Needs of policymakers and responsibilities of nutritionists

Maintaining and improving the health of populations are major concerns of governments, and a primary factor in health is what people consume. Policymakers increasingly must set standards, formulate regulations, and write statutes in this area (1,2). It is crucial that nutrition researchers provide the information and tools necessary to define healthy intakes and predict the consequences of over- and underconsumption. The fundamental problem is that of estimating the prevalence of adverse health effects (AHEs)3 for a population that consumes

(either currently or potentially) specific diets or dietary components (3,4). Although traditionally requirements and toxicity levels were estimated in different ways, with standard statistical analyses used for minimum levels (requirements) (5,6) and risk-analysis models for upper levels (toxicities) (7–9), this article shows how new work on requirements has permitted the formulation of a general model for safe ranges that includes both minimums and maximums.

Prevalence of any factor or condition is defined as the fraction of a specific population that has that condition. For each specific AHE, each dietary component, and each population, the estimation of prevalence depends directly on two factors: how much of the component the population does or might consume (food intake), and what this consumption would do to them (physiological response). Estimating these factors is difficult; the research nutritionist is primarily involved with measuring the effects of varying intake levels of specific nutrients and components, and the nutritional epidemiologist works on determining what people really do eat.

With respect to estimation of physiological response, different approaches exist for setting lower and upper limits of the range of healthy intakes. Requirements, which are minimal levels that are consistent with health, were primarily based on data from feeding experiments on individuals and were described by average requirement levels and variability between individuals (10,11). Typically, toxic levels were treated differently: the upper levels were primarily set to be consistent with health based on either animal experiments or, occasionally,

ABSTRACT Nutrition researchers are called upon to inform policymakers of the consequences of consuming different amounts of various nutrients: how much is enough, how much is too much, and why. Traditionally, requirements were described by population-average levels and some measure of between-individual variation, whereas toxicities were described by single levels above which a toxic response was likely to occur. These statistics were used to model the prevalence of various adverse health effects in individual populations. Recently, accumulated data and improved understanding have led to more complete descriptions of the utilization of a few important nutrients and to the formulation of a model that includes both requirement and toxicity as extremes of the healthy-intake spectrum. This approach is based on estimating the full statistical distributions of individual nutritional deficits and surfeits within a population and using this to derive prevalence curves for specific adverse health effects. This information promises to provide policy planners with better tools to judge the health of populations, design feeding programs, and predict the sequelae of modifications in food supply and dietary habits. J. Nutr. 134: 1610S–1616S, 2004.

KEY WORDS: • nutritional planning • assessment • nutrient requirement • adverse health event • prevalence • curve • deficit • protein • risk analysis
epidemiological population studies, and the final estimates of safe upper levels were provided by risk-analysis methods (7,12).

Recently accumulating data and increased interest have led to the development of more complete statistical descriptions of the full response curves for a few important nutrients: protein requirement (6) and energy requirement and toxicity (where obesity is viewed as a toxic effect of consuming excess energy) (5). These estimates permit us to refine the concept of prevalence and provide policy planners with additional information and tools for estimating the societal cost of food alternatives. Moreover, although so far most of this attention focused on the lower limits of intake (requirements) consistent with health, existing statistical and mathematical methodologies readily transfer to estimation of the upper limits.

This article summarizes present efforts to build a unified model of requirements and toxicities and the limitations and benefits of this approach. A major restriction on the progress in this area is the lack of data that are available; however, developments in modern biology (metabolomics, etc.) give promise that we may soon face a surfeit rather than a deficit of data (13,14). As the amount of data increases, there is increasing need for more comprehensive models in part to give direction to the application of bioinformatics to the problems of nutrition (15).

**Traditional approach to estimating prevalence**

To determine the prevalence of a specific AHE in a specific population and as a function of some dietary component, it is necessary to have information about the following: 1) the physiological response curve relating the intake of that component to the risk of the AHE (the risk that individuals who consume at or below that level will have that AHE); 2) the intake distribution (i) for the population of interest (the fraction of the population that is or might be consuming specific amounts of the component); and 3) the relationship between these two functions, although this last contribution is usually ignored. Traditionally, intake and response were assumed to be independent and were estimated independently, and the prevalence was estimated as the sum (more precisely the integral) for each possible level of intake for the fraction of the population that was consuming at that level multiplied by the risk that consumption at that level would lead to some AHE. This is formally described as the weighted average of the population intake (I) with weighting by the risk of the AHE (3,16):

\[
\text{Prevalence} = \Sigma \text{risk}(I) \times \text{consumption}(I)
\]

\[
\int \text{risk}(I) \times \text{consumption}(I)
\]

In practice, evaluating this integral with any precision is difficult. There are difficulties in both defining response functions and measuring population intake, and moreover, the assumption that intake and requirement are independent is not true for requirement (both are usually related to body size, and most nutrient requirements are expressed per unit of body weight in an attempt to minimize this effect) and often not true for toxicity, where tolerances may be developed. As a step toward improving this model, we need to consider how the formalization captures the inherent and pervasive fact of biological variability—the conceptual problems of dealing with random variables.

**Variability in biology**

Variability is at the core of biology, and many statistical procedures were developed to deal with this (17). The lack of precise reproducibility of biological data is formalized by stating that most biological measurements or observations are fundamentally random variables, and information about them is best captured by probability distributions (mathematical functions that contain the essential information about their expected values, variabilities, and interrelationships) that describe how individual variables are distributed (such as normal or lognormal) and how they are related (such as by correlation and bivariate distributions).

For example, the statement that population intake has a distribution, written as a mathematical function i(x), formally says that the fraction of the population consuming between any two levels a and b is the area under i(x) or the integral of i(x) from a to b. We need to consider distributions and cumulative distributions, where I(X) = \( \int_0^X i(x) dx \) to intake level X, which represents the fraction of the population consuming below the level X. The importance of these concepts is highlighted by the fact that the cumulative distribution is just the dose-response curve for a nutrient. If we assume that every individual has a unique requirement and toxicity level for a nutrient (levels below or above which various AHEs are expected to occur), then the dose-response curve is simply a function of the cumulative distribution of these requirements or toxicities. If we let r(x) be the distribution of individual requirements and t(x) be the distribution of individual toxicities, then the risk (probability) of an individual having the AHE if he or she consumes some level X is the \( I_r(x) \) for x above X (or one minus the cumulative distribution of requirement), and the risk of the AHE occurring due to toxicity if he or she consumes at the level X is \( t(x) \) for x below X (the cumulative distribution of toxicity). That we can go back and forth between individual-response distributions to population dose-response curves by simple integration and differentiation is key to the estimation of prevalence, which depends on our being able to estimate these distributional curves.

There are two approaches to estimating the dose-response curve. One method is to estimate the risk or cumulative distribution directly using data from feeding experiments and epidemiology to determine responses for a few levels of intake and then fit these data to a cumulative distribution model to interpolate/extrapolate the whole risk curve. This approach is most often used for toxicities and is employed in formal risk analysis. The alternative method, which is used most frequently for requirements, is to estimate the response distribution and then integrate it to obtain the dose-response curve. For requirements, this involves determining the minimal level that is consistent with health for a number of individuals, a process that is again made more difficult by the inherent and introduced variability. Because we have yet to come up with an experiment that will determine an individual’s requirement with a single observation, we often must study individuals over several different levels of intake and interpolate the results to find a level that we can call requirement. Our data contain error; therefore, this interpolative process uses regression analysis to estimate the effects of diet and individual phenotype and then uses ANOVA to partition the residual variability into sources such as between and within individuals. These two procedures (based on a general paradigm called the general linear model) have special relevance for nutritional data. Extensive descriptions of the varieties and procedures of regression and ANOVA and their practical applications exist at both the introductory (18,19) and more advanced (20) levels. Many statistical programs exist for the actual work; SPSS (21) and SAS (22) are two of the more complete and widely used programming systems.
Regression allows us to examine data with a specific model (equation) in mind, find the parameters of a model that best fit the data, and separate the “signal” (the variability explained by the model) from what is left over, the “noise” (23). The particular model choice is critical to the regression. Two approaches are currently used: fitting a smooth regression curve and estimating where it intersects the zero-balance line (24), and fitting a regression curve with a discontinuity and estimating where that discontinuity occurs (25,26).

Using ANOVA allows us to deal with the fact that not only do individuals differ from one another, but so do measurements taken at different times on the same individual due to observer or measurement error and inherent biological variability. Because the ultimate goal is to estimate the distribution of the population response, it is critical that we remove this intra-individual variability from the data. ANOVA is a procedure that was developed to compare differences between groups to differences within groups and proceeds by partitioning variability into different sources. It is used in the estimation of population requirements and intake distributions to isolate between-individual variability (6,27). Although the procedure is based on the assumption of normality of error, it is a robust procedure in that its results are insensitive to minor deviations from normality.

**General model of prevalence**

A general model of prevalence is based on the concept of a joint distribution. In the same manner that intake and requirement or toxicity can be considered as random variables with individual distributions described as mathematical curves and cumulative distributions defined as areas under these curves, so we can define the joint distribution of intake and requirement or toxicity as a surface over the intake and response plane and cumulative joint distributions as volumes under that surface. Prevalence is then defined as the fraction of the population that has intake below the requirement or above the toxic level and is calculated as the volume to the left (right) of the line where intake is equal to response level (Fig. 1).

Denoting the mathematical function representing the joint distributions as \( p(i,r) \) for the joint intake and requirement distribution or \( p(i,t) \) for the joint intake and toxicity distribution, we write this volume as the integral:

\[
\text{Prevalence of intake below requirement} = \int \int p(i, r) \, \text{for } i < r
\]

and, symmetrically

\[
\text{Prevalence of intake above toxic level} = \int \int p(i, t) \, \text{for } i > t.
\]

We cannot directly estimate these joint distributions because at least at present, we cannot simultaneously determine intake and either the requirement or toxic level for the individuals in a population. The traditional approach, as described above, is to estimate this joint distribution on the basis of independent estimates of the intake and response distributions while assuming that intake and response are not correlated. From this we can rewrite Eq. 2 (for requirement) as the following:

\[
\text{Prevalence} = \int \int p(i, r) = \left( \int r(i) \times i(i) \right) = \int R(i) \times i(i).
\]

This expression is the expected value of intake weighted by its risk. In general, this prevalence integral has proven difficult to calculate, and various ad hoc techniques have been used based on certain simplifying assumptions. The most successful of these are the probability and the shortcut methods (16). However, one of the required assumptions is that intake and response are uncorrelated; although most nutrients are both needed and consumed in amounts that are related to body weight, there is evidence that toxicity may be related to usual levels of intake. Appreciation of this problem leads to many nutrient requirements being expressed in units per body weight. For toxicities, this is obviously not a solution.

An alternative approach to this procedure for estimating prevalence is to consider that more fundamental than an...
individual’s intake and requirement is the difference between the two, or how much an individual’s intake is below requirement or above toxicity levels (Fig. 2). This transforms the joint bivariate distribution into a univariate one and changes prevalence from a function of each individual’s intake and response to an explicit function of their difference. We can define the nutritional deficit \( d \) as \( d = i - r \) and \( d(d) \) as its distribution and rewrite the prevalence integral in terms of its cumulative distribution as follows:

\[
\text{Prevalence} = \int \int p(i, r) = \int d(d) \text{ for } d = i - r < 0 = D(0). \quad (5)
\]

With this formulation, prevalence becomes the cumulative distribution of the nutrient deficit evaluated for \( d = 0 \), and moreover, adjusts for correlation between intake and requirement. Where normal distributions describe the variables involved, as is often the case (5), this leads to the well-studied bivariate normal distribution and thereby reduces the estimation of the prevalence integral to the evaluation of the standard normal curve (28). This is based on the fact that the difference between two normally distributed random variables itself has a normal distribution with a mean that is equal to the difference of the two means and a variance that is simply the sum of the two individual variances corrected for the correlation between the two random variables. In this case, prevalence can be readily evaluated using tables, hand calculators, or spreadsheet formulas.

In addition to simplifying the calculations, rewriting the prevalence integral as a function of nutrient deficit permits us to define a prevalence level for each degree of deficit in a population. We can further define a prevalence curve to describe the nutritional status of a population in terms of fractions that have different degrees of deficit:

\[
\text{Prevalence}(D) = \int \int p(i, r) \text{ over } i - r < D
\]

\[
= \int d(d) \text{ over } d < D = D(D). \quad (6)
\]

This embodies the fact that different degrees of deficit have different consequences in terms of impact on the individual and society. Being mildly malnourished is not as serious as being severely malnourished, and the prevalence curve captures this aspect of a population (Fig. 3).

Once the planner has available the nutrient-deficit distribution within a population, the impact of this particular nutrient pattern can be estimated from the cost of each level of deficit. Defining a cost function \( C(d) \), which quantifies the consequences of various degrees of deficit (measured as, e.g., mortality, morbidity, or hospital days), that function can be used to weight the amount of deficit in the population and the total cost or load on a society due to a specific AHE that results from a specific food-component intake pattern:

\[
\text{Load} = \text{total cost} = \int C(d) \times D(d) \text{ over all deficit levels within a population } \quad (7)
\]

Although Eq. 7 is framed in terms of requirement, exploration of energy requirements and toxicities (risk of obesity) shows that it is equally applicable to upper limits consistent with health.

**Example**

Although this theoretical model presents some difficult manipulations in the general case, it is easy to apply in situations where we have enough data that we can actually estimate distributions. This is illustrated for protein requirement, for which an explicit distribution has been estimated (6).

The protein requirement for adults is log-normally distributed such that the natural logarithm of requirement follows a normal distribution with a mean of \(-0.43\) and a standard deviation of 0.12. It should be noted that the units here and throughout are natural logarithms of \( g \cdot kg^{-1} \cdot d^{-1} \), and that the means involved are negative because the actual values tend to be smaller than \( e \), the base of the natural logarithm \[ \exp(-0.43) = 0.65 \], the estimated median of protein requirement. Examining the US intake data [using unweighted National Health and Nutrition Examination Survey III data (29)], for example, for women \( > 50 \) yr of age, we find that the logarithm of

---

**Cumulative Distribution of Protein Deficit**

![Graph showing cumulative distribution of protein deficit.](https://example.com/graph.png)

**FIGURE 2** Nutrient deficit is defined as the amount by which individual intake differs from requirement. The intensity of each circle (which increases as the circle is below the requirement line) represents the clinical significance of that individual’s intake relative to his or her requirement.

**FIGURE 3** The prevalence curve is the cumulative distribution of the nutrient-deficit distribution. For each level of intake (as a percent of requirement; \( x \)-axis) the fraction of the population that is consuming below that level (\( y \)-axis) is estimated. The derivation of this curve is outlined in the text.
intake is also normally distributed with a mean of −0.29 and a standard deviation of 0.54. Correcting the intake variability to remove within-individual variability (3) reduces the (between-individual) standard deviation of the logarithm of intake to 0.40, and thus we can describe this segment of the US population as log-normally distributed with a mean of −0.29 and a standard deviation of 0.40.

From these data, the distribution of the logarithm of individual deficits (intake minus requirement) can be estimated as also log-normal with a mean deficit of 0.14 (mean intake minus mean requirement) and a standard deviation of 0.42 [square root of the sum of (0.12)² and (0.4)² and explicitly assuming zero correlation between intake and requirement]. Because this distribution is of the difference of two logs, which is the ratio of the original variables, we are able to calculate the fraction of individuals estimated to be consuming less than a specific fraction (F) of their requirement as the probability (P) that a unit normal (z) is less than [0.12 − ln(F)]/0.42. This gives the prevalence for varying degrees of protein deficiencies as shown in Figure 3. To estimate the cost of this protein-intake pattern, we need to assume a cost function that associates a societal cost with each level of deficit. This approach agrees with the traditional approach in that the cost function is 1 for a deficit below 0 and 0 for a deficit above or equal to 0 (no deficit).

As a specific example of estimating a specific prevalence, we calculate the percentage of women in the U.S. over age 50 y who consume <90% of their requirement. The setting is as follows:

\[
F = 0.9
\]

\[
\ln(0.9) = -0.105 \text{ (transforming into log units)}
\]

\[
Z = (0.12 - (-0.105))/0.42 = -0.225
\]

\[
\times \text{ (transforming into unit-normal standard deviations)}
\]

\[
P = 0.30 = 30\% \text{ from the unit-normal table (shown in Fig.3)}
\]

[This can be calculated using the Excel function NORMDIST (x, mean; standard deviation, 1) which calculates the area to the left of a specific value of x when x is normally distributed with specified mean and standard deviation (30). Here the function NORMDIST(−0.105, 0.14, 0.42, 1) was used.]

Limitations and needs

The probabilistic modeling procedure outlined above estimates the risks of populations that can be attributed to consumption levels of food components outside the healthy range. The procedures incorporate correlations between intake quantities and safe limits and make provision for incorporating costs into the calculations. Although much more can be done with the current data, the general implementation of this approach requires more and better data concerning each of the three major components: estimations of intake distributions, requirements and toxicities, and costs.

Dose-response curves. Accurate distributions of physiological requirements and toxicities are essential. Although we have good data on energy (requirement and toxic levels) and total protein requirements (defined by nitrogen balance) and some data on essential amino acid requirements, we have little data elsewhere, especially for upper ranges and toxicities. We especially need information on the interactions such as how the safe limits of one component are influenced by other dietary components.

Population intake distributions. To apply physiological dose-response curves in the real world, tools for accurate intake data are essential. These need to be multivariate to capture how intake levels of various nutrients are correlated within the individual.

Health responses. To be most useful to planners, better definitions of response in terms of health outcome are needed. Protein requirement is currently defined in terms of achieving and maintaining balance, whereas planners need to be able to translate deficits and surfeits of intake into well-defined health risks (or societal costs).

Multivariate data and models. We obviously need more multivariate data on safe levels and intake levels including simultaneous observations for different nutrients. We also need to develop and test more applicable multivariate techniques and models. Although general theoretical multivariate machinery presently exists, we need to work on its application in realistic situations. Progress often occurs when we discover the exceptions and simplifications necessary to move from the theoretical to the practical.

Bioinformatics

In general, we are dealing with complex situations and have limited data, models, and machinery. Although it is possible to make assumptions to fill some of the data gaps and satisfy some of the necessary model restrictions, as reasonable as these assumptions may appear, the full implications are unknown. Moreover, there are not enough data to guide the choice of assumptions. However, modern nutritional study appears to be poised to provide us (and perhaps even overwhelm us) with lots of data.

One approach to providing planners with useful information involves identification of multiple biomarkers, such as biomarkers for health responses, so that samples of individuals with varying health status can be examined to characterize multivariate physiological dose-response curves, and biomarkers for intake, so that samples of populations can be assayed to determine what is being consumed. Given these data, the above statistical modeling provides guidelines for linking the two. However, focused searches for such biomarkers have not yet resulted in great breakthroughs, and the complexity and inherent multidimensional nature of the problem has led to developing interest in metabolomics as a way of pursuing biomarkers (14,31,32).

Movement in this direction takes us quickly from the problems associated with a paucity of data to those that come with a surfeit of data. The problems of deriving information from the massive datasets that are coming from molecular biology researchers has given rise to a whole new field of bioinformatics, which combines the various computer, mathematical, and statistical techniques for working with large databases. Although there are many different definitions of bioinformatics, the study has two broad areas of concern: data management and exploration. Data management includes computing and information technology for dealing with the nontrivial problems of capturing, coding, cataloguing, storing, and searching large datasets (33). Data exploration of masses of data includes primarily visualization, identification, and testing of multivariate patterns using various techniques of clustering, data mining, and pattern recognition (34). The basic statistical problem lies in the quantity of data to be sifted through, but we also don’t quite know what we are looking for. Traditional statistics techniques were developed primarily to estimate and test parameters of given models and to partition variability into
given sources. Bioinformatics moves the primary interest into generating models and finding sources of variability.

Once sets of potential biomarkers for intake and response are found, statistical methods must be used to relate multivariate exposures with multivariate outcomes. Whereas classical statistical techniques explore how single-dependent variables are related to sets of independent variables, now we must move to techniques that explore the relationships of sets of dependent variables (health biomarkers) with sets of independent variables (intake biomarkers). These include multivariate ANOVA, factor analysis, canonical correlation, and principal component analysis (35). Statistics has sketched out theoretical approaches in these areas, but few successful applications in the field of nutrition exist due largely to complexity and data limitations. Although these tools are beginning to be used in nutritional study; full utilization awaits development of further insights into the underlying processes and organized programs for collection of relevant data.

Summary

A dominant feature of any biological phenomenon is its variability or lack of precise reproducibility. Connecting diet and health in any way that is useful to planners means describing complex relationships that are somewhat concealed by an incredible overlay of variability from a large number of potential sources. Available data differ from one another because they represent different individuals with different consumption patterns measured in different ways at different times. The tools of statistics can be used to estimate the responses of individuals to different intakes and different combinations of food components. Additionally, a major task is to estimate how individuals differ from each other and factor out within-individual and between-measurement variabilities. Probabilistic models can be used to integrate these data into useful information for planners, but the parameters of these models need to be estimated. Regression is the major tool used to estimate response curves, and ANOVA is used to tease out the variability that can be ascribed to dietary levels and individual characteristics from variability due to within-individual and measurement variations. Like all statistical techniques, these procedures need models and data that span the full range of the variables of interest and they need replicated measurements at all levels.

Formally, the distribution of nutrient deficit (or equivalently, surfeit) is the convolution (weighted average) of the intake distribution for any specific population. These cumulative distributions describe the fraction of a population that consumes below (above) the individual requirements (toxic levels) as a function of the degree of deficit (or surfeit). Although in general the calculation of the convolution of two functions is a complex procedure, for normally distributed variables, it is again a normal distribution that can be easily specified and evaluated. This permits the user/planner to easily introduce a cost or load function into assessment or planning, which relates the degree of deficit (or excess) to the societal impact, namely, the impact of individual substances in the diet.

We now have the models and machinery to provide planners with better information with which to assess health status and plan for the supply and control of food supplies. These procedures involve linking intakes and physiological responses, especially minimums (requirements) and maximums (toxicities), and factoring in the cost or load that limited or excess food supplies impose on society. We have much of the machinery, but we need the data to test and refine our models, and we need to push into the next big challenge: linking the nutrients together.

Much more can be done with the present data and with extensions of the present approaches; however, to proceed much further, we need data on a grander scale, data that are more extensive and more precise, data that are hinted at by metabolomics. Analyzing these huge databases and integrating the results to arrive at information that the planners can use is a major challenge of nutrition and statistics.

LITERATURE CITED

    John Wiley and Sons, New York, NY.

29. National Center for Health Statistics (1994) Plan and Operation of the
    Third National Health and Nutrition Examination Survey, 1988–94. Series no. 1,
    Report no. 32, National Center for Health Statistics, Hyattsville, MD.

    Press, Redmond, WA.


    potential for assessment of adequacy and safety of amino acid intake. J. Nutr. 133:
    2097S–2100S.

    York, NY.


    Chapman and Hall, New York, NY.