The so-called ‘criteria’ of causation are arguably the most familiar component of the epidemiologist’s approach to causal inference. These criteria and other aspects of causal inference have been described in textbooks and methodological papers for nearly 50 years, emerging from discussions in the late 1950s of the inadequacy of the Henle-Koch postulates for causal inference in chronic disease epidemiology.1 In two recent papers, very different recommendations regarding the use of these criteria appeared. Wynder2 argued that epidemiologists should use criteria (more than we do) to determine whether a relation is causative. These criteria are used for causal assessments as well as for making public health recommendations. In theory, causal criteria can be used to either refute or predict causal effects. In this context, wish bias is the tendency for investigators to make causal conclusions (or not) primarily on the basis of their own published results of the association in question or other inherent tendencies to ‘wish for’ a desired result.5

Although these two papers were written from vastly different perspectives and represent only two voices out of many, the sharp discrepancy regarding a central methodological problem is noteworthy. While controversy may be important for methodological progress, it may also confuse epidemiologists searching for guidance in the practice of causal inference. Simply put, both views cannot be correct. Interestingly, the contrast between these two views is evaluable in practical terms. During the same brief time period in which these two papers appeared, a third paper, describing the practice of causal inference, was published.4

The purpose of this commentary is to examine how well the two discrepant views square with current practice. Some theoretical issues central to Charlton’s arguments will also be discussed. The paper concludes with some recommendations for practice and for research in this important methodological area.

BACKGROUND—TWO DISCREPANT VIEWS

Wynder argues that criteria should be used more often than they are for the purpose of assessing causation. (Nowhere does he mention their use in making public health recommendations.) He advocates the more frequent use of the criteria of consistency, dose-response, temporality and biological plausibility when epidemiologists are faced with weak associations. The use of these criteria, in Wynder’s view, will reduce ‘wish bias’ and thereby increase the validity of causal inferences. In this context, wish bias is the tendency for investigators to make causal conclusions (or not) primarily on the basis of their own published results of the association in question or other inherent tendencies to ‘wish for’ a desired result.5

In contrast, Charlton argues that criteria should not be used for the purpose of assessing causation because their use diminishes the validity of causal inferences. He believes that Hill’s criteria are only useful for making public health recommendations (what he calls policy decisions) and that their use makes epidemiology less a
science and more a multidisciplinary technik producing unreliable and ungeneralizable results. He argues that Hill’s criteria are incapable of testing (by which he means refuting) hypotheses of necessary cause. Because Charlton sees such tests as the essence of establishing causality in science, the use of Hill’s criteria in his view is unscientific.

The Current Practice of Causal Inference

One way to evaluate these vastly different views on the use of causal criteria is to examine the extent to which they are reflected in the current practice of causal inference. In a recent review,4 two series of review papers were examined for the use of causal criteria in determining causation and in making public health recommendations. The associations examined in these series were alcohol and breast cancer and vasectomy and prostate cancer, both weak associations.

To what extent do epidemiologists use the criteria of strength of association, consistency, temporality, dose-response, and biological plausibility in order to evaluate evidence regarding weak associations, a practice Wynder claims is underutilized?

In the 14 alcohol and breast cancer reviews, the most commonly used criteria were: consistency (in all), strength of association in 13 reviews, and dose-response and biological plausibility in nine and eight reviews respectively. A similar finding was reported from the vasectomy and prostate cancer reviews. Thus, Wynder’s concerns about four of Hill’s criteria may be misplaced. On the other hand, the criterion of temporality was infrequently used in these reviews (four out of 14), consistent with Wynder’s argument that epidemiologists should use this criterion more than they do.

Wynder also makes suggestions regarding where epidemiologists should assess causation: specifically, in discussion sections of reports of single studies. The ‘review of reviews’4 did not examine the use of causal criteria in other types of publications. It seems reasonable, nevertheless, to argue that ‘wish bias’ could best be reduced if causal inference were reserved for review papers written by investigators who had not themselves published a study on the same topic. A similar strategy is used by the National Institutes of Health (NIH) for choosing panel members for NIH Consensus Conferences wherein causal inferences and public health recommendations often emerge. General experts are chosen, excluding those whose research could potentially bias their viewpoint.6 Interestingly, the ‘review of reviews’ did not tend to make causal claims.

To what extent do epidemiologists use Hill’s criteria for concluding causation (a practice Charlton denies as valid)?

The ‘review of reviews’ also examined the extent to which Hill’s causal criteria were used to make causal conclusions as distinct from public health recommendations.4 While all 14 alcohol and breast cancer reviews used all or subsets of Hill’s criteria, six of these reviews did not address public health or policy recommendations. In other words, nearly half the reviewers used criteria to assess causation alone.

Some Theoretical Issues

Given that epidemiologists practice what Charlton’s arguments suggest is invalid on theoretical grounds, a look at his theoretical arguments will be important. They rest on two types of claims: one regarding the nature of causation, the other—more methodological—regarding how causes are tested.7 Charlton’s claim is that causation in epidemiological science is defined in terms of necessary causes. In this view, sufficient causes are (virtually) never seen and components of sufficient causes, as described in a classic paper and more recent textbook,8,9 are apparently unimportant. Whether these related claims regarding the relative frequency of different types of causes are reasonable or not, the central methodological claim is that Hill’s criteria cannot test (meaning refute) hypotheses of causal necessity. If true, then the claim that Hill’s criteria cannot be used for scientific purposes may be correct, as long as we accept the companion claim that science can be demarcated from non-science or pseudo-science on the basis of refutability, a central concept in the philosophy of critical rationalism10 sometimes called ‘Popper’s philosophy’.

It will not be necessary to enter the domain of philosophical criticism—expanding on recent discussions regarding weaknesses in a Popperian perspective—to show that Charlton’s assertion regarding the use of causal criteria is severely flawed. In direct contrast, at least one of Hill’s criteria, temporality, can refute a hypothesis of causal necessity at either the individual or population level. If exposure to a potential causal factor occurs after rather than before disease onset (which itself occurs prior to the diagnosis of the disease or condition) then that factor cannot be considered necessary for the occurrence of that disease. Hence, causal necessity is refutable using the criterion of temporality. Interestingly, although this criterion appears to be very
important for testing causal hypotheses by refutation, it is not often used in the practice of causal inference perhaps because it is considered a moot point if cohort evidence is available.4

Even if an exposure factor precedes disease occurrence, temporal order is nevertheless insufficient for inferring cause.12 This idea, reinforced every time the fallacy of post hoc, ergo propter hoc is invoked, anticipates the need for other causal considerations (i.e. criteria).13 In more theoretical terms, Hill’s criteria must be interpreted not only in terms of refutability, but also in terms of predictability. Two criteria, strength of association and consistency, are important predictors of causal effects.7 Large and consistent relative risks are predicted from studies involving a necessary cause, if that cause has been accurately measured (given that the effects of confounding and other biases are ignored). As discussed below, better measurements, in turn, are most likely to arise from biological and molecular (i.e. basic) sciences.

Charlton’s broad criticism of Hill’s criteria as unscientific suggests that the criterion of biological plausibility may not play an important role in the practice of causal inference. Yet it is precisely the consideration of plausibility that bridges the gap between human epidemiological evidence and the many forms of biological evidence (including animal experiments) that Charlton appears to so highly value. Plausibility may become increasingly important in causal assessments as molecular epidemiology permits more precise measurements of intracellular causal effects. For an example of the impact of increasingly precise measurement technology on inferential capacity, consider the viral aetiology of cervical cancer.14 Early epidemiological studies examined sexual habits as a risk factor. An evolving research programme revealed that papillomavirus infection could explain the increased risk due to sexual activity. Eventually, the rapidly expanding capability to measure HPV subtypes 16 and 18 and a worldwide effort to collect tumour specimens, led to the conclusion that HPV infection may be a necessary cause of cervical cancer.15

The criterion of experimentation also falls under Charlton’s critical eye. Yet prevention trials are generally considered to be strong (and scientific) tests of a causal hypothesis, examining the impact of removing or reducing exposure to a potential cause under controlled conditions. The on-going Women’s Health Initiative testing the effects of a low-fat diet on breast cancer incidence is a good example. In his paper, Charlton refers to the idea of experimentation solely in terms of the morally bankrupt trial wherein a potential cause is randomly allocated to a healthy population.3

SUMMARY
It seems that two recent recommendations regarding the use of causal criteria have missed the mark in practice and in theory. Contrary to Wynder’s assertions, epidemiologists use many causal criteria in their assessments of weak associations, primarily in reviews and editorials, where ‘wish bias’ may be less likely to occur. Contrary to Charlton’s assertions, the existing causal criteria can be used to critically test—through refutation and prediction—causal hypotheses of necessary cause. Other types, such as sufficient causes and components of sufficient causes can also be tested.7

While we may disagree with their solutions, Wynder and Charlton have brought important problems to the table. Weak associations are an inferential challenge. Continued success in molecular epidemiology may herald the end of an era for the oft-cited adage that ‘only weak associations remain to be discovered’.2 However, if the future reveals even a few remaining weak associations, e.g. in diet and nutrition, then epidemiologists will continue to face the problem of deciding which additional criteria (beyond strength of association) should be used. In practice, epidemiologists tend to use consistency, dose-response, and plausibility along with considerations of confounding and bias although they also exhibit considerable methodologic variability.16

Is this the ‘best’ way to practice causal inference in epidemiology?

Looking for answers to such methodological questions in theory is another important challenge. A theoretical approach could be useful for explaining and for informing the current practice of causal inference. Interestingly, Charlton’s view appears to be centred on the idea that epidemiology is a science and only a science, narrowly defined not only in causal terms but also in methodological terms relying exclusively upon refutation. Unfortunately, this narrow view of the discipline exempts epidemiologists from public health endeavours, in direct conflict with the way in which most epidemiologists practise causal inference,4 recent definitions of the discipline,17 and perhaps most importantly, the development and application of ethics guidelines and ethical reasoning in epidemiology.18 The role of ethics in making public health recommendations has recently been examined.19 Professional epidemiologists who profess (i.e. promise) to prevent disease have a moral obligation to consider public health recommendations when examining evidence for causality.20 Central to this view is the idea that public health recommendations are made in light of judgments about causation, thus tying together science and its application within the professional discipline of epidemiology.
It follows that theoretically-inclined epidemiologists will need to find a balanced perspective somewhere between the best available scientific theories and the most appropriate ethical theories if we endeavour to develop a theory general enough to cover causality, scientific methodology, and the obligatory public health role epidemiologists undertake when reviewing evidence. Metaphorically, we must chart a course on the ‘sea of person-time’ where so many theoretical approaches exist. A comprehensive approach to a theory of causal inference may need to consider the many different notions of cause, the views of relativists and realists, other views in contemporary philosophy of science as well as the profound pluralism found in postmodern bioethics, where utilitarian theory, Kantian theory, feminist bioethics, and virtue theory vie for attention with specified principlism and casuistry. As Labarthe and Stallones suggest, selecting among these is a challenging problem in the philosophy of epidemiology.

RECOMMENDATIONS FOR THE FUTURE
In the end, what should epidemiologists do? Should we follow the current methodological paradigm—a somewhat subjectively applied criteria-based causal inference method—or should we challenge it with new approaches? I believe we should do both. This approach, more evolutionary than revolutionary, is consistent with the way in which epidemiologists have traditionally used practical problems to drive theoretical and methodological research and, by an indirect and socially-influenced feedback loop, used theory and methodology to drive practice. In sum, I recommend that we practise causal inference as it is practised today, using as a guide the general patterns found in careful descriptions. Note, however, that current practice leaves considerable room for personal preference regarding the choice of criteria and other issues, a situation that can produce very different judgments about causation from the same evidence. A recent controversy about induced abortion and breast cancer provides a compelling example. In light of these concerns, I recommend that we prepare ourselves for improvements that could emerge from research on causal inference methodology using the following rational problem-solving scheme: Identify the strengths and the many problems in the practice of causal inference. Then, identify theories, concepts, and approaches capable of explaining the strengths and solving the problems. Propose tentative solutions from those sources and translate the findings into practice. Criticize these solutions and their sources. Hope for progress.

Some Practical Problems
The general problem-solving scheme described above will only succeed if specific problems are identified. What follows are a few practical problems to get things started:

(i) In practice, definitions and accompanying rules of evidence for the same criterion differ across users of causal inference methods. For example, there are at least three different definitions for consistency, each used by a different reviewer of the alcohol and breast cancer literature. The practice of causal inference would improve if a broader consensus could be achieved on definitions and rules.

(ii) To what extent does meta-analysis, which provides a summary effect measure over several similar studies, provide a quantitative approach to assessing consistency?

(iii) The interconnectedness of criteria should be more carefully examined. Consider a simple hypothetical example involving strength of association, biological plausibility, and consistency. Suppose a ‘weak’ association is consistent across nearly all human epidemiological studies. Suppose that advances in biological measurements provide increasingly precise measures of a factor that had been considered a weak confounder. When substituted into the analysis, the factor reduces the measured effect to the null for a majority of studies. As a result, the association is no longer consistent. Given that most associations involve unknown confounders, to what extent should causal conclusions or public health recommendations be delayed until all (or nearly all) potential confounders are discovered or better measured? What are the benefits and risks of such a strategy?

CONCLUSIONS
A misunderstanding of how and why epidemiologists use causal criteria provides an opportunity to identify methodological problems directly related to the fundamental goals of epidemiology: studying the determinants of disease in populations and applying the knowledge gained for the benefit of public health. These problems include the interrelationships of causal criteria, the basic definitions of criteria and their inferential properties. Solutions to these problems may arise from theoretical investigations and from further empirical studies describing the current practice of causal inference.

ACKNOWLEDGEMENTS
Suggestions on an earlier version of this manuscript from Drs Karen Goodman, Marshall Joffe, and Darwin Labarthe are greatly appreciated.
REFERENCES


(Revised version received January 1997)