ARIC Manuscript Proposal # 3775

PC Reviewed: 2/9/21	**Status: _____**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: What's the difference? Clinical answers from standardized logistic models.

b. Abbreviated Title (Length 26 characters): Risks, ORs, RRs, & Differences

2. Writing Group:
Russell Localio, James Henegan, Anne Meibohm, Eliseo Guallar, Eric Ross, Steve Goodman, David Couper, Michael Griswold

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _RL_ **[please confirm with your initials electronically or in writing]**

F**irst author**: **Russell Localio**
Address: 617 Blockley Hall, Univ Pennsylvania, 423 Guardian Dr. Philadelphia PA 19104-6021
Phone: 215-681 7855 (cell and pandemic phone) Fax: NA
E-mail: rlocalio@pennmedicine.upenn.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

Name: **Michael Griswold**
Address: The MIND Center, 2500 N State St, Jackson, MS 29215
Phone: (601) 984-4933 Fax: NA
E-mail: mgriswold@umc.edu

3. Timeline:
Manuscript will be completed in 3-6 months.

4. Rationale:

In a randomized controlled trial with two treatment groups and a binary outcome, a typical report can consist of a concise 2 by 2 table that cross classifies binary exposures and outcomes. The reader expects to find estimates of risks, relative risks and/or risk differences, confidence bounds and often a p-value. But in observational designs, and even in many originally randomized designs, the balance from randomization is absent and potential confounders abound. For that reason, investigative teams often turn to multivariate logistic regression as the method of choice, and to estimate and report odds ratios, a metric that can be difficult to explain to clinicians and patients. Why odds ratios? Because the estimate that flows directly from logistic regression is the (log) odds ratio. Alternatively, investigators sometimes try alternative models that directly produce relative risks or risk differences such as such as log- or identity-link generalized linear model approaches, but these are inherently misspecified and often fail to converge. This brief review relates long-known concepts of standardization and prediction to offer simple methods by which investigators can used conventional and properly specified models to report “adjusted” estimates of risks, their differences, and their ratios to answer their clinical questions. In the process we link these methods to 19th and early 20th century methods of direct standardization and to current concepts of causal inference. We use real-world data from the ARIC study to demostrate the ideas.

Our goals are to translate the value of modern, model based standardization methods for primary results to a clinical / epidemiologic audience. We show that computational advances now permit more sophisticated statistical methods to be more approachable, allowing investigators to use the statistical model that best fits the data and respects the underling biological or clinical process, but simultaneously allowing re-expression of the estimated results into the best metric to support decision making.

We will begin with an example from ARIC on the well-known association of smoking and mortality; The basic unadjusted 2 by 2 table with risks, risk differences, risk ratios and odds ratios are easily estimated (as well as confidence intervals), and methods are largely transparent to readers. We will then show models adjusting for sex (binary), age (continuous) and hypertension (HTN: binary), sex+age, and sex+age+HTN as adjustor sets. While logistic regression converges for all these adjustor sets, log-binomial and identity-binomial models fail to converge when using any but the univariate adjustors. Hence, adjusted odds ratios (OR) are simple to produce, but adjusted relative risks (RR) and absolute risk differences (RD) are less immediate. Investigators often resort to approximate solutions such as log-Poisson models for RRs and (all too often) even identity-Gaussian models for RDs. These models strive to arrive at clinically useful measures of risk but have fundamental shortcomings since they can produce expected risks outside of the possible data range with some risk estimates below 0 or over 1.

We will then (re)-introduce direct standardization methods of obtaining risk estimates on subroups and pooling those estimates using appropriate weighting techniques to allow results that apply to alternative representative samples. While standardization is a common technique used today, standardization in its original form applies only to categorical confounders and usually in large datasets. Fortunately, standardization as implemented using logistic regression, can apply to one or more continuous as well as categorical confounders, and will result in risk estimates that always fall within the bounds of 0 and 1, and can additionally be used to re-express logistic results into adjusted risk differences and risk ratios. Advances in current statistical software makes this process imminently more accessible.

Lastly, we will show that the theoretical link between the standardization approach to re-express logistic regression estimates into RRs and RDs is not necessarily new, but has even greater implications today. The link became the subject of two papers (Kalton 1968) in the context of surveys, then Lane and Nelder (1982) more formally linked standardization with prediction and adjustment for covariates and coined the term “predictive margins”. Additional formal linkages have continued (Little 1982, Rosenbaum 1987), and the connection is present and especially important in the context of more recent advances in causal inference and weighted estimation based on propensity scores (Vittinghoff; Korn and Graubard 1999; Hernan and Robins (2020)).

We will then return to our simple example of smoking on mortality from ARIC. We will re-express the results from the logistic models into RRs & RDs by (1) estimating individual risks, (2) marginalizing by applying appropriate weighting functions to obtain pooled adjusted risk estimates for smokers and non-smokers, and (3) express the results as relative risks, absolute risk differences (and obtain related standard errors, confidence intervals and p-values). We will also use propensity score models to estimate the probability of smoking as a function of the other covariates, sex in the simplest case, following by the estimation of the inverse probability of exposure weights, and finally, re-express weighted logistic regression models of the association of mortality and smoking, into estimates of risk difference and ratios.

Programming code for common statistical software packages (e.g., R, SAS, Stata) will be presented.

5. **Main Hypothesis/Study Questions**:

- This is a **methods translation** paper and no substantive hypotheses are under investigation. A small set of ARIC data with 1 outcome (death), 1 primary exposure (smoking) and a couple adjustors (sex, age and HTN) will only be used motivate the example and allow reproducible code generation.

- The goals are to produce a tutorial-type paper demonstrating how to use existing standardization (weighting) methods and widely used software to re-express estimates from logistic regression models into more clinically interpretable estimates of absolute and relative differences in probabilities ("risk differences" and "relative risks"), thereby avoiding inherent estimation issues with log-link or identity-link methods.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodological limitations or challenges if present).

Study design: Prospective observational study of ARIC participants seen at Visit 1 and followed for approximately 30 years (i.e. through Visit 7).

Exclusion criteria: None

Outcome: 30-year Death status (i.e. through most available data via status71 file).

Exposure: Current Smoking at Visit 1

Covariates: Sex, Age, hypertension status, all at Visit 1

Statistical analyses:
We will demonstrate and compare pros and cons of the following approaches:

- Two-by-two tables (unadjusted ORs, RRs, RDs)

- Log-binomial regression to estimate adjusted RRs with binary outcome data
- Log-Poisson regression to estimate adjusted RRs with binary outcome data

- Identity-binomial regression to estimate adjusted RDs with binary outcome data
- Identity-Gaussian regression to estimate adjusted RDs with binary outcome data

- Logistic (logit-binomial) regression to estimate adjusted ORs with binary outcome data
    - Weighting and standardization after logistic regression to re-express results as:
        - Adjusted marginal probability estimates with Confidence Intervals (CIs)
        - Adjusted RRs with CIs and p-values
        - Adjusted RDs with CIs and p-values

We will also connect the standardization approach to newer causal estimation approaches which target estimands such as the “average treatment effect among the treated”, or the “ATT”. Such estimates address clinical (and counterfactual) questions such as: “What would one expect to happen if the patients who smoked had instead been non-smokers, but without changing their other characteristics?”

References (partial list)

Kalton, G. (1968). Standardization: A technique to control for extraneous variables.
Applied Statistics, 23, 118-136.

Kleinman LC, Norton EC. What's the Risk? A simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. Health Serv Res. 2009 Feb;44(1):288-302. doi: 10.1111/j.1475-6773.2008.00900.x. Epub 2008 Sep 11. PMID: 18793213; PMCID: PMC2669627.

Graubard BI, Korn EL. Predictive margins with survey data. Biometrics. 1999 Jun;55(2):652-9. doi: 10.1111/j.0006-341x.1999.00652.x. PMID: 11318229.

Lane PW, Nelder JA: Analysis of covariance and standardization as instances of prediction. Biometrics 38: 613-621, 1982

Little, R. J. A. (1982). Direct standardization as a tool for teaching linear models for
unbalanced data. American Statistician, 36(1), 38-43.

Ian C. Marschner, Alexandra C. Gillett, Relative risk regression: reliable and flexible methods for log-binomial models, *Biostatistics*, Volume 13, Issue 1, January 2012, Pages 179–192, https://doi.org/10.1093/biostatistics/kxr030

Louise-Anne McNutt, Chuntao Wu, Xiaonan Xue, Jean Paul Hafner, Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes, *American Journal of Epidemiology*, Volume 157, Issue 10, 15 May 2003, Pages 940–943, https://doi.org/10.1093/aje/kwg074

1.

7.a. Will the data be used for non-CVD analysis in this manuscript?
____ Yes _X_ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = “CVD Research” for non-DNA analysis, and for DNA analysis RES_DNA = “CVD Research” would be used? N/A
____ Yes ____ No

(This file ICTDER has been distributed to ARIC PIs, and contains
the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript?
____ Yes __X__ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"? N/A
____ Yes ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php
__**X**__ Yes ________ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

None

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? ____ Yes __X__ No

11.b. If yes, is the proposal N/A
___ A. primarily the result of an ancillary study (list number* _________)
___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* ____________________ __________)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PUBMED Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to Pubmed central.