Reproducible Statistical Inference

James L Rosenberger

July 10, 2017

Outline

- Reproducibility
- Design of Experiments Principles
- Good statistical Practice
- Example of Statistical Practice
  - Court challenge of accuracy of BAC testing
  - Use of Inter-laboratory testing to establish accuracy

Reproducibility

Three basic concepts of reproducibility in research

1. Reproducible experiments

2. Reproducible analysis

3. Reproducible inference

Reproducibility

Three basic concepts of reproducibility in research

1. Reproducible experiments
   - Another lab can repeat your experiment and get similar results
2. Reproducible analysis
   - Another analyst (or yourself) can reanalyze your data and obtain the identical results
3. Reproducible inference
   - Another lab can reproduce your experiment and come to similar scientific conclusions

Publication Standards

- To accomplish reproducible experiments
  - Must provide clear a description of how your experiment was done
  - Requires a detailed protocol
    - Population sampled
    - Handling of materials
    - Measurement of response
  - Recording of data should be auditable from instrument to raw data file
    - Avoid Excel which allows changes without documentation

Reproducible Analysis

- Create a script of all steps from reading raw data to producing the report
  - Markdown in Rstudio
  - Sweave
  - SASweave
  - Minitab Journal
  - JMP script

Policy: NIH plans to enhance reproducibility

Francis S. Collins& Lawrence A. Tabak, Nature 505:612-613, (27 January 2014)

… a complex array of other factors seems to have contributed to the lack of reproducibility. Factors include poor training of researchers in experimental design; increased emphasis on making provocative statements rather than presenting technical details; and publications that do not report basic elements of experimental design. Crucial experimental design elements that are all too frequently ignored include blinding, randomization, replication, sample-size calculation and the effect of sex differences.

Design of Experiments

- Sources of Error
- Basis for Inference
  - Random Samples
  - Randomization
  - Nested sources of variation
  - Independence of observations
  - Blindness / Avoidance of Biases
- Generalizability

What is the Difference Between Repeatability and Reproducibility?

- https://www.labmate-online.com/news/news-and-views/5/breaking-news/what-is-the-difference-between-repeatability-and-reproducibility/30638
- What is Repeatability
  - Repeatability practices were introduced by scientists Bland and Altman. For repeatability to be established, the following conditions must be in place: the same location; the same measurement procedure; the same observer; the same measuring instrument, used under the same conditions; and repetition over a short period of time.

Reproducible inference

- What is Reproducibilty
  - Reproducibility, on the other hand, refers to the degree of agreement between the results of experiments conducted by different individuals, at different locations, with different instruments. Put simply, it measures our ability to replicate the findings of others.

Principles of DOE

- Randomization
  - Random selection
  - Random assignment
- Replication
  - SEM = SD/SQRT(n)
- Blocking
  - Local control
  - Stratification e.g. Gender, race, species

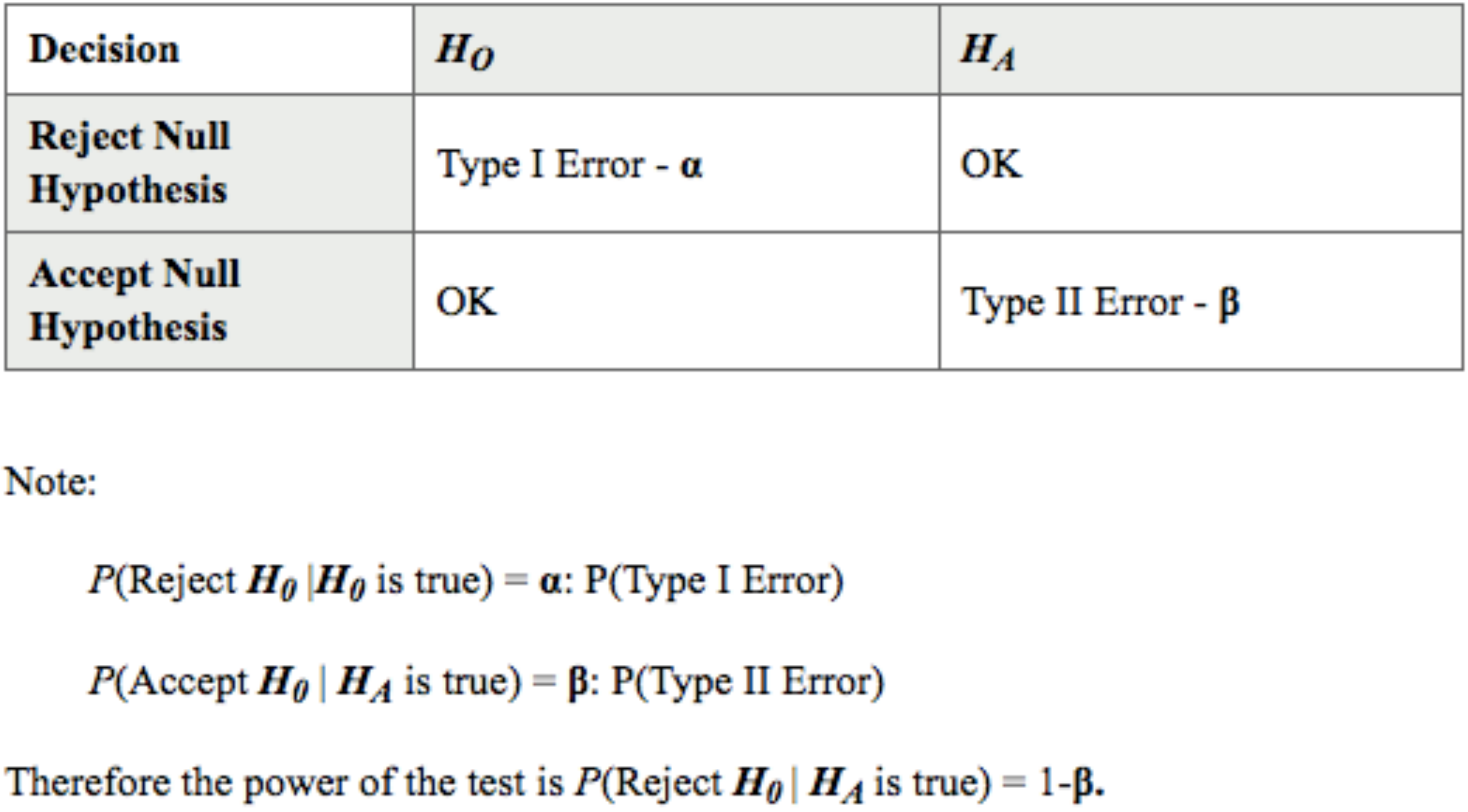
Additional principles

- Multifactor Designs
  - Use factorial structure to broaden the scope of inference and investigate interactions
  - More efficient than one-factor-at-a time designs
- Confounding
  - Can be a curse and provide efficiency
- Blindness
  - Those making measurements should not know the "preferred outcome"

Steps in conducting an experiment

- Recognition and statement of the problem
- Choice of factors, levels, and ranges
- Selection of the response variable(s)
- Choice of design
- Conducting the experiment
- Statistical analysis
- Drawing conclusions, and making recommendations

Determining Power



Decision	***H_O***	***H_A***
Reject Null Hypothesis	Type I Error - **α**	OK
Accept Null Hypothesis	OK	Type II Error - **β**

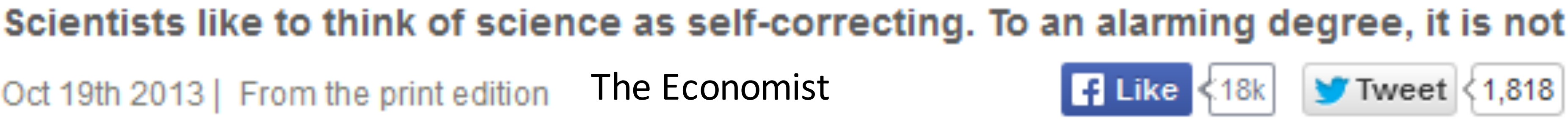
Note:

P(Reject $\boldsymbol{H_0}$ |$\boldsymbol{H_0}$ is true) = **α**: P(Type I Error)

P(Accept $\boldsymbol{H_0}$ | $\boldsymbol{H_A}$ is true) = **β**: P(Type II Error)

Therefore the power of the test is P(Reject $\boldsymbol{H_0}$ | $\boldsymbol{H_A}$ is true) = 1-**β.**

Reproducible Research?

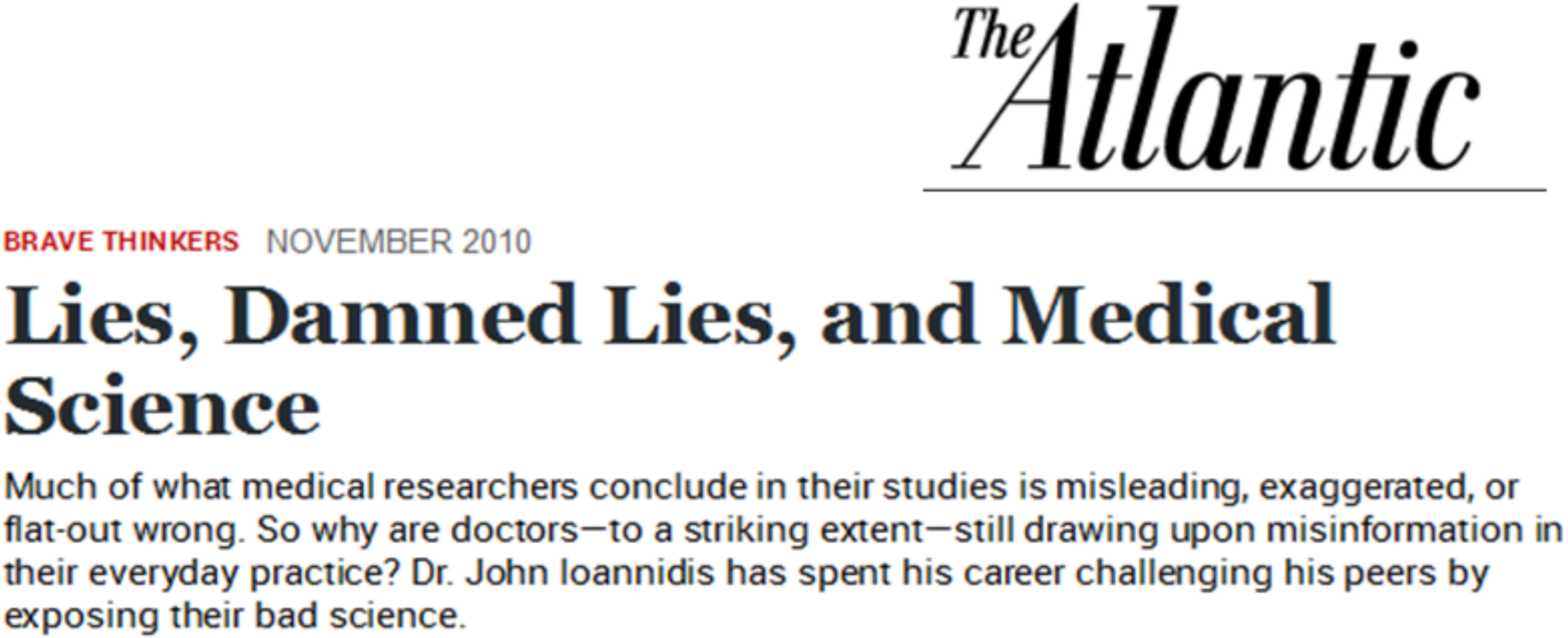


Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

The Economist

Can we count on science to produce reliable results?



The Atlantic

BRAVE THINKERS NOVEMBER 2010

Lies, Damned Lies, and Medical Science

Much of what medical researchers conclude in their studies is misleading, exaggerated, or flat-out wrong. So why are doctors—to a striking extent—still drawing upon misinformation in their everyday practice? Dr. John Ioannidis has spent his career challenging his peers by exposing their bad science.

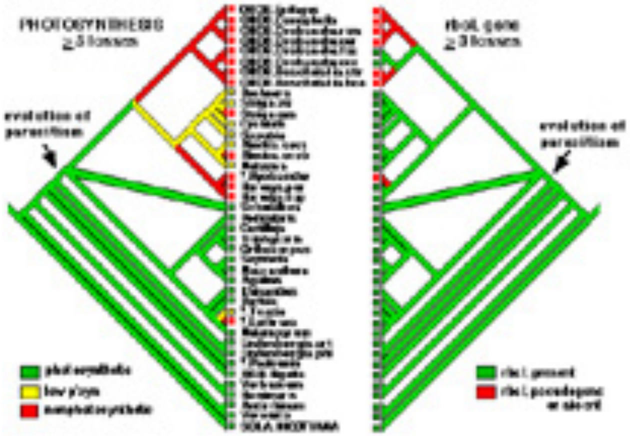
Two aspects of reproducibility

Reproducibility in the lab

data measurements

Paulson lab

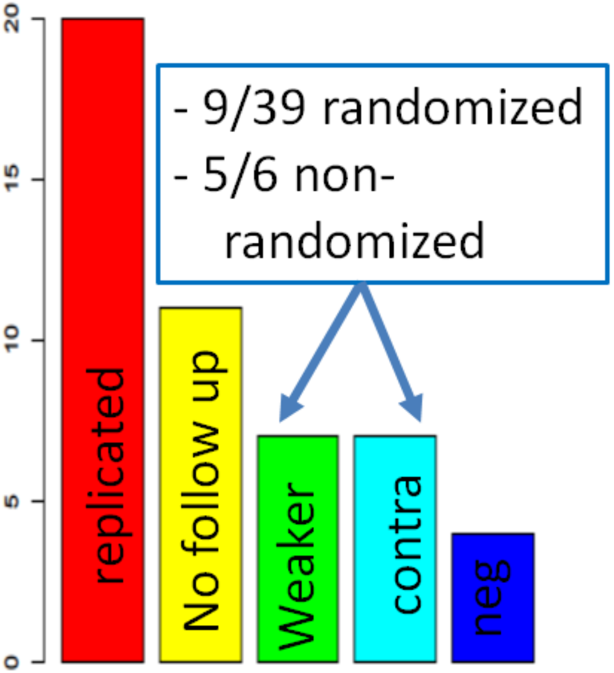
Reproducibility of the bioinformatics analysis



From dePamphilis lab

Reproducibility in the Lab

- Ioannidis 2005 *JAMA*: “Contradicted and initially stronger effects in highly cited clinical research.”
- Ioannidis 2005 *PLoS Medicine* “Why Most Published Research Findings are False”



- 9/39 randomized
- 5/6 non-randomized

Reproducibility in the Lab

Ioannidis' arguments:

- Focusing on rare events – most discoveries are false
- Publication bias: Only "interesting" events are published.
- Detection bias: Selective or distorted reporting, conflicts of interest, deliberate manipulation
- Lack of independent replication
- Selection of most significant events instead of proper meta-analysis when there is replication

Reproducibility in the Lab

High false positive rates should be expected when:

- Low power
- Small studies: if the false positive rate is controlled, the false negative rate is high
- Small effect sizes:
- Large number of relationships tested without preliminary findings: lower prior probability of effect
- High flexibility in designs, definitions, outcomes and analyzes: the search for significance
- **Rewards of research: winner takes all, so it pays to be first and to find something**
- Hot areas of research: research is rushed, lots of studies with low prior probability of effect

Altman, 11/15

Reproducibility in the Lab

Good experimental designs for obtaining reliable data:

1. Determine the population to which results are to be applied.
2. Be sure the experimental units are a random sample from this population.
3. Design measurements to avoid biases.

Altman, 11/15

Reproducibility in the Lab

Good habits for obtaining reliable data :

1. Follow lab procedures and protocols.
2. Document everything including failures and irregularities.
3. Keep a good lab book.
4. Every number should be auditable.

Reproducibility in the Lab

Reproducible ≠ Correct

We can ALWAYS get reproducible results if we:

- Do biased experiments
- Always accept (reject) our hypothesis

Correct is more important!!

Reproducibility in the Lab

Statistical testing:

We do a test (e.g. t-test) and reject if $P<\alpha$

(Reject means that we declare a "discovery")

Usually we pick $\alpha=0.05$ or 0.01

What is the probability that we have **actually** discovered something if we reject?

Reproducibility in the Lab

Statistical testing:

Prob(correct|P<α)=

↑

Power

Prob Null is true	α	Power	Prob(correct|P<α)	Prob(correct|P>α)	Prob(reproducible |Correct reject)
95%	0.05	0.80	46%	99%	64%
95%	0.01	0.47	71%	97%	22%
50%	0.05	0.80	94%	83%	64%
50%	0.01	0.47	98%	68%	22%

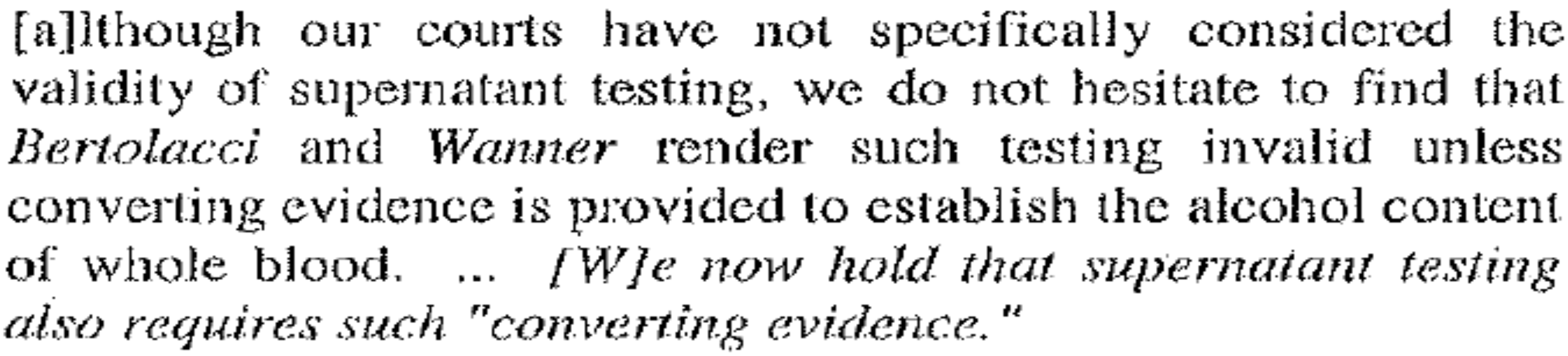
A Story

- A court case challenged the accuracy of routine clinical lab measurements of BAC
- Claimed that although measurements were correlated, they were not accurate
  - Indeed correlated measurements does not establish equivalence between two methods
- Analysis of an inter-laboratory study showed that the WB method routinely used, was accurate compared to the GC method, considered the gold standard

Com v Daughenbaugh et al

- Brief describing the proceedings…

Example of establishing accuracy

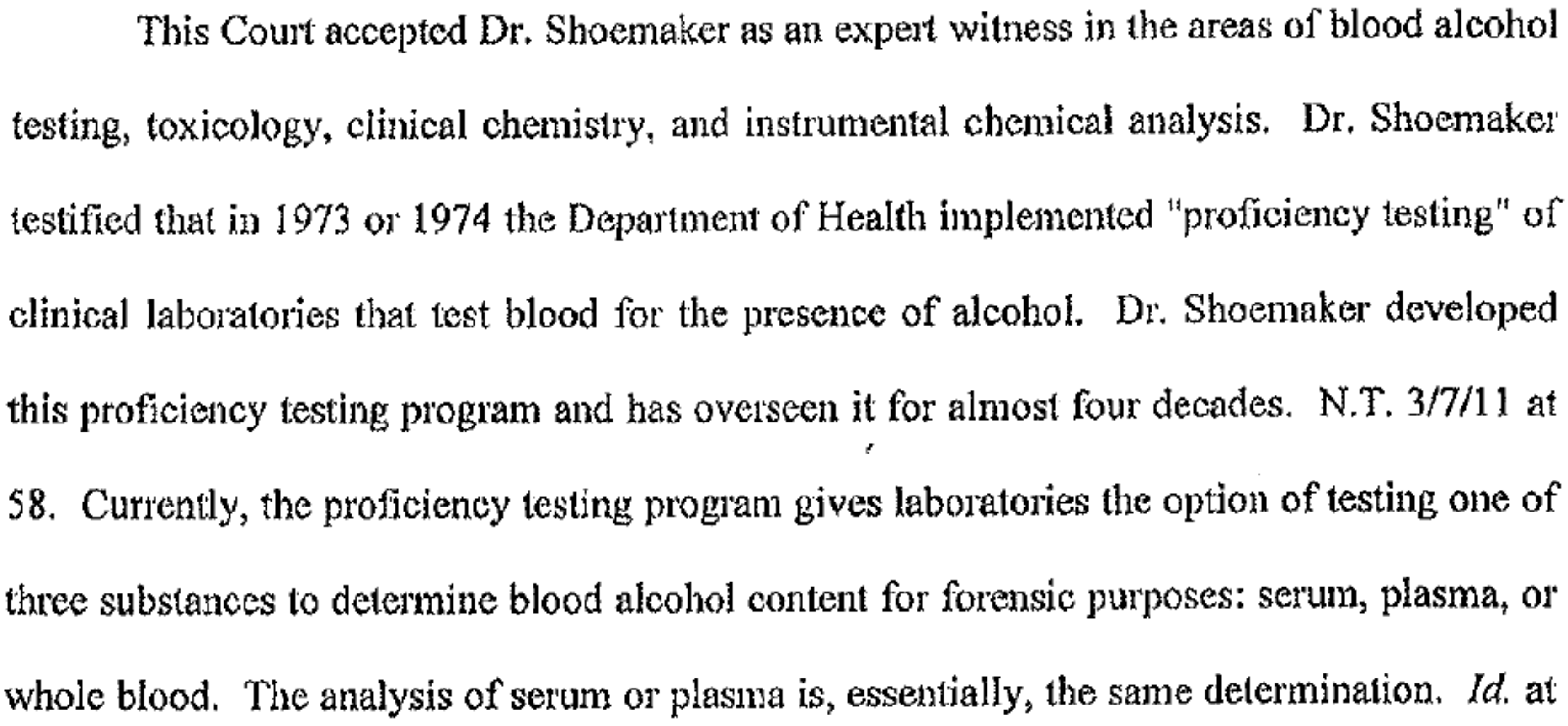


[a]lthough our courts have not specifically considered the validity of supernatant testing, we do not hesitate to find that *Bertolacci* and *Wanner* render such testing invalid unless converting evidence is provided to establish the alcohol content of whole blood. ... *[W]e now hold that supernatant testing also requires such "converting evidence."*

This was the challenge to the accuracy of the local laboratory testing procedure.

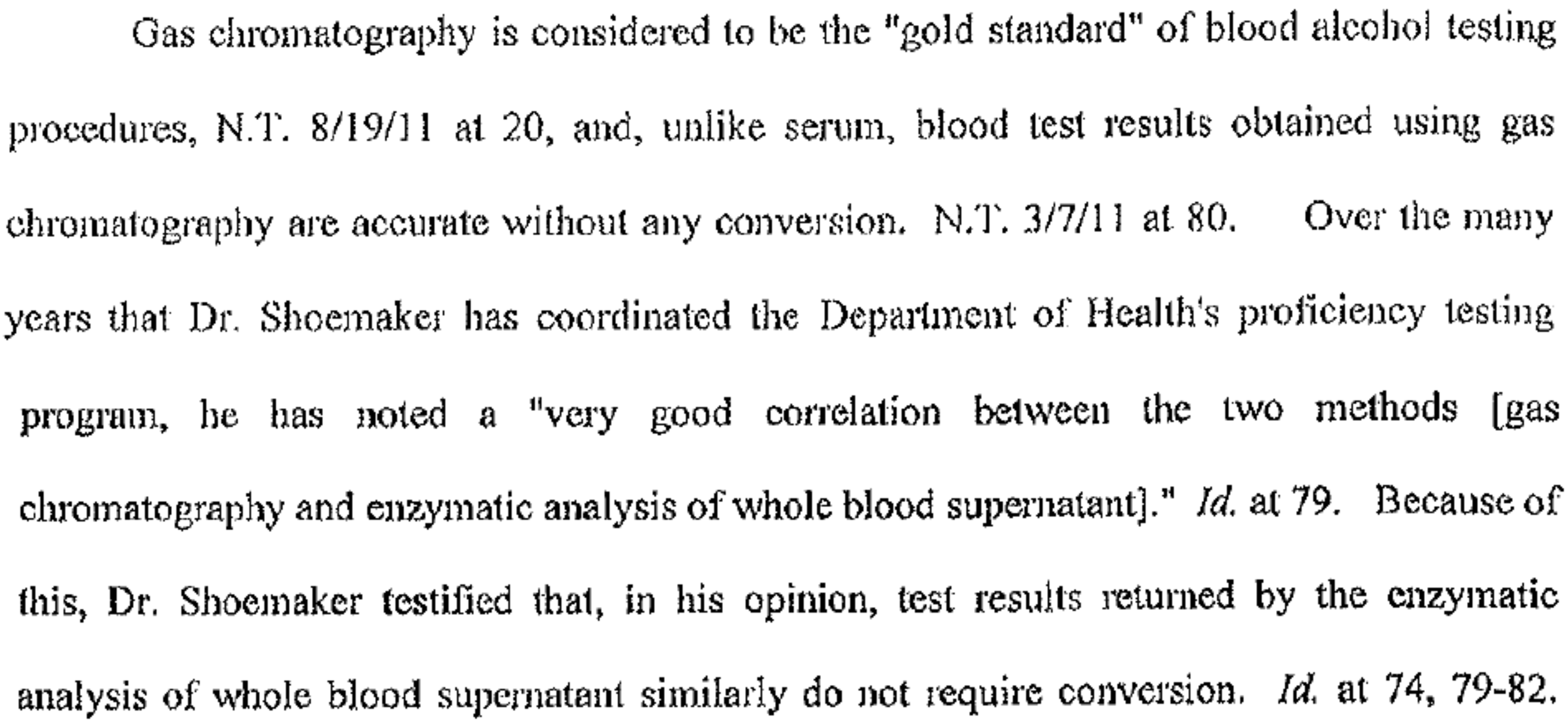
The defendant dredged up an old ruling that suggested that the laboratory procedure was flawed and required a conversion factor.

Testing of Clinical Laboratories



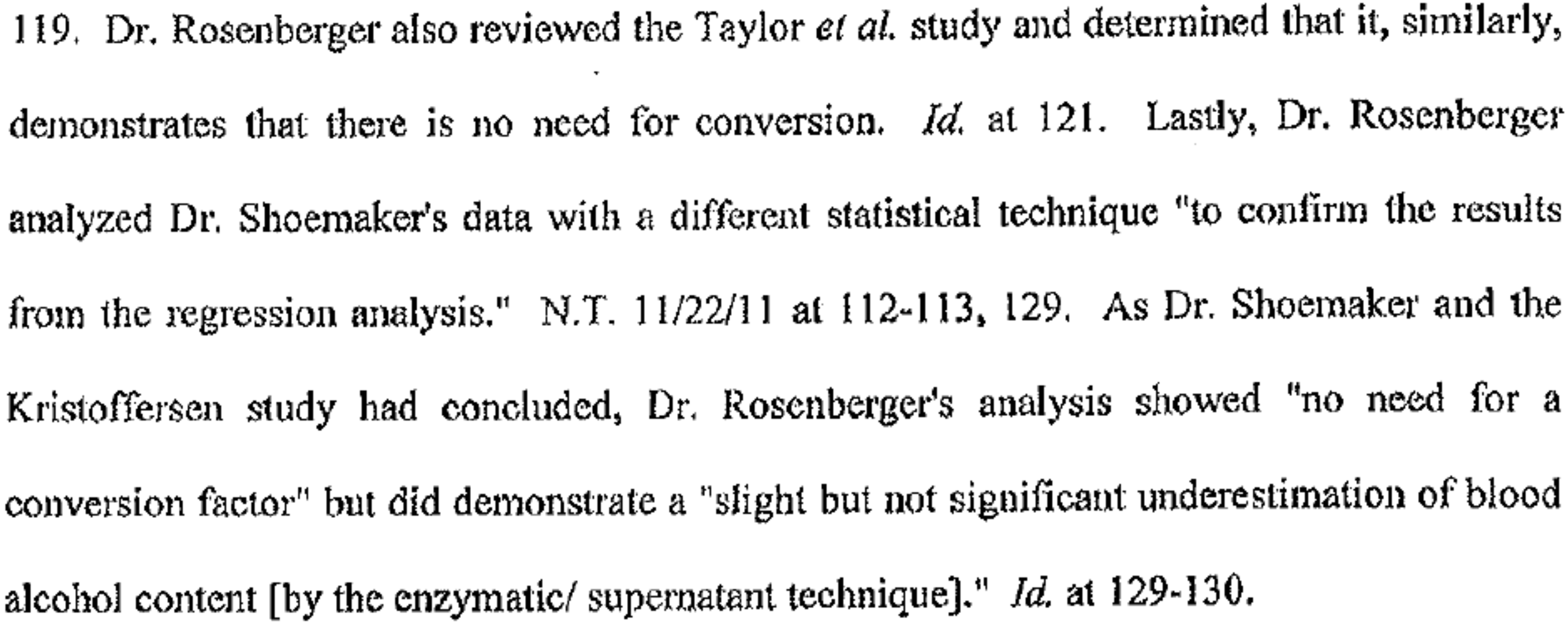
This Court accepted Dr. Shoemaker as an expert witness in the areas of blood alcohol testing, toxicology, clinical chemistry, and instrumental chemical analysis. Dr. Shoemaker testified that in 1973 or 1974 the Department of Health implemented "proficiency testing" of clinical laboratories that test blood for the presence of alcohol. Dr. Shoemaker developed this proficiency testing program and has overseen it for almost four decades. N.T. 3/7/11 at 58. Currently, the proficiency testing program gives laboratories the option of testing one of three substances to determine blood alcohol content for forensic purposes: serum, plasma, or whole blood. The analysis of serum or plasma is, essentially, the same determination. *Id.* at

Gold Standard



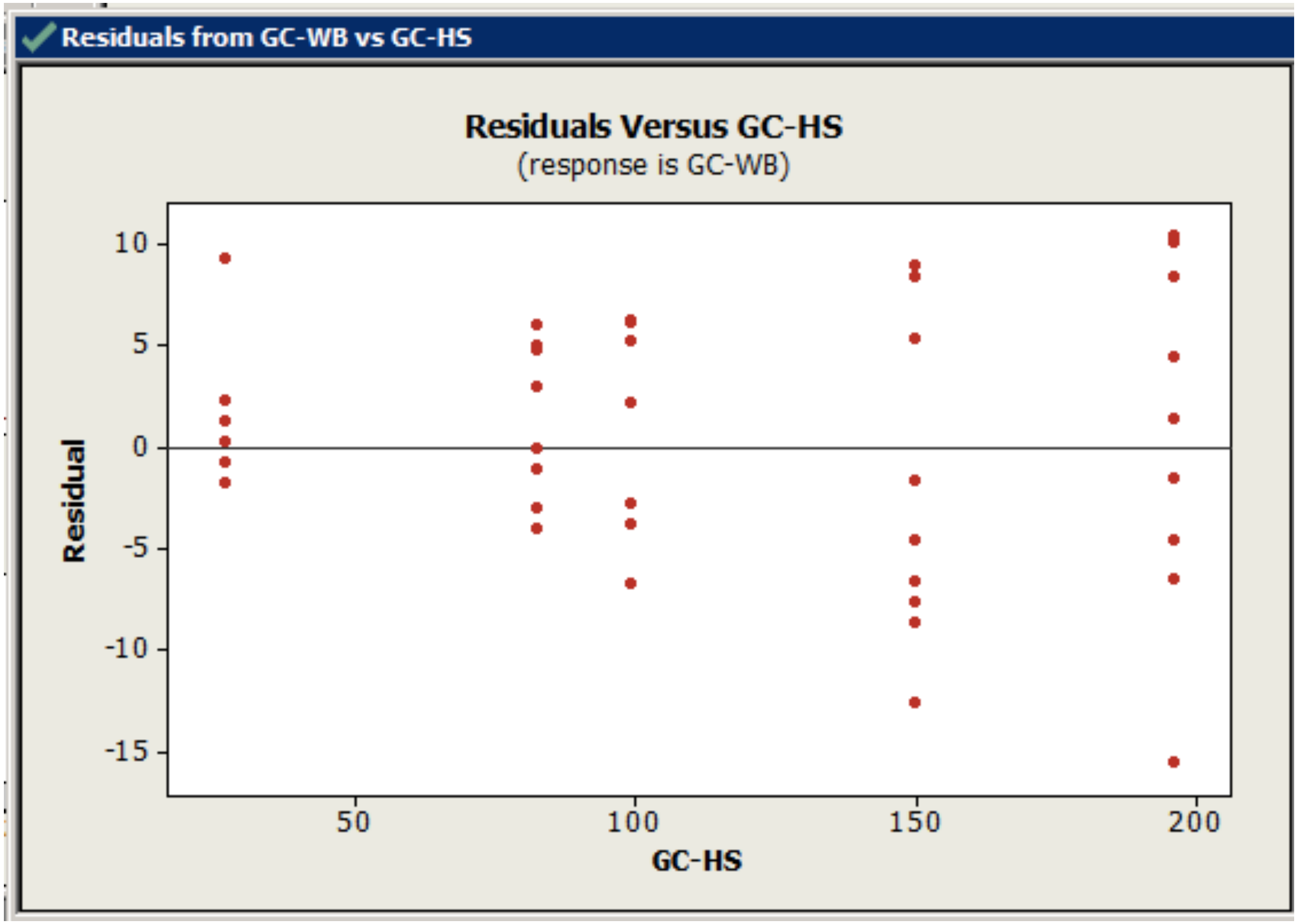
Gas chromatography is considered to be the "gold standard" of blood alcohol testing procedures, N.T. 8/19/11 at 20, and, unlike serum, blood test results obtained using gas chromatography are accurate without any conversion. N.T. 3/7/11 at 80. Over the many years that Dr. Shoemaker has coordinated the Department of Health's proficiency testing program, he has noted a "very good correlation between the two methods [gas chromatography and enzymatic analysis of whole blood supernatant]." *Id.* at 79. Because of this, Dr. Shoemaker testified that, in his opinion, test results returned by the enzymatic analysis of whole blood supernatant similarly do not require conversion. *Id.* at 74, 79-82.

Conclusion: No need for a conversion factor to use plasma methods

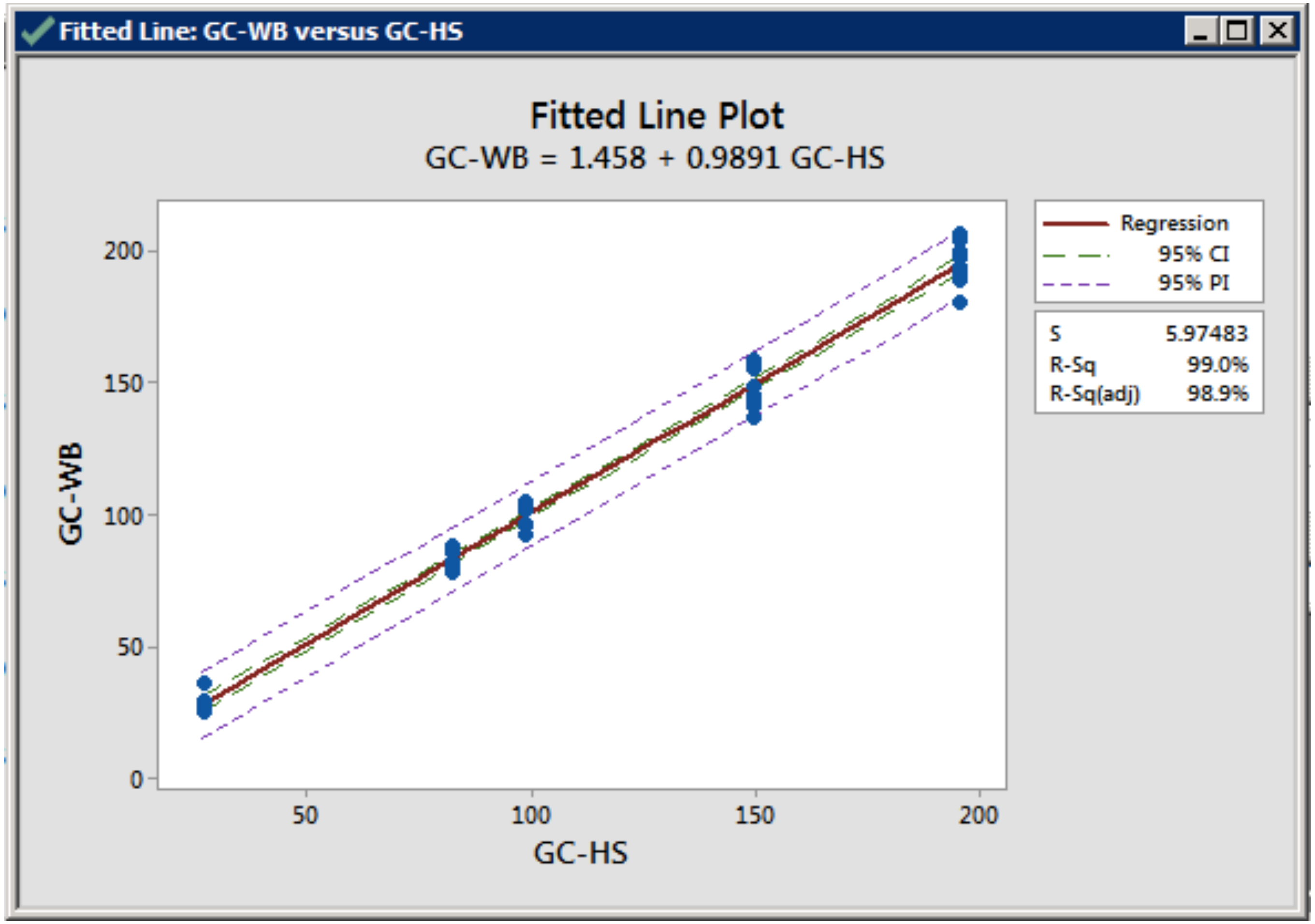


119. Dr. Rosenberger also reviewed the Taylor *et al.* study and determined that it, similarly, demonstrates that there is no need for conversion. *Id.* at 121. Lastly, Dr. Rosenberger analyzed Dr. Shoemaker's data with a different statistical technique "to confirm the results from the regression analysis." N.T. 11/22/11 at 112-113, 129. As Dr. Shoemaker and the Kristoffersen study had concluded, Dr. Rosenberger's analysis showed "no need for a conversion factor" but did demonstrate a "slight but not significant underestimation of blood alcohol content [by the enzymatic/ supernatant technique]." *Id.* at 129-130.

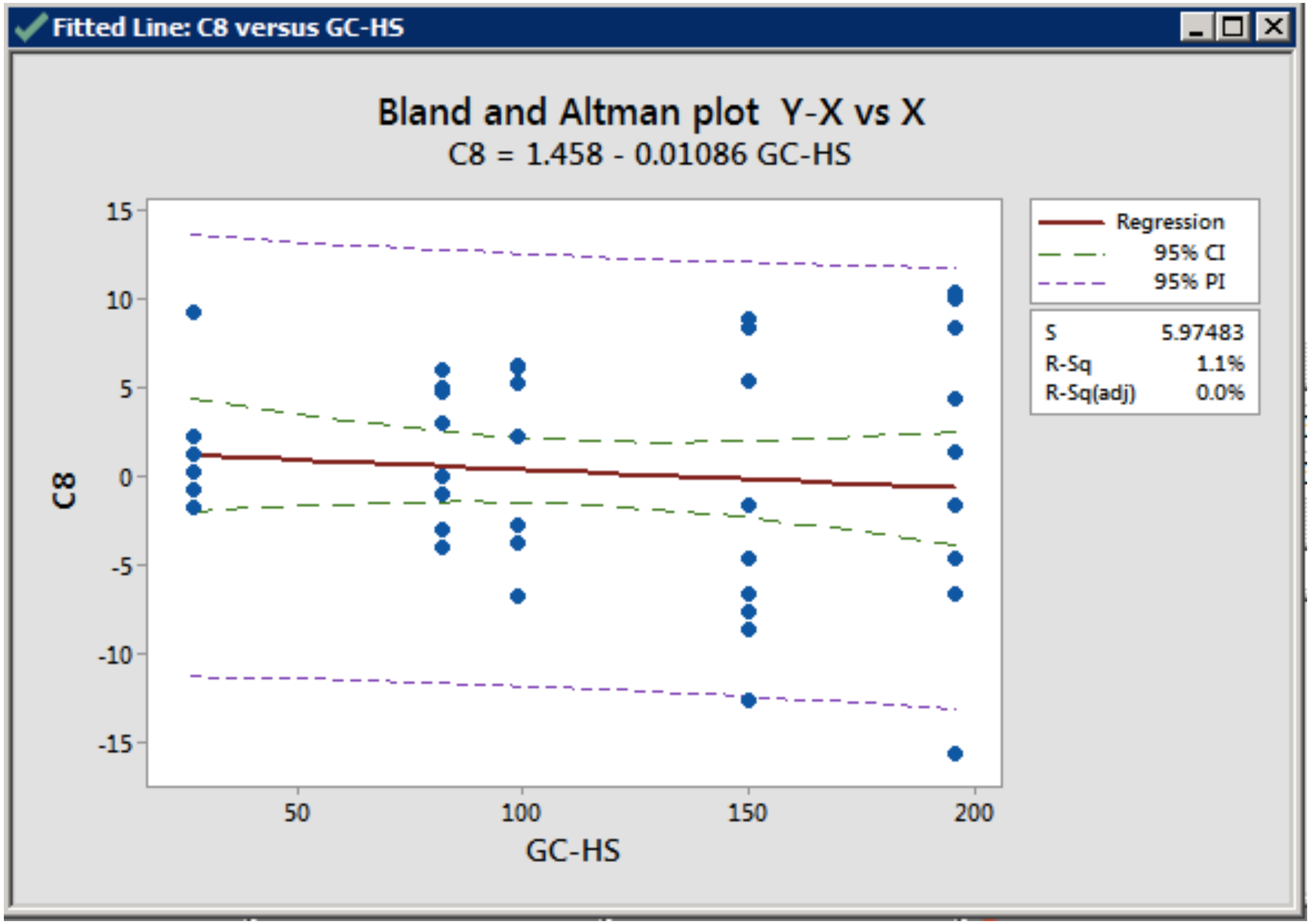
Results from Bland and Altman



Regression of Y on X



Expanded view of regression result



Fitted Line: C8 versus GC-HS

Bland and Altman plot Y-X vs X

C8 = 1.458 - 0.01086 GC-HS

S	5.97483
R-Sq	1.1%
R-Sq(adj)	0.0%