Introduction and Background

The medical literature has experienced intense criticism in the past few years [1-9]. A recent editorial in *The Lancet*, for example [1], suggested that many of the conclusions in the literature, "perhaps half, may simply be untrue." Observational studies are particularly problematical in that, while such studies generally include disclaimers that the identified associations do not imply causality, frequently there is the general sense that causality is implied. Inappropriate use or interpretation of statistics, e.g., misunderstanding the meaning of p values [2,3] and questionable practices such as intention-to-treat [4] appear as common targets of criticism. The problem, however, is more global and rests with interpretation of the data and what standards can be used to decide when a causal link has been established between variables. Bradford Hill [10] proposed nine criteria although was clear that, rather than precise standards, they were extensions of common sense. In our view, uncritical reliance on statistics at the expense of understanding the underlying biology constitutes the major problem in the medical literature. Emphasizing the biological import of the data implies a statistical approach that allows different observers to make different interpretations, an idea appreciated in recent renewed interest in Bayesian methods which identify probability with the observer's belief, strengthened by data. We think that one of the problems in the medical literature is the failure to appreciate and utilize this aspect of research, relying, instead on identifying statistical significance which by itself is not sufficient to decide on causal relations.

In some cases, the mode of presentation of the statistics can lead to incorrect conclusions. For example, reporting relative values is widely criticized as misleading in many cases [7-9]. Odds ratio (OR) for two events will have the same value regardless of the absolute value of the odds for the two events. Relative risk (RR; "risk" is probability) and hazard ratio (HR; risk for a fixed time period) similarly hide information about the two events or experiments that are to be compared. The widely cited study of Mediterranean Diets [11], for example, shows a 30% reduction in HR compared to a low fat diet. The two diets, had reductions in risk of 6.4 % and 4.8 %, respectively or an absolute difference of 1.6 %. (The HR of 30 % would have been true if the observed values were 64 % and 48 % where the difference of 16 % would have to be taken more seriously). Other factors enter into the decision as to which method of presentation is most meaningful. An absolute difference of 1.6 % might well be meaningful in a vaccine trial where it is known exactly who received vaccine and who didn't. No such level of reliability obtains in most dietary trial where food records may have large error. Here we address the presentation of the data on risk that may provide a more intu-

itive feeling for what happens in an epidemiological experiment. It is likely that most readers seek straightforward information and want to know what their actual risk is in following a particular intervention.

Here we show the value of a simple method for analyzing cohort studies by casting the problem in the form of a diagnostic test. A diagnostic test will be characterized by its *sensitivity*, the probability that a positive test predicts a positive outcome and its *specificity*, the probability that a negative outcome indicates a true failure. There is, in fact, nothing unique in the mathematical formalism of a diagnostic test. We describe how the independent variable, food consumed or drug taken may be considered as diagnostic of, or prognostic for outcome; as an example how fat intake can be predictive of some disease. We use a simple 2 x 2 matrix, commonly, if not universally, taught to students, that allows rapid calculation of conditional probabilities. The term "natural frequency" is sometimes used [5-6] but "diagnosis method" seems more explanatory. The technique is completely equivalent mathematically to formal Bayesian statistics and can be derived from it, although this is not required [5-6].

We use the diagnosis method to re-analyze a published paper on the effect of red meat on risk of mortality and specifically a risk for cancer [12]. We show that the original conclusion of the paper, that high consumption of red meat is diagnostic, or predictive, of mortality risk, is unfounded.

Definition of Odds and Risk.

The *probability* of a particular outcome in a statistical trial is the number of subjects with that outcome divided by the total number of subjects or, in the language of games of chance, the number of winners divided by the number of all outcomes. Probability is also referred to as *risk*.

The *odds* of winning is the number of winners divided by the number of losers. At low probability the two measures are very close. (The probability of drawing the ace of spades is $1/52 = 0.0192$ or 1.92 %, The odds are $1/51 = 0.0196$ or 1.96 %

When comparing two trials, the ratio of the two outcomes, the *relative risk (RR)* or *odds ratio (OR)* is determined. Hazard rate is similar to the other measures but as the name implies, includes a time factor, that is the number of winners per particular time and, obviously, the *hazard ratio (HR)* compares two trials. The important point in what follows is that when there are a small number of winners, for example low prevalence of disease or small number of cases in the time period studied, the three measures, HR, OR and RR are essentially the same and we will use them interchangeably. RR is generally easiest to calculate because

the number of cases and the total n is immediately available from most cohort studies.

Analysis of Cohort Studies - Does Red Meat Cause Cancer?

Pan, *et al.* [12] followed 37, 698 men from the Health Professionals Follow-up Study (HPFS) for 22 years and followed 83, 644 women from the Nurses Health Study (NHS) for 28 years. Food intake was determined from food frequency questionnaires (FFQs). Red meat was defined as "unprocessed," including beef, pork, or lamb as main dish, sandwich or mixed dish, or as "processed," bacon, hot dogs and "sausage, salami, bologna and other processed red meats." The raw data were subjected to multivariate adjustment for major lifestyle and dietary risk factors and hazard ratios (HRs) were calculated for the effect of one serving/d increase of red meat. The raw data found HRs for total mortality, cardiovascular disease (CVD) and cancer mortality in the range of 1.10 to 1.16. Age-adjustment or multivariate-analysis gave higher HRs, 1.39 and 1.17 for the NHS for total red meat and mortality due to cancer. (An HR of 1.0, or 50-50 odds, means that there is no difference between highest and lowest red meat consumption. The conclusion was: "Red meat consumption is associated with an increased risk of total, CVD, and cancer mortality." We will re-assess the results from the NHS part of the study and focus, in particular, on the risk of cancer.

Table 1. Baseline Age-Standardized Characteristics of Participants According to Quintiles of Total Red Meat Consumption

Characteristic	Q1	Q2	Q3	Q4	Q5 "POSITIVE TEST"
			Nurses' Health Study		
Participants, No.	16 499	17 247	16 461	16 603	16 834
Age, mean, y	47.3	46.0	45.8	45.3	46.0
Total red meat intake, mean, servings per day	0.53	1.04	1.52	2.01	3.10
Physical activity, mean, MET-h/wk	16.9	13.9	13.8	13.3	12.4

Table 4. Cancer Mortality According to Red Meat Intake in the Nurses' Health Study

Variable	Q1	Q2	Q3	Q4	Q5	HR (95% CI) "TRUE POSITIVE"
			Nurses' Health Study			
Total red meat						
Cases/person-years, No.	1264	1191	1185	1263	1488	
Age-adjusted model	1 [Reference]	1.04 (0.96-1.13)	1.08 (1.00-1.17)	1.19 (1.10-1.29)	1.39 (1.29-1.50)	1.21 (1.17-1.25)
Multivariate model	1 [Reference]	1.05 (0.97-1.14)	1.10 (1.01-1.19)	1.15 (1.06-1.25)	1.17 (1.08-1.28)	1.09 (1.05-1.13)

Table 1. Data from Tables 1 and 4 from Pan, *et al.* [12] Subjects in quintile 5 (most red meat) are considered to have had a "positive test." Those subjects who died from cancer are "true positives."

Table 1 collates excerpts of two tables of the NHS data from Pan, *et al*. [12]. Data from the people in the study were divided into five approximately equal groups on the basis of total red meat intake. Table 1 from the original paper provides the number of people in each quintile. Table 4 shows the number of people who had died from cancer.

	CANCER	no CANCER	TOTALS	
(Q5) High red meat	1488	15346	16834	PPV = 1488 / 16834 = 0.088 or 8.8 %
∑(Q1-Q4)	4903	61907	66810	
TOTALS	6391	77253	83644	
	sensitivity = 1488 / 6391 = 0.23 or 23 %	specificity = = 6197 / 77253 = 0.80 or 80 %	Likelihood ratio = = 0.23 / (1 - 0.80) = 0.23 / 0.20 = 1.15	

Table 2. Diagnosis matrix for data from Pan, *et al*. [12] as shown in Figures 1 and 4 from that paper (Table 2). True positive (high red meat consumption and cancer) shown in green. True negative (not high red meat consumption, no cancer) shown in red.

sensitivity = true positives / (true positives + false negatives) = 1488 / 6391 = 0.23 or 23 %
specificity = true negatives / (true negatives + false positives) = 61907 / 7723 = 0.80 or 80 %
PPV = true positives / (true positives + false positives) = 1488 / 16834 = 0.088 or 8.9 %
likelihood ratio = = (sensitivity / (1 - specificity)) = 0.23 / 0.20 = 1.15

Sensitivity and Specificity

In a diagnostic test, patients who test positive and who actually have disease are referred to as *true positives*. Patients with disease who test negative are *false negatives*." The *sensitivity* of a diagnostic test tells you how likely the test is to identify people with disease and is defined as the true positives divided by all patients that actually have the disease.

sensitivity = true positives / (true positives + false negative (have cancer))

The 2 x 2 matrix shown in **Table 2** can be used to determine the characteristics of red meat consumption as diagnostic for dying from cancer. The columns

show the number of subjects with or without cancer and the rows, the number of people with or without high meat consumption. The total number of big red meat eaters are those in quintile 5 = 6,391. Of this group, those who also died of cancer (true positives) = 1,488 as shown in green.

The sensitivity of the test is 1,488 / 6,391 = 0.0234 or 23.4 %.

Thus red meat consumption misses more than ¾ of people whose death was caused by cancer. This constitutes a very poor sensitivity. (If the cohort were divided into quintiles at random, each quintile would have a 20% sensitivity). In diagnostic medicine, colonoscopy has a sensitivity of 80-100% for detection of colorectal cancer[14]. If colonoscopy had a sensitivity of only 20% it would not be considered for use or recommended as a screening procedure.   Intuitively, a positive result on a high sensitivity test is rarely wrong but, as brought out in the Discussion, high sensitivity is not a sufficient criterion for a reliable test and other factors must be considered.

Specificity in a diagnostic test refers to the ability of the test to identify those patients without disease. Specificity is given by the ratio of the *true negative*s to all patients who do *not* have the disease, (true negatives + false positives), that is, the probability that someone who does not have the disease is identified by the test.

specificity = true negatives / (true negatives + false positives (no cancer))

In Pan, *et al*., 77,253 patients did not die of cancer. The red meat test showed that 61,907 of these were not in the high red meat group.

Specificity = 61,907 / 77,253 = 0.8013 or 80.1 %.

This number indicates the diagnostic value of a negative test. Specificity and sensitivity tend to go in opposite directions — if a positive test is not reliable you may be confident about a negative result. Participants who had not been high red meat eaters were very unlikely to have died from cancer. Of course, even if they had been big meat eaters, they are not likely to have died from cancer. The data in Pan, *et al*. show that, because of the low prevalence of cancer and the low sensitivity of the test, there is little predictive value in red meat consumption in contradiction to the authors' conclusions.

Positive Predictive Power.

The most intuitive measure of the value of a diagnostic test may be its *positive predictive value (PPV)* which is defined as the probability that a positive test actually indicates disease.

PPV = true positives / (true positives + false positives)

Similarly for *negative predictive value, NPV*

NPV = true negatives / (true negatives + false negatives)

The positive predictive value (PPV), from the matrix in **Table 2** = 1,488 / 16,834 = 0.088 = or 8.9 % indicating that that high consumption of red meat is not predictive at all.

Although the positive predictive value is very low, it is not zero and it is important to decide if it provides a meaningful number. Does it represent an accurate but small effect that, if scaled up to a large population, would provide a meaningful health outcome? Alternatively, does it simply reflect the inherent variability in the trial? Comparing the effect of high red meat consumption with low (Quintile 1 in **Table 1**), we get a sense of the real effect of red meat. The probability of having died of cancer in Q1 = 1, 264 / 16,499 = 0.077 or 7.7 %. The difference between the *extremes* of red meat consumption, then, is 1.2 %. (Alternatively, one might consider that there is a 91 % chance that a participant's death was *not* due to cancer if they were in Q5 and 92 % chance in Q1).

The small difference in positive predictive values follows from the low prevalence of cancer in the population sample. However, the real practical effect depends, as well, on the reliability of the independent variable, the proposed risk factor. Values for red meat appear as "servings" from a food frequency questionnaire (FFQ). Such data are subject to very large error [14-16]. The tables on experimental consumption in **Table 1** indicate that e.g., thinking that you had two savings of red meat when you only had one, can move you from one quintile to another. Several investigations have shown the high variability in data from FFQs and it is claimed that because one value, total caloric intake, is constrained by a range of known normal average consumption, error is frequently obvious but one cannot reasonably assess the degree of error at all [16]. It is important to emphasize that the error may be systematic — people tend to under-report fat consumption — but it is impossible to say the nature of the error, that is, whether or not it is in favor of the hypothesis. This error in assigning quartiles, in combination with

the very low difference in positive predictive values suggests that nothing can be decided from this data.

Likelihood ratio

Less widely used than the other parameters, *the likelihood ratio* compares the probability of having the disease if testing positive compared to testing negative.

likelihood ratio = sensitivity / (1- specificity)

Likelihood is described as a measure of the amount of information that a medical test provides about individual patients [15]. Values close to 1 are considered of little practical diagnostic value. In the current case the likelihood ratio = 0.23 / (1 - 0.8) = 0.23/0.2 = 1.15

As previously described, under conditions here, where there is low prevalence and small changes the odds ratio (OR), relative risk (RR) and hazard ratio (HR) are essentially the same. The OR can be calculated from these numbers = 8.9/7.7 = 1.15 close to the value found by Pan, *et al*. in their multivariate model (**Table 1**). While this might be reported as a 15 % difference, the 1.2 % absolute difference found by the diagnosis method tells the story better. It should be emphasized that, as noted below an HR in the range of 1.2, even if statistically significant, is too small to be considered practically significant.

Discussion

Does red meat cause cancer? How do we know? "In what circumstances can we pass from…observed association to a verdict of causation? Upon what basis should we proceed to do so?" The questions were posed by Bradford Hill whose cigarette-lung cancer study remains the classic case of causation implied by observational studies. His Presidential Address of 1965 provided nine criteria [10].

Hill's criteria give a point of departure to distinguish biologic importance from statistical significance.Statistics may tell us what to look for in the experiment. "Beyond that they contribute nothing to the 'proof' of our hypothesis" Hill insisted [10]. He made it clear that these criteria were not meant as hard-and-fast rules but rather general principles following from common sense and that

> "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*."

How then *can* statistics be used? "First upon my list I would put the strength of the association." Hill pointed to the fact that the OR for cigarette smoking and lung cancer was 20 and, for heavy smokers, 30. The OR in the red meat-cancer study was in the range of 1.2, close to 1.0 or equal odds. From another perspective, epidemiological data in a toxic tort case will rarely be introduced into evidence with an OR of less than two [18].
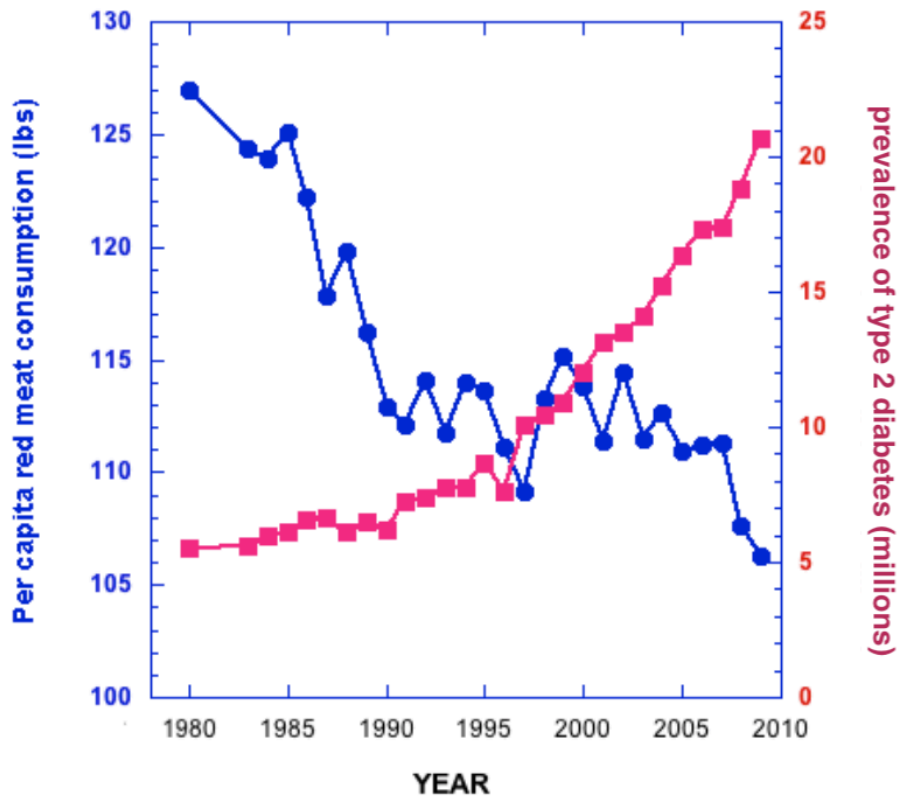

Figure 1. Red meat consumption and prevalence of type 2 diabetes. Data from USDA (e.g. http://bit.ly/2J51hlG) and CDC (https://www.cdc.gov/ diabetesstatistics/slides long_term_trends.pdf), respectively.

It must be emphasized that "size" means size in the biological sense. This rests with understanding of the experiment and should be distinguished from "effect size" as used in statistical analysis. The latter only tells you about the relations of

statistical parameters. Only the experimental details can tell you about the biological magnitude. The study discussed here was based on "validated" food frequency questionnaires (FFQ's) which are known to have great error [14-16] — "validated," however, did not mean certified as accurate; the questionnaires were validated to have *definable error*. Comparison of diet records with subsequent answers to FFQ's show great variability in perception and recall of red meat consumption, and typically identify a Pearson's correlation coefficient of 0.5 - 0.6) [14]. Because errors are due to memory lapses, inaccurate assessment of portion size and unconscious attempt to conform to some standard it is likely that people under-report consumption of fat and other forbidden fruit. It is not possible, however, to know whether the error is in one direction or the other, that is, whether it can be considered to support or weaken the hypothesis being tested. Thus the minimal sensitivity of red meat as diagnostic does not mean that there is a small risk that can be scaled up. It simply means that the results are too imprecise to decide on the role of red meat in cancer.

Evidence on Red Meat as a Causative Agent

Describing results in terms of the "diagnosis procedure" may provide a more intuitive and familiar form for many readers. The limited value of red meat to diagnose/predict cancer, like any statistical parameter, does not, by itself, exclude risk — a vaccination trial may have very low odds ratio but the small benefit *can* be scaled up because you know who got the vaccine and who didn't Here, the great error in the independent variable *supports* the conclusion that red meat has little effect on cancer development.

While both popular and medical opinion try to dissuade the population from dietary red meat consumption — can this be justified? [19-20]. We have previously shown that a paper on changes in red meat consumption and diabetes was even more problematic than the current case: The NHS showed a similar low sensitivity of 8.0 % for those subjects with large changes in red meat consumption and 6.1 % in the low group. In a published letter to the editor, it was pointed out that any positive association was implausible given the opposite trajectories of the incidence of diabetes and the consumption of red meat the past thirty years [21] (**Figure 1**).

A recent analysis of the effect of red meat on colorectal cancer and subtypes found, like Pan, *et al*., low HRs, low positive predictive values and absolute differences in risk in the range of 1 % [22]. Those who advocate for reduced meat ingestion cite epidemiological studies of the type considered here. We suggest that these be re-evaluated before rushing to accept this kind of judgment.

Summary and implications for quality in the medical literature

Statistics as a discipline has been criticized for evolving in the direction of internal rigor at the expense of fundamental comprehension of the underlying data (e.g. [23]). In our view, the medical literature has followed this trend with arbitrary "levels of evidence" and "gold standards" and a willingness to accept *statistical significance* regardless of magnitudes of biological change, while ignoring normal practice, intuition and even common sense. At the same time, a high degree of bias is palpable in the literature and statistics are presented in a way to favor the authors' point of view. The most general description of the problem is that authors do not challenge their conclusions but rather strive to demonstrate consistency.

The diagnosis procedure presented here is intended to make the statistics of cohort studies more accessible. *Positive predictive value* is easily and intuitively appreciated. The poor predictive value of a mammogram is a staple of statistical education [5,6] and the number of demonstrations on YouTube greatly exceeds the number of lists of the world's worst aircraft or greatest operatic tenor. By itself, the method is not definitive for deciding on causality but a poor PPV points to the need for other evidence from Hill's criteria or other principles, particularly the need for biological mechanism.

In the end, it is the implausibility or even foolishness of the that red meat might be a cause of cancer, diabetes and/or everything else. The lack of mechanism is probably dispositive; It is unlikely that even sub-types of red meat are well enough defined mechanistically to be considered as causative agents. The poor predictive value shows how high the stakes are for other evidence Finally, it is not clear what health benefits of any kind can be shown to be associated with the large decline in red meat consumption shown in **Figure 1**.

We do not propose that the diagnosis approach is required or even the best method of deciding on the causal nature of an association. By itself, no statistical analysis can be definitive. Low statistical risk, however, raises the stakes on other evidence, particularly biological plausibility. We must know the question that is being asked and the magnitude of the effect that is acceptable. Acceptable means that it reflects real physical change (as distinct from the statistical effect size). In our view, the use of statistical methodology, and of a particularly dubious kind, has come to dominate and distort scientific judgment and research practice. *Statistics should be servant rather than master.* The inversion of this principle accounts for much of the lack of reliability and reproducibility noted and lamented about the state of the medical literature.

Bibliography

1. Horton R: Offline: **What is medicine's 5 sigma?** *The Lancet* 2015, **385**:1380.

2. Colquhoun. **Review article: An investigation of the false discovery rate and the misinterpretation of *p*-value**. 2014 *Royal Society Open Science*. **19** November.

3. Chavalarias D, Wallach JD, Li AH, Ioannidis JP: **Evolution of Reporting p Values in the Biomedical Literatur**e, 1990-2015. *JAMA* 2016, **315**(11): 1141-1148.

4. Feinman RD: **Intention-to-treat. What is the question?** *Nutr Metab (Lond)* 2009, **6**:1.

5. Gigerenzer, G. *Calculated risks: how to know when numbers deceive you*. New York: Simon & Schuster; 2002.

6. Gigerenzer, G. **What are natural frequencies**? *BMJ* 2011; **343**: d6386.

7. Gigerenzer G. **Misleading communication of risk. Editors should enforce transparent reporting in abstracts**. *BMJ* 2010;**341**(c4830).

8. Feinman RD, Keough SM: **Ethics in Medical Research and the Low-Fat Diet-Heart Hypothesis.** *Ethics in Biology, Engineering and Medicine* 2014, (2):149-159.

9. Feinman RD(2018) **What's Really Wrong with Medical Research and How to Fix It**. Journal of Evolution and Health: **2 (3)**:, Article 10. https://doi.org/10.15310/2334-3591.1069

10. Hill AB: T**he Environment and Disease: Association or Causation?** *Proceedings of the Royal Society of Medicine* 1965, **58**:295-300.

11. Estruch R, Ros E, Salas-Salvado J, Covas MI, Corella D, Aros F, Gomez-Gracia E, Ruiz-Gutierrez V, Fiol M, Lapetra J, Lamuela-Raventos RM, Serra-Majem L, Pintó X, Basora J, Muñoz MA, Sorlí JV, Martínez JA, Martínez-González MA; **PREDIMED Study Investigators. Primary prevention of cardiovascular disease with a Mediterranean diet**. N Engl J Med. 2013;**368**(14):1279–90.

12. Pan A, Sun Q, Bernstein AM, Schulze MB, Manson JE, Stampfer MJ, Willett WC, Hu FB: **Red meat consumption and mortality: results from 2 prospective cohort studies**. *Arch Intern Med* 2012, **172**(7):555-563.

13. Hu FB, Rimm E, Smith-Warner SA, et al. **Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire**. *Am J Clin Nutr*. 1999; 69 (2) :243-249.

14. Salvini S, Hunter DJ, Sampson L, *et a*l. **Food-based validation of a dietary questionnaire: the effects of week-to-week variation in food consumption**. *Int J Epidemiol.* 1989;**18**(4):858-867.

15. Frank E. Harrell Jr, PhD; Robert M. Califf, MD; David B. Pryor, MDKerry L. Lee, PhD; Robert A. Rosati, MD, *et al.* **Evaluating the Yield of Medical Tests** *JAMA.* 1982;**247**(18):2543-2546.

16. Archer E, Marlow, ML, Lavie, CJ J Clin Epidemiol. **Controversy and debate: Memory-Based Methods Paper 1: the fatal flaws of food frequency questionnaires and other memory-based dietary assessment methods.** 2018 **104**:113-124. doi: 10.1016/j.jclinepi.2018.08.003. Epub 2018 Aug 17.

17. Garborg K., Holme,Ø. Løberg, M. Kalager , M. Adami, H. O. Bretthauer. M. *Annals of Oncology*, **24**(8),1963–1972

18. Lin A. Beyond Tort: **Compensating Victims of Environmental Toxic Injury**.*Southern California Law Review*,. 2004;**78**(8).

19. Williams KA, Sr., Patel H. **Healthy Plant-Based Diet: What Does it Really Mean?** *J Am Coll Cardiol.* 2017;**70**(4):423-5.

20. Pan A, Sun Q, Bernstein AM, Manson JE, Willett WC, Hu FB: **Changes in Red Meat Consumption and Subsequent Risk of Type 2 Diabetes Mellitus**: **Three Cohorts of US Men and Women.** *JAMA Internal Medicine* 2013:**74(**1-8.

21. Feinman R: **Red Meat and Type 2 Diabetes Mellitus**. *JAMA Internal Medicine* 2013, **174(**4):646.

22. Bernstein AM, Song M, Zhang X, Pan A, Wang M, Fuchs CS, et al. **Processed and Unprocessed Red Meat and Risk of Colorectal Cancer: Analysis by Tumor Location and Modification by Time**. *PLoS One*. 2015;**10**(8):e0135959.

23. Ziliak, ST, McCloskey. *The Cult of Statistical Significance. How the standard error cost us jobs, justice, and lives*. Ann Arbor: The University of Michigan Press. (2009)