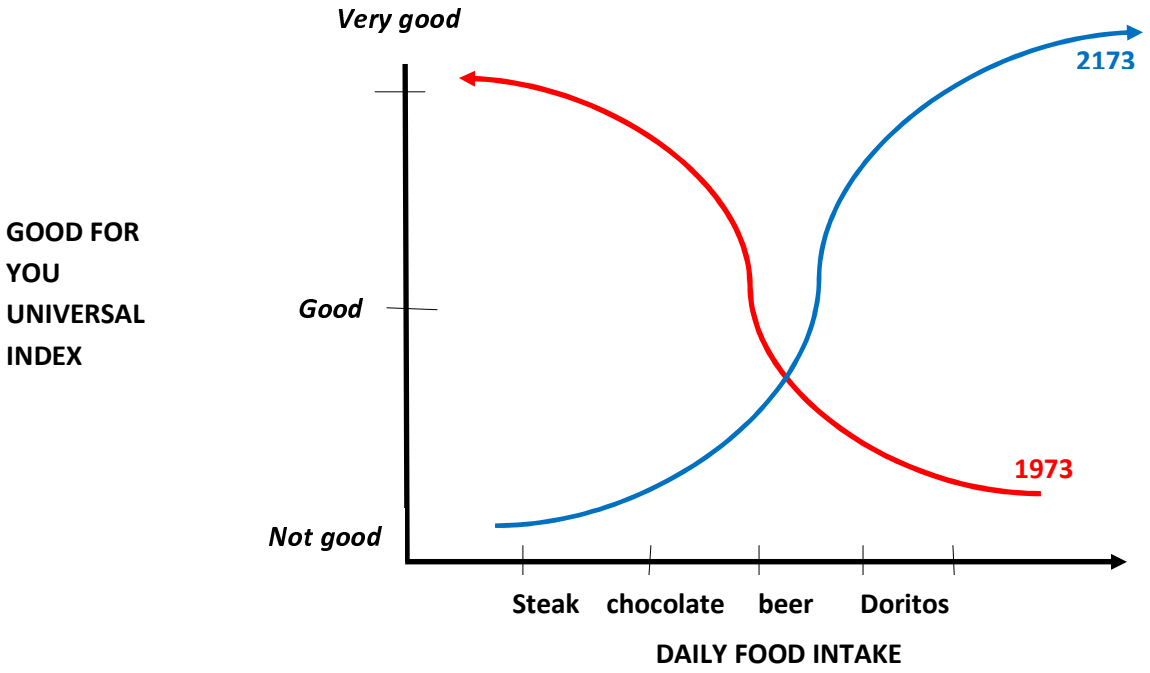
MEDICAL RESEARCH AND THE SLEEPER CURVE

The following is a transcript from the Woody Allen movie *"Sleeper*" in which Woody Allen is cryogenically frozen in 1973 and awakens in the new American society of 2173. Allen, former owner of the Happy Carrot health food store in Greenwich Village was requesting breakfast from his doctors after his long ride through time. The ensuing dialogue, presented below, establishes the framework for our discussion of what is valid and invalid research.

Dr. Melik: Yes, this morning for breakfast. He requested something called wheat germ, organic honey and tiger's milk.
Dr. Agon: [laughs] Oh, yes. Those were the charmed substances...That some years ago were felt to contain life-preserving properties.
Dr. Melik: You mean there was no deep fat? No steak or cream pies? Or hot fudge?
Dr. Agon: Those were thought to be unhealthy, precisely the opposite of what we now know to be true.
Dr. Melik: Incredible.

The funny events in the movie Sleeper parallel what is happening today in our modern society in which almost everything is decided by what researchers tell us...what is good for us, and what is bad for us; what will make us stronger and better looking; what will not hurt us and what will . Many times we are a little shocked or surprised when our most accepted and researched findings are reversed, as happened in Woody Allen's' movie *Sleeper.* In the year 2173, Woody learns that health food is now bad for you and chocolate fudge, steak and cream pies are now good for you. This phenomenon has a name thanks to Steven Johnson, and its hallmark is the "Sleeper Curve" which I envision as looking like this:



This graph demonstrates the Sleeper principle of research: what was once bad for you is now good for you, conversely, what was once good for you, is now bad for you.

Guide to healthcare research and predict a Sleeper Curve:

There is truth within every absurdity, and the truth in the Sleeper Curve is that not all research is equally valid, and therefore will not stand the test of time or use; in healthcare, we have our own Sleeper Curves.

Medical literature and journals have become saturated with research studies that explain the advantages or disadvantages of new or existing treatments and interventions. More and more, we are witnessing prehospital care being integrated into hospital care through advance notification systems (alerts) and more sophisticated diagnostic and treatment modalities such as STEMI and stroke recognition. With this recognition of EMS as an important provider in the patient care continuum, there is a noticeable increase of research being directed toward EMS care; and with the importance of evidence-based medicine, we can anticipate even more research directed our way.

This development, while acknowledging the contributions of EMS to good patient care, also brings elements of uncertainty with it in the determination of what research is valid, or good research, and which is spurious and will not contribute to excellence in patient care we all seek. Our number one goal in EMS is to serve our patients and provide the best care possible, and in order to provide that care we have become very dependent on research studies to help determine the best possible course of treatment for our patients. The examination of research is not just for scholars anymore; there is such saturation in the medical literature that we all must become familiar with its evaluation and make judgments on its merits.

Current evaluation strategies may actually be more complicated than the actual research itself. These methods require a large amount of study time to understand and even more time to utilize effectively in the analysis of research. Fortunately, there are excellent researchers that have simplified the process and we have borrowed from their recommendations.

THE ONE STUDY SIGNIFICANCE PROBLEM:

Results of new research should not be considered as a true, conclusive finding if it is the only study reporting a result based on a significant statistical value, or a "p value" This statistical "p value" is the probability of obtaining a result that would be opposite of what the researcher hypothesized. Usually the threshold of significance is set at 0.05, so any p value equal to that or less is considered statistically significant. There is nothing wrong with setting any standard for significance, the issue becomes that this means nothing conclusive when only one study is involved. Results from one study should be viewed with skepticism until there are other studies performed that will either support or refute the conclusion of the "new" study. And, whenever possible, we should search for studies that contradict the new finding as this will provide insight into why the "new" finding may not really be true after all. It is difficult to find contradictory studies as they usually are not published, a form of "negative" bias or publication bias. Example: The 1997 study conducted by UK doctor Andrew Wakefield in which he announced evidence of the vaccine-autism link provides good evidence of this problem. This small study published in the prestigious British medical journal *Lancet* was not discredited completely until after approximately twenty studies found no relationship between autism and vaccines. It was also found that Dr. Wakefield took money from attorneys representing autistic children and was disciplined for professional misconduct. However, this did not take place until years after, and this misconception persists today due to persistent citations of his study in other literature.

SMALL EFFECT SIZE:

Although a little more difficult to understand, there should be lots of suspicion around a study that has small relative risk values. Just to understand relative risk, a treatment or drug that results in a relative risk (e.g. death) of 1 means no increase or decrease of risk of death. A small difference either way may be meaningless, but is reported as meaningful. Relative risks of 1.2, or 0.98 are very close to 1, and even with a very large sample size, this may not be meaningful. Example: A cancer drug causes a patient's relative risk of death to drop from 3% to 6%; therefore the relative risk has decreased 50% according to the researcher. However, one can easily see this is misleading; 50% decrease sounds much better than a decrease from 6% to 3%!

THE DISGRONIFIER TEST:

Research designs or statistical methodologies that use outcome measures which are not universal or commonly understood should be viewed with suspicion. For example, scales that measure impairment, cognitive abilities, or motor skills etc. should be carefully evaluated to ensure that they are commonly used and understood. If not, scale outcomes can be reported, appear to show real results, but will not

be true findings and therefore mislead the reader. Example: If studies use lesser known and/or complicated outcome measures to assess clinical effectiveness, then the results are difficult to evaluate and call into question if a finding is true. Measurement scales should be straightforward. Consider that probability of survival models for trauma in research studies may use the TRISS scale or the ASCOT scale to predict the survivability of a trauma patient based on injury severity. Both of these scales are derived from other scales; the TRISS scale is derived from the ISS scale and the Revised Trauma Score (RTS), and the ASCOT uses the RTS, GLASGOW COMA scale and the Abbreviated Injury (AIS scale) for its composition. This heterogeneity makes comparisons very difficult across studies and makes the possibility of an actual true finding less likely and not generalizeable to other populations. The outcome measure itself diminishes the likelihood of a true finding due to its complexity.

Statistical methodology or analysis can be used to "prove" almost anything. Perhaps the most egregious example comes from the last century when statistician Sir Francis Galton "proved" that 100 upper class babies would benefit society more than 1000 babies born to the lower classes. He used regression modeling, which he pioneered and developed, and was a new statistical analysis technique in 1901. (Appendix A)

SELECTIVE REPORTING:

Data manipulation either of methods or reporting of results is an indicator of bias in a study and will call into question its results. When researchers do not report all their results, and opt to report only the favorable or "best" results, this is called selective reporting and can be a sign of a biased study. Research that employs statistical methodologies that are not customary and accepted, and use instead experimental methods, make it much easier to engage in selective reporting; for example, utilizing a little known statistical test, reporting all the numerical results, but emphasizing the "best" results. Example: Between 1966 and 1995, 47 studies were conducted on acupuncture in China Taiwan, and Japan, and all found acupuncture was a valid treatment. During the same time frame, 94 studies on acupuncture were conducted in the UK, Sweden and the United States. Of those, only fifty-six percent were yielded significant positive results. Obviously, some researchers were selectively reporting the parts of their studies, reporting significant favorable results and ignoring the insignificant results.

PUBLICATION BIAS:

This is common in drug studies and means that only studies that have positive significant findings are published. What happens to the studies that do not have positive results? Could they not be valid as well? Yes! Can they be accessed? Not easily. Example: It is now being shown that the benefits of some anti-depressant drugs were exaggerated and trials which showed their ineffectiveness were never published. Out of 24 trials testing the clinical effectiveness of such well-known drugs such as Zyprexa and Seroquel, four were not published due to negative results. Of the twenty that were published, four had negative results, but were not reported as negative by the study authors.

LOW POWER:

This means that in order to save money and time, many researchers will not obtain enough sample size. Large sample size randomized studies have a greater chance of being true with more subjects being studied; however, sometimes these large studies produce very small effect sizes. Example: A hypothetical study of ten thousand subjects found in a registry data base may produce an odds ratio of 1.12 with a significant p value; this is a small effect size and means that the odds of a desired or undesired effect are only slightly greater than 1, and therefore not very big. This small effect size calls into question the actual validity of the finding.

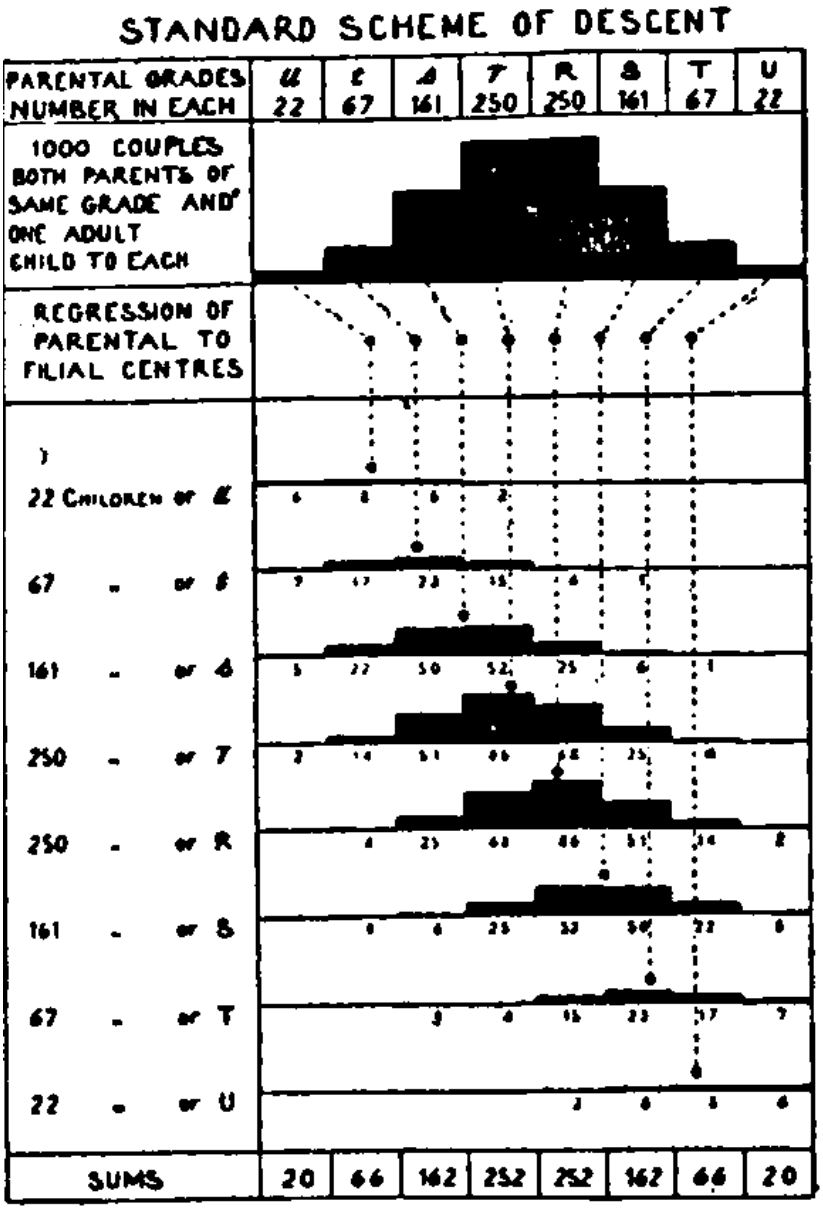
FOLLOW THE MONEY:

Private or government researchers and academic institutions, et. al. have financial interests and these should not be dismissed: positive research findings are directly proportional to the financial interests of the researchers in many cases. Also, researchers from academic institutions many times do research for career advancement and so there exists much pressure for positive results that are not actual, true findings. Even more insidious is the practice of esteemed primary investigators or researchers of using peer review in order to suppress results that are contradictory to their interests. Example: In 2009, the FDA did not approve polyheme, a hemoglobin-based oxygen carrier that could be transfused into trauma patients by EMS providers prior to arrival at the trauma facility. The authors of the polyheme study had underreported the risk of MI with polyheme and patients infused with polyheme had slightly higher death rates than those that had blood products. This study was financed and sponsored by Northfield Labs, who had potential of great financial gain if polyheme was proven to be an effective blood substitute for trauma patients. Although correct disclosures and regulatory processes were followed with this study, any study which will lead to tremendous profits by "someone" need to be examined carefully.

Hopefully this article will lead to a more scrupulous examination of any research: healthcare in general, EMS in particular, nutrition, pharmaceutical, or psychological. All research has the potential to one day be part of a Sleeper curve, and we can all think of other examples: hormone replacement therapy being more unsafe than reported, Vitamin C curing cancer, coffee causes breast cancer, faked research for Celebrex relieving knee pain after surgery, etc. etc. etc. The results of well conducted studies with good study design, data analysis and presentation may not lead to significant findings, but may contribute the body of research on a particular subject, and in that sense, all research can be good research if properly designed, reported, and used.

Appendix A

Sir Francis Galton's research on babies born to lower class parents CIRCA 1900-1911



The result is shown to the nearest whole per thousand in the diagram up to 'U and above.' It may be read either as applying to fathers and their sons when adult, or to mothers and their daughters when adult, or, again, to parentages and filial couplets. I will not now attempt to explain the details of the calculation to those to whom these methods

are new. Those who are familiar with them will easily understand the exact process from what follows. There are three points of reference in a scheme of descent which may be respectively named 'mid-parental,' 'genetic' and 'filial' centers. In the present case of both parents being alike, the position of the mid-parental center is identical with that of either parent separately. The position of the filial center is that from which the children disperse. The genetic center occupies the same position in the parental series that the filial center does in the filial series. 'Natural Inheritance' contains abundant proof, both observational and

theoretical, that the genetic center is not and cannot be identical with the parental center, but is always more mediocre, owing to the combination of ancestral influences—which are generally mediocre—with the purely parental ones. It also shows that the regression from the parental to the genetic center, in the case of stature at least, would amount to two thirds under the conditions we are now supposing. The regression is indicated in the diagram by converging lines which are directed towards the same point below, but are stopped at one third of the distance on the way to it. The contents of each parental class are supposed to be concentrated at the foot of the median axis of that class, this being the vertical line that divides its contents into equal parts. Its position is approximately, but not exactly, half-way between the divisions that bound it, and is as easily calculated for the extreme classes, which have no outer terminals, as for any of the others.

References

1. Hordig, M. Briese, T., Bueie, T., Bauman,M. et al Lack of Association between Measles Virus Vaccine and Autism with Enteropathy: A Case-Control Study PloS One 3 (9) 2008 http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0003140#pone.0003140-Barlow1

2. Ioannidis, J. Commentary: Grading the credibility of molecular evidence for complex diseases Int. J. Epidemiol. (June 2006) http://ije.oxfordjournals.org/content/35/3/572.full

3. Lieberman, M., Thou shalt not report odds ratios Language Log July 30, 2007 http://itre.cis.upenn.edu/~myl/languagelog/archives/004767.html

4. Meta-analysis of continuous data The Cochrane Collaboration 2002 http://www.cochrane-net.org/openlearning/html/modA1-4.htm

5. Songer, T. Measuring Injury Severity: A brief introduction University of Pittsburgh http://www.pitt.edu/~epi2670/severity/severity.pdf

6. Galton Essays; retrieved from http://galton.org/essays/1900-1911/galton-1902-pop-sci-improvement.pdf

7. Polyheme: Artificial Blood and Emergency Medicine Research. Indiana University Center for Bioethics 2010 http://bioethics.iu.edu/reference-center/polyheme/

8. Gialanella,D. Securities Fraud Action Alleging Pfizer Fudged Celebrex Data Gets Green Light

New Jersey law Journal May 15, 2012

http://www.law.com/jsp/nj/PubArticleNJ.jsp?id=1202553738717&slreturn=1

9. Ioannidis, J. Why Most Published Research Findings are False 2005 PloS Med 2(8)

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/pdf/pmed.0020124.pdf

10. Johnson,.S. Everything Bad Is Good for You: How Today's Popular Culture Is Actually Making Us Smarter Riverhead Trade May 2, 2006

11. Kirshner, I. , Deacon, B., Huedo-Medina,.T. , Scoboria,T. Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration PLoS Med 5(2): e45. doi:10.1371/journal.pmed.0050045

http://www.plosmedicine.org/article/citationList.action;jsessionid=B6AA6F98BD0D10AD01E68D3849CDEEC4?articleURI=info%3Adoi%2F10.1371%2Fjournal.pmed.0050045

12. Lehrer,J. The truth wears off: Is there something wrong with the scientific method?

The New Yorker December 13, 2010

http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer?currentPage=all

Special Thanks to Steven Berlin Johnson, the originator of the Sleeper Curve.