METHODS, RESULTS AND DISCUSSION

Angela Sauaia, MD, PhD

September, 2013

San Francisco, California, 2013 AAST
Disclosure

• Dr. Sauaia works for the Journal of Trauma and Acute Care Surgery
• Nothing else to disclose
Why should surgeons care about clinical research methodology?


Dept. of Surgery, Johns Hopkins School of Medicine

**BACKGROUND:** A low prevalence of high-level clinical studies in the surgical literature has been reported previously. We reviewed a recent sample of surgical publications to assess the current status of clinical research.

**STUDY DESIGN:** A 3-month sample of journal articles in Archives of Surgery, Surgery, and Annals of Surgery in 2005 was evaluated by two independent reviewers to determine the distribution of articles in established evidence classes.

**RESULTS:** A total of 133 publications were identified in the three journals during the time periods reviewed, including 101 clinical articles and 30 basic science articles. Among the clinical papers, there were 8 class I studies (7.9%), 34 class II studies (33.7%), and more than half were class III studies (59 of 101, or 58.4%).

**CONCLUSIONS:** The low prevalence of high-level evidence to guide surgical management of patients persists in major general surgery journals. We believe that education about proper research methodology is not only important for researchers, but is also important for practicing surgeons, and can have important health policy implications as well.
A competent surgeon reads the literature, but not all of it. Be discerning. For decisions about therapy for an individual patient, focus on RCTs; evidence-based, systematic reviews; and EBM guidelines. Using professional wisdom, apply what you read, understanding the limitations and applicability to your unique patient.

Ronald V. Maier, MD (Arch Surg. 2006)
Another reason: better chances of publishing your study in the Journal

Reviewer 1

Reviewer 2

ASSESSING THE MANUSCRIPT

In an effort to standardize the review process for the *Journal of Trauma and Acute Care Surgery*, we ask that you consider the following questions when assessing a manuscript for possible publication:

- Why was the study done? Does it address either an important unsolved problem of clinical relevance or a basic scientific topic relevant to trauma and acute care surgery? Do you think that there is sufficient evidence to justify the study? Have the authors explicitly stated a study purpose or a hypothesis?

- How was the study done? What is the design and is it explicitly stated by the authors in the methods?

- Is the study population defined well? Do the authors explicitly define inclusion and exclusion criteria? Are all of the patients accounted for in the results section?

- Are the outcome measures appropriate? Are the selected variables suitable to the study purpose or hypothesis? Are confounding variables assessed?

- Are the analytical methods (e.g. statistical analyses, laboratory diagnostics) appropriate? Is there hypothesis testing? Was a power analysis done?

- What is the significance of the work? Are the results compared with previous similar work? Are potential study limitations addressed?

- Are the conclusions warranted by the data?
Checklist for statistical assessment of general papers

<table>
<thead>
<tr>
<th>Item</th>
<th>YES</th>
<th>Comments (at the end of document by number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriate study design used to achieve the objective(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source of subjects/data appropriately described</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling/sample size appropriately described</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entry and exclusion criteria clearly defined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data exclusions are stated/explained and impact on results are explored</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N reported at the start of the study, for each data set and for each analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrepancies in value of N between analyses clearly explained/ justified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing data are explained, and impact on findings minimized/ explained</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfactory follow-up/ response rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequate uni/bivariate statistical analyses used/described</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequately multivariate statistical analyses used/described</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confounding and bias explored and minimized</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assumptions of tests applied met (particular attention paid to non-normal data sets or small sample sizes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjustments made for multiple testing explained</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unit of analysis given for all comparisons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alpha level given for all statistical tests</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tests clearly identified as one or two-tailed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual P values are given for primary analyses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate measure(s) of center (e.g. mean or median)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate measure(s) of variability (e.g. standard deviation or range)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unusual/complex statistical methods clearly explained for JT’s wide readership</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method of treatment assignment (randomization etc) explained and justified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any data transformations clearly described and justified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confidence intervals given for the main results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conclusion drawn from the statistical analysis is justified</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adapted from
• BMJ statistical checklist at [http://resources.bmj.com/bmi/authors/checklists-forms/statisticians-checklist](http://resources.bmj.com/bmi/authors/checklists-forms/statisticians-checklist) (accessed 10/08/11)
• Nature statistical adequacy checklists [www.nature.com/nature/authors/gta](http://www.nature.com/nature/authors/gta) and [www.nature.com/ncomms/.../Checklist_of_statistical_adequacy.doc](http://www.nature.com/ncomms/.../Checklist_of_statistical_adequacy.doc) (accessed 10/08/11)
Purpose of the J Trauma Std. Methods & Statistical Review

- To assist editor and reviewers in reviews
- To ensure the high methodological quality of the Journal’s publications
- To provide authors with information to address concerns regarding methods/stats
- To adjudicate levels of evidence
- To identify common statistical issues in submissions and address them through Podcasts, opinion articles, AAST meetings
Reporting Standards: Highly Encouraged by the Journal

- Each reporting standard has a checklist
- Include the checklist with your article submission
Methods Section

1. Study design
2. Sample
3. Study groups description
4. Variables
5. Biases
6. Statistical analysis:
   – Descriptive
   – Inferential
Study design

• **RCT:** groups defined at random minimizing bias

• **Cohort:** groups defined by exposure, outcomes compared
  – **Prospective:** hypothesize first, collect data later
    • What about Trauma Registries, NTDB, etc.?
  – **Retrospective:** collect data first, hypothesize later

• **Case-control:** groups defined by outcome, risk factors compared
  – what if we compare risk factors of survivors vs. non-survivors
    Answer: study is a cohort if both survivors/non-survivors were
    consecutive patients with a common risk factor (e.g., trauma).

• **Case series:** no comparator group, rare events or conditions,
  feasibility of new interventions, first evidence of innovations

• **Systematic Review/Meta-analysis/ Guidelines**
Methods Section

1. Study design
2. Sample
3. Study groups description
4. Variables
5. Biases
6. Statistical analysis:
   – Descriptive
   – Inferential
The sample

- **Purpose**: allows reader to know what sample looks like and compare to his/her population (clinical/research)

- **Setting, time and geography**
  - *All in-hospital deaths from 2000-2011, S. Diego Scripps Mercy*

- **Source of subjects/data**
  - *Scripps Mercy Trauma Registry + Medical Examiner + MAC*

- **Participants:**
  - **Entry and exclusion criteria**
    - *admission without trauma consultation were excluded; Results: 23 (2.7%) not managed by trauma service excluded*
  - **Enrollment process, refusal rates, reasons for refusal**
  - **Length and Methods for follow-up, Loss to follow-up**
Methods Section

1. Study design
2. Sample
3. Study groups description
4. Variables
5. Biases
6. Statistical analysis:
   – Descriptive
   – Inferential
Study groups description

• RCT: randomization process, blinding
• >=2 study groups in cohort, case-control, cross-sectional
  – Precise description of how study and comparator groups were defined:
    • Elderly: age>65 years
    • Problems: when interventions define groups (e.g., received massive transfusion: yes or no)
  – When matched: describe matching criteria (e.g., propensity scores)
Methods Section

1. Study design
2. Sample
3. Study groups description
4. Variables
5. Biases
6. Statistical analysis:
   – Descriptive
   – Inferential
Variables

• Precise definitions that allow reader to reproduce them
• Readers need to be able to reproduce the measurement
  – Details on the measurements:
    – ...time from arrival to drawing the specimen in the trauma bay was 4-18min
    Worst daily values of each variable are defined in SDC-Table 1
• Types of variables:
  – Outcome (s)
  – Effect variables (the variable(s) of interest)
  – Covariates (confounders, effect modifiers)
• If non-objective variable: use scores, set of criteria
  – Patient comorbidities ... were evaluated by Charlson Comorbidity Index\textsuperscript{17}...
  – MOF was measured using the Denver MOF score
  – The committee definition of exsanguination included all deaths in the operating room (OR) due to uncontrollable hemorrhage.”
Methods Section

1. Study design
2. Sample
3. Study groups description
4. Variables
5. Biases
6. Statistical analysis:
   – Descriptive
   – Inferential
Biases

- **Selection bias:** patients/data excluded differ from included patients
- **Survivor bias (special type of selection bias):**
  - only patients who survived long enough to receive the intervention are included;
  - use of censoring diminishes it but does not eliminate it
  - Example: early studies of 1:1 RBC:FFP ratio
- **Intervention bias:** “at the discretion of the attending physician”
  - Solutions: SOPs, classify patients based on exposures not interventions
Biases: Missing data

• Missing data: explore and report!! why are data missing?
  – Missing at Random (MAR): may not need specific treatment
  – Missing Not at Random (MNAR): Informative!!

• Satisfactory follow-up/ response rate
Methods Section

1. Study design
2. Sample
3. Study groups description
4. Variables
5. Biases
6. Statistical analysis:
   - Descriptive
   - Inferential
Descriptive statistics

• Measures of central tendency and dispersion: they must allow the reader to “see” the sample
  – Categorical and ordinal variables: N, %
  – Continuous
    • Normally distributed (e.g., SBP, HR)
      – Mean and SD/ SEM
    • Not normally distributed (e.g., RBC units, LOS)
      – Median and Interquartile Range (25%/75%)
      – Median and Range
      – Categorize using meaningful cutoffs: ISS>=25
      – Transformations: e.g., log of LOS
Mean (SD) 9 (10)
Median (IQR) 5 (2.5-10)

Mean (SD) 43 (19)
Median (IQR) 42 (27-56)

Mean (SD) 13 (12)
Median (IQR) 9 (4-17)
Inferential Statistical analysis

DID RESEARCH

GOT SIGNIFICANT RESULTS
Inferential Statistical analysis

• Study results versus Truth

• **Significance:** controls the Type I error or “the chance of finding a difference in the sample when there is NOT a difference in the universe from which this sample was taken. ”
  
  – Usually <0.05 (thanks Dr. Fisher 😞)
  
  – If multiple comparisons, it must be adjusted to a lower value through post-hoc tests, FDR q-value, etc.
  
  – Statistical difference versus Clinical difference

“Give me a large enough sample size and I will give you a p-value.”
Inferential Statistical analysis

• Study results versus Truth
• **Statistical power**: controls the Type II error or “the chance of NOT finding a difference in the sample when there is a difference in the universe from which this sample was taken.”
• If you find a significant difference, a discussion on power is IRRELEVANT
• Setting the statistical power:
  – The OLD ways: <0.80
  – The New ways: it depends
Negative studies and Statistical Power or What happens when P>0.05

...and this is where we put the non-significant results.
Negative studies and Statistical Power or What happens when P>0.05

- Hypothesis: there is a difference
- Interpretation: There may/may not be a difference; we failed to find a difference
- Statistical power can be low

- Hypothesis: there is NO difference (Equivalence/Non-inferiority trials)
- Interpretation: the treatments are equivalent
- Statistical power must be high
Medicine is the science of temporary truth
Naim Sauaia, MD, PhD 1928-2001

• It takes a lot of good quality and convergent scientific evidence before we can be reasonably sure of a “scientific fact”

• The “p-hacking”: maneuver by which a researcher might attempt to modify their data or analyses until they reach the revered $p = .05$ level.

(Simmons, Nelson, Simonsohn, 2012)
Just because saying it once may never be enough

<table>
<thead>
<tr>
<th>Result of statistical test</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Null hypothesis false</td>
</tr>
<tr>
<td>Statistically significant</td>
<td>True positive</td>
</tr>
<tr>
<td>Not statistically significant</td>
<td>False negative Type II error (β)</td>
</tr>
</tbody>
</table>
Inferential Analysis: Uni/Bivariate

• Fundamental questions:
  – Type of outcome variables: continuous or categorical?
  – Type of distribution: normal or not-normal/UNK?
  – Independent or paired/repeated?
  – Censoring (e.g., subjects did not live long enough to experience intervention or outcome in question)?
  – Time to event analysis?

• Three additional important questions:
  – Multiple comparisons?
  – Cluster effects: observations from one practice/center are more alike than from two different centers?
  – Are the variances the same or different?
Inferential Analysis: Uni/Bivariate

- Two Unpaired Samples
  - Continuous Outcome Measure
    - Normally Distributed (parametric)
      - Student t-test
      - Wilcoxon rank-sum (Mann-Whitney U)
  - Dichotomous Outcome Measure
    - Small Sample Size
      - Fisher’s Exact test
      - Chi-square test (Pearson)
  - Ordinal Outcome Measure
    - Wilcoxon rank-sum (Mann-Whitney U)
  - Nominal Outcome Measure
    - Fisher’s Exact test

- Two Paired Samples
  - Continuous Outcome Measure
    - Normally Distributed (parametric)
      - Paired t-test
      - Wilcoxon signed-rank test
  - Dichotomous Outcome Measure
    - McNemar’s Test
  - Ordinal Outcome Measure
    - Wilcoxon signed-rank test
  - Nominal Outcome Measure
    - McNemar’s Test
Inferential Analysis: Uni/Bivariate

Correlation between Two Variables

- Continuous Outcome Measure
  - Normally Distributed (parametric)
    - Pearson’s correlation
    - Spearman rank correlation
  - no Normally Distributed (parametric)
    - Kendall’s coefficient of concordance
- Ordinal Outcome Measure
  - Kappa statistic
Predicting vs. Risk-adjusting

**Predicting**

Goal: To develop a TOOL to predict health outcomes

Best predictors may be chosen using stats software

Can NOT study cause-effect

**Risk-adjusting**

Goal: To determine the INDEPENDENT EFFECT of a variable, adjusted for confounders

Confounders must be chosen by CLINICIAN

Can study cause-effect
- Confounding
- Interaction or effect modification
- Model performance measures
Multivariate Analysis: assumptions, assumptions, assumptions

- Linear relationship?
- If linear, what kind of line?
Multivariate analysis:

- Testing for an interaction or effect modifier: a variable that changes the direction of the effect of another variable
A final word of caution
Odds ratios and Relative risks

• Logistic regression: gives you odds ratios
• Odds ratios are reasonable estimates of relative risks when the incidence is 20% or below. Anything higher than this and odds ratios are poor estimates of relative risk
• Poisson regression (used in the Death trends article) gives you relative risks
• Logistic regression is NOT the only technique for multivariate analysis
<table>
<thead>
<tr>
<th>Performance component</th>
<th>Definition</th>
<th>Examples of Measure(s)</th>
</tr>
</thead>
</table>
| Overall               | Measure of how well curve fits the data | • R-square  
|                       |            | • AIC                  |
| Discrimination        | Ability of distinguishing (discriminates) individuals with and without the outcome | • C-statistic  
|                       | Great discrimination: All with probability > 20% died | • ROC curves |
| Calibration           | Ability to correctly estimate the probability of a given event across the whole range of prognostic estimates | • Hosmer-Lemeshow (HL) statistic |
|                       | Great calibration: predicted 20% mortality, observed 20% mortality | |
| Validation            | Ability to produce same accuracy in different samples | • Validation datasets  
|                       |            | • Simulations |
Statistics for Selecting Variables

• Principal components analysis: a type of factor analysis to select groups of variables that are correlated amongst themselves but not correlated with the variables in another group

• Once components are identified, clinicians can look at the composition of the component and make inferences about possible mechanisms that “unify” those variables
Systematic Reviews and Meta-analysis

• Require specific review protocols
• Meta-analysis: combines results of individual studies, not always indicated
  “Never meta-analysis I like.”
• Sensitivity analyses: repeating the analyses under different assumptions
• Testing for heterogeneity and publications bias
• The best source: http://handbook.cochrane.org/
Guidelines

• GRADE: Grading of Recommendations Assessment, Development and Evaluation.

• http://www.gradeworkinggroup.org/
Results

1. Participants
   – Flow diagram

2. Descriptive data

3. Inferential data
   – Bivariate comparison of the two groups
   – If this RCT, this will demonstrate that randomization was effective in avoiding a differential distribution of confounders and will examine the difference in outcomes
   – If not RCT, this may be done in SDC or text
   – Multivariate if possible confounding or effect modification

4. Other analyses
Effect of video laryngoscopy on trauma patient survival: A randomized controlled trial

Dale J. Yeatts, MD, Richard P. Dutton, MD, MBA, Peter F. Hu, MS, Yu-Wei W. Chang, MS, Clayton H. Brown, PhD, Hegang Chen, PhD, Thomas E. Griswol, MD, Joseph A. Kufera, MA, and Thomas M. Scalea, MD, Baltimore, Maryland

Figure 2. Patient flow diagram.
The anatomy of an article: The discussion section

“How does the article I read today change what I will recommend to my patients tomorrow?”

Angela Sauxa, MD, PhD, Ernest E. Moore, MD, Jennifer Creels, Ronald Maier, MD, David B. Hoyt, MD, and Steven R. Shuckford, MD, Denver, Colorado

“Medicine is the science of temporary truths.”
—Naim Sauxa, MD, PhD, 1928–2001

Exactly 30 years ago, Sauxa and Sauxa1–3 (the first author and her father) published a series of articles on “editing a scientific article.” The idea was to assist the reader to incorporate the knowledge brought by an article into what he or she already

This article, the first of a series on the “anatomy of an article,” concentrates on the Discussion section, about which little has been written in terms of standardization. We believe that the Discussion is crucial to optimally assist the reader to translate the results of the article to his or her clinical practice or research. Reporting standards for the Discussion section currently exist (Table 1), but they give little direction to what should be included. Indeed, these recommendations for the Discussion section are much less explicit than the recommendations for the Methods and Results sections, which specifically guide
Discussion

1. Re-state purpose and give three main findings

2. Study in the context of current evidence: “How did the study findings fill the existing gap identified in the Introduction?”

3. Evidence level of the individual study (ELIS).
   – Type of study: (therapy, epi, diagnostic, economic, systematic review/meta-analysis or guideline)
   – Study design: prospective vs. retrospective
   – Strengths/limitations

4. Application: use PICO framework (patients, interventions, comparators, outcomes)

5. Next steps

6. Conclusions. one paragraph conclusion or take-home message supported by the current study
<table>
<thead>
<tr>
<th>TABLE 2. Limitations and Strengths of a Study Affecting the Level of Evidence (for all Studies and Specifically for Systematic Reviews/Meta-Analysis)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>For All Type of Studies</strong></td>
</tr>
<tr>
<td>1. &lt;80% follow-up</td>
</tr>
<tr>
<td>2. &gt;20% missing data or missing data not at random without proper use of missing data statistical techniques</td>
</tr>
<tr>
<td>3. Limited control of confounding (e.g., mortality comparisons with inadequate risk adjustment)</td>
</tr>
<tr>
<td>4. More than minimal bias (selection bias, publication bias, report bias, etc.)</td>
</tr>
<tr>
<td>5. Heterogeneous populations (e.g., institutions with distinct protocols/patient volume, conditions caused by distinct pathogenic mechanisms)</td>
</tr>
<tr>
<td>6. For RCT only: no blinding or improper randomization</td>
</tr>
<tr>
<td>7. Inadequate statistical power: this only applies to studies not finding statistical differences, and it is defined as power &lt; 80% for declaring “failure to detect a significant difference” or power &lt; 90% for declaring “bioequivalence or noninferiority or comparative effectiveness” or area under the Receiver Operating Characteristics curve or both sensitivity and specificity of &lt;80%</td>
</tr>
<tr>
<td><strong>For Systematic Reviews and Meta-analysis</strong></td>
</tr>
<tr>
<td>1. No or inadequate standard search protocol</td>
</tr>
<tr>
<td>2. More than minor chance of publication bias or publication bias nonassessed</td>
</tr>
<tr>
<td>3. Moderate heterogeneity of included studies and/or populations (e.g., elective surgery and acute surgery)</td>
</tr>
<tr>
<td>4. Predominance of Level III or lower studies</td>
</tr>
<tr>
<td>5. No measures or inappropriate measures of pooled risk (for meta-analysis only)</td>
</tr>
<tr>
<td><strong>Strength: Large Effect</strong></td>
</tr>
<tr>
<td>• Study with large relative risk (&gt;5 or &lt;0.2) about condition of low-to-moderate morbidity/mortality.</td>
</tr>
<tr>
<td>• Study with moderate to large relative risk (&gt;2 or &lt;0.5) about condition of high morbidity/mortality.</td>
</tr>
</tbody>
</table>
Collectively Unconscious

News from the Frontiers of Science

NOVEMBER 3, 2012

New version SPSS will include ‘celebratory fireworks’ for significant results

An official press release has confirmed that the newest release of SPSS will be equipped with ‘performance-rewarding features’. The new installment of the popular data-analysis package will light up with song, dance and fireworks whenever a statistical test is significant. ‘We want to provide a package that is in line with the day-to-day experiences of researchers. We understand the pressure the publish, and the relief that is felt by many when those Stars of Significance appear in the results table.’

The level of significance will determine the abundance of the celebrations. If the p-value is below 0.05, researchers will automatically hear what is described as ‘a cheerful tone’, according to a company spokesman. ‘But if your p-value is below 0.01, the software package will play a series of congratulatory videos, complimenting your experimental design and choice of analysis. And if it is very highly significant, or below 0.001, your extra order of magnitude is rewarded by a lavish display of fireworks, clinking of champagne glasses and a showtune that plays ‘Tenure is here to stay’.

Dr. Hellst from the University of Ontario thinks it is a logical step: “Research is hard work. It can take months, sometimes even years, to collect the data. It’s such an anticlimax when are in your office, you run the analysis, the results are significant and the computer is completely and utterly silent. As if it doesn’t even care that my three-way ANOVA came out exactly the way I predicted. I’m so glad that the new edition of SPSS captures my feelings of elation in a suitable, yet professional, manner.”