtually certain that they increase the risk to some degree	>99.5
Strongly believe that they increase the risk to some degree	90 to 99.5
Prone to believe that they increase the risk to some degree	60 to 90
Close to the dividing line between believing or not believing that EMFs increase the risk to some degree	40 to 60
Prone to believe that they do not increase the risk to any degree	10 to 40
Strongly believe that they do not increase the risk to any degree	0.5 to 10
Virtually certain that they do not increase the risk to any degree	< 0.5