Northwestern Medicine®

StatTag: A Practical Approach to Reproducibility in Clinical and Translational Science

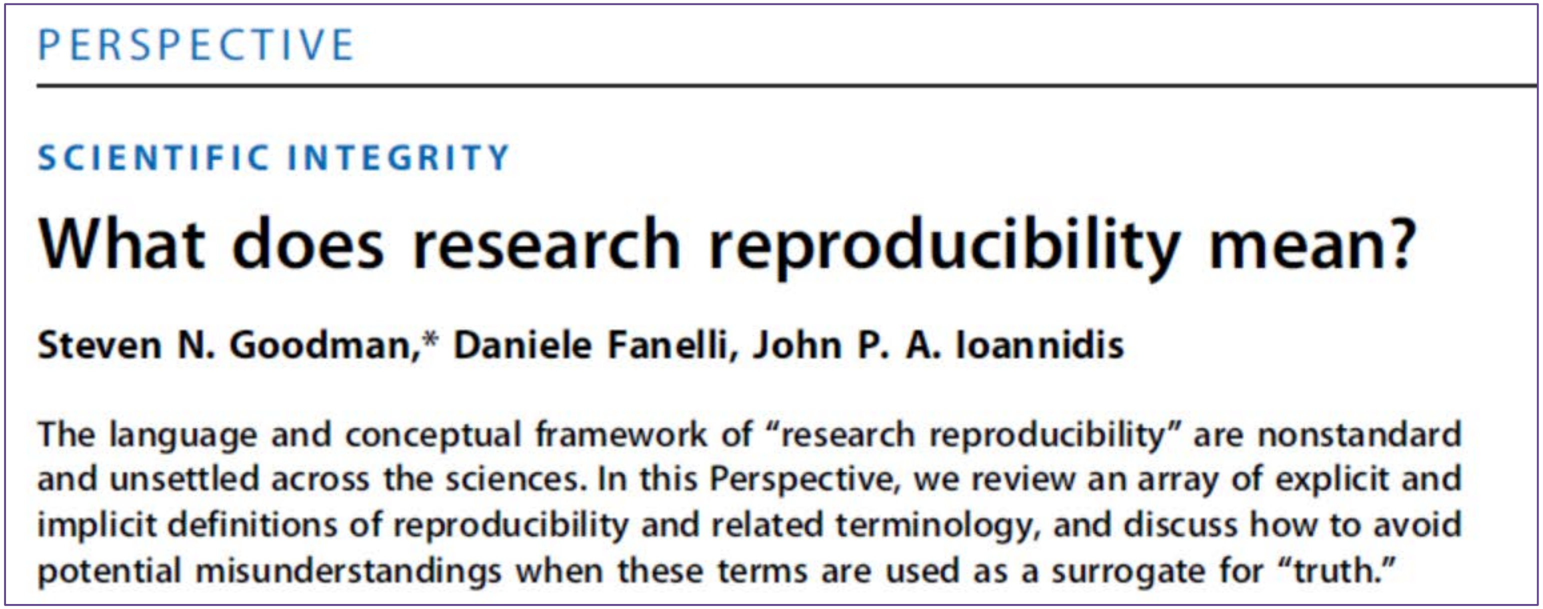
ACTStat 2017

Baltimore, MD

Leah J. Welty, PhD

Reproducible Research

Some confusion



PERSPECTIVE

SCIENTIFIC INTEGRITY

What does research reproducibility mean?

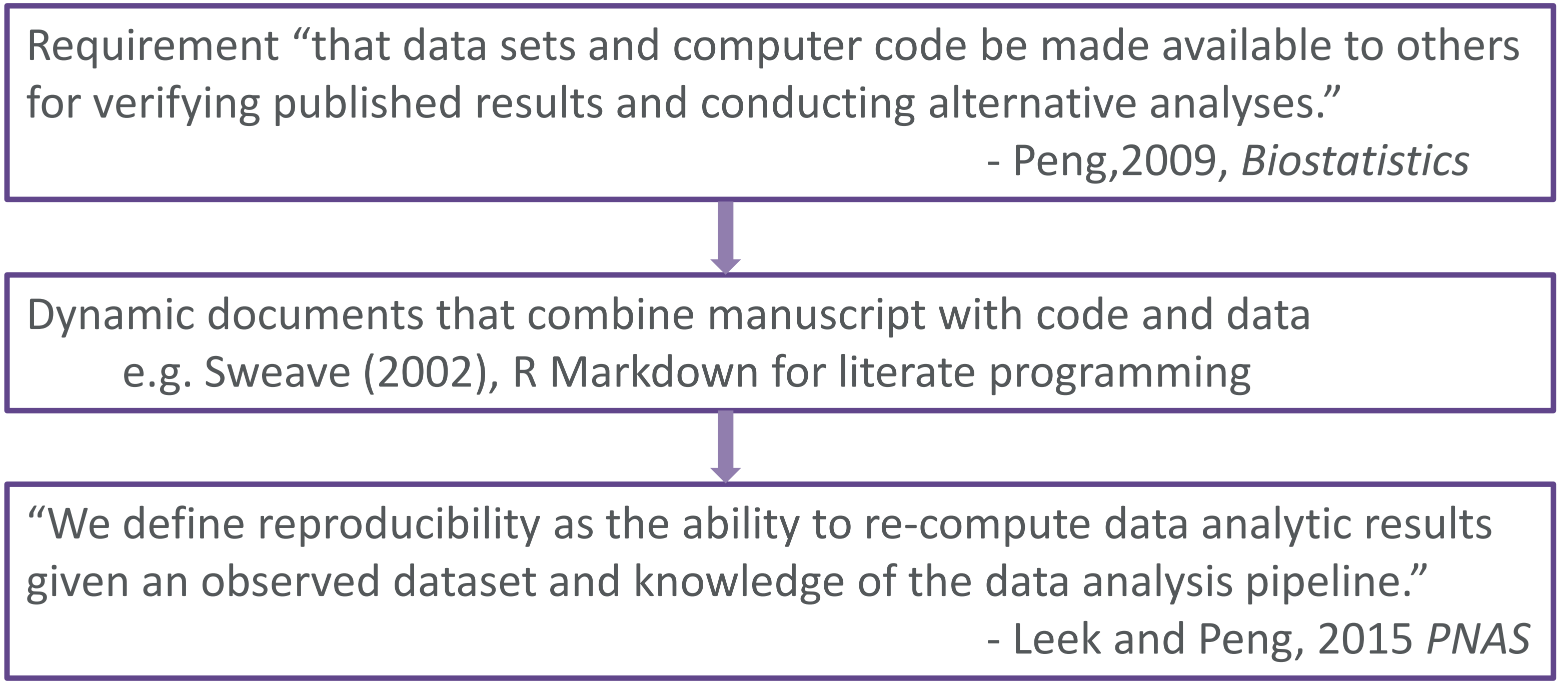
Steven N. Goodman,* Daniele Fanelli, John P. A. Ioannidis

The language and conceptual framework of “research reproducibility” are nonstandard and unsettled across the sciences. In this Perspective, we review an array of explicit and implicit definitions of reproducibility and related terminology, and discuss how to avoid potential misunderstandings when these terms are used as a surrogate for “truth.”

Goodman, Fanelli, and Ioannidis. “What does research reproducibility mean? *Science Translational Medicine* 01 Jun 2016: Vol. 8, Issue 341, pp. 341ps12

What is reproducible research?

An Evolution in Biostatistics



Requirement “that data sets and computer code be made available to others for verifying published results and conducting alternative analyses.”

- Peng,2009, *Biostatistics*

Dynamic documents that combine manuscript with code and data

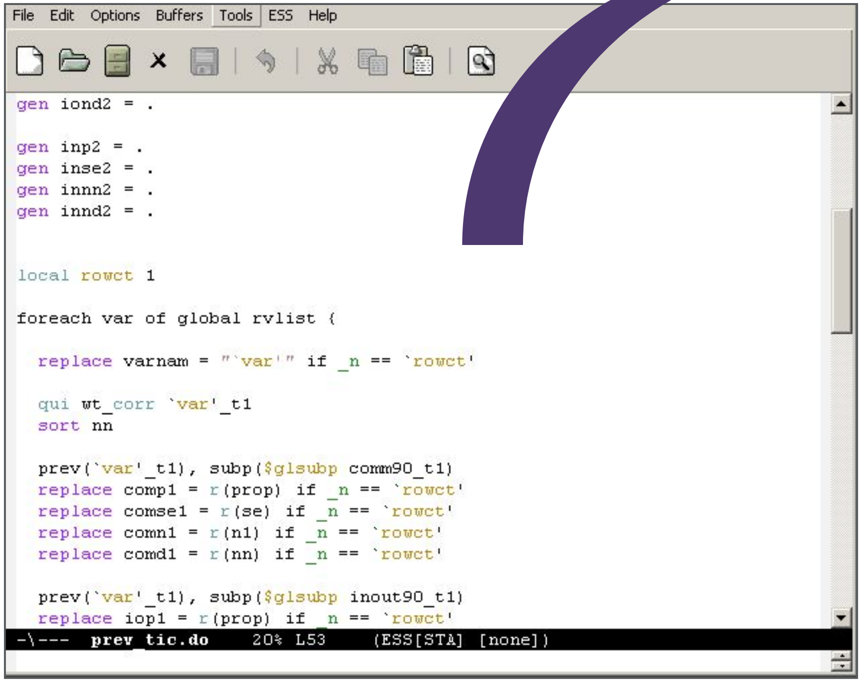
e.g. Sweave (2002), R Markdown for literate programming

“We define reproducibility as the ability to re-compute data analytic results given an observed dataset and knowledge of the data analysis pipeline.”

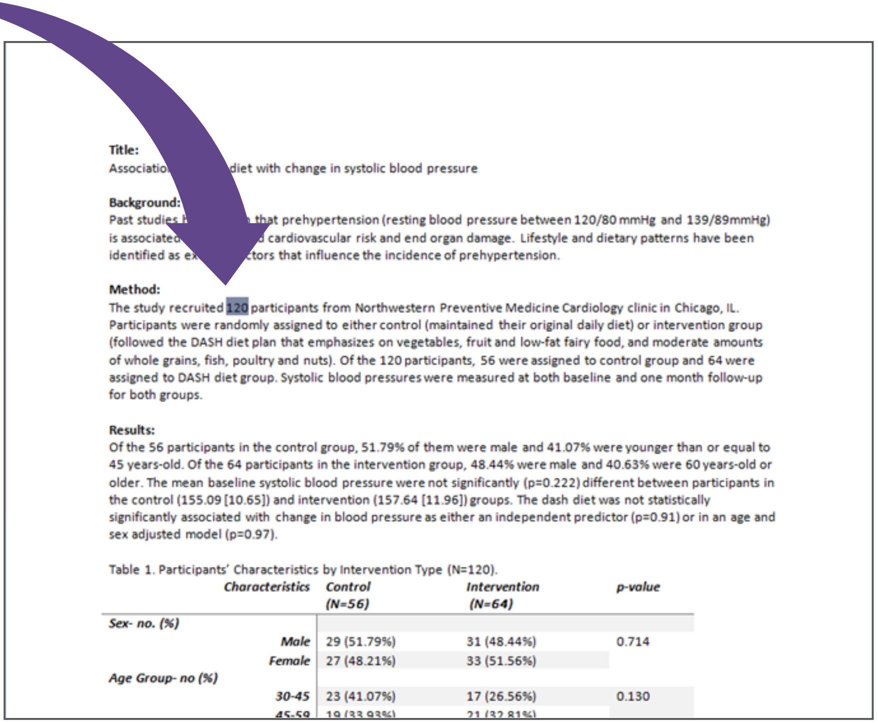
- Leek and Peng, 2015 *PNAS*

Tools for Reproducible Research: Manuscript Prep

Dynamic Documents



```
gen iond2 = .

gen inp2 = .
gen inse2 = .
gen innn2 = .
gen innd2 = .


local rowct 1

foreach var of global rvlist {

  replace varnam = "`var'" if _n == `rowct'

  qui wt_corr `var'_t1
  sort nn

  prev(`var'_t1), subp($glsubp comm90_t1)
  replace comp1 = r(prop) if _n == `rowct'
  replace comse1 = r(se) if _n == `rowct'
  replace comn1 = r(n1) if _n == `rowct'
  replace comd1 = r(nn) if _n == `rowct'

  prev(`var'_t1), subp($glsubp inout90_t1)
  replace iop1 = r(prop) if _n == `rowct'
-\---  prev_tic.do     20% L53     (ESS[STA] [none])
```



Title:
Associatio [illegible] diet with change in systolic blood pressure

Background:
Past studies [illegible] that prehypertension (resting blood pressure between 120/80 mmHg and 139/89mmHg) is associated [illegible] cardiovascular risk and end organ damage. Lifestyle and dietary patterns have been identified as ex[illegible]ctors that influence the incidence of prehypertension.

Method:
The study recruited 120 participants from Northwestern Preventive Medicine Cardiology clinic in Chicago, IL. Participants were randomly assigned to either control (maintained their original daily diet) or intervention group (followed the DASH diet plan that emphasizes on vegetables, fruit and low-fat fairy food, and moderate amounts of whole grains, fish, poultry and nuts). Of the 120 participants, 56 were assigned to control group and 64 were assigned to DASH diet group. Systolic blood pressures were measured at both baseline and one month follow-up for both groups.

Results:
Of the 56 participants in the control group, 51.79% of them were male and 41.07% were younger than or equal to 45 years-old. Of the 64 participants in the intervention group, 48.44% were male and 40.63% were 60 years-old or older. The mean baseline systolic blood pressure were not significantly (p=0.222) different between participants in the control (155.09 [10.65]) and intervention (157.64 [11.96]) groups. The dash diet was not statistically significantly associated with change in blood pressure as either an independent predictor (p=0.91) or in an age and sex adjusted model (p=0.97).

Table 1. Participants' Characteristics by Intervention Type (N=120).

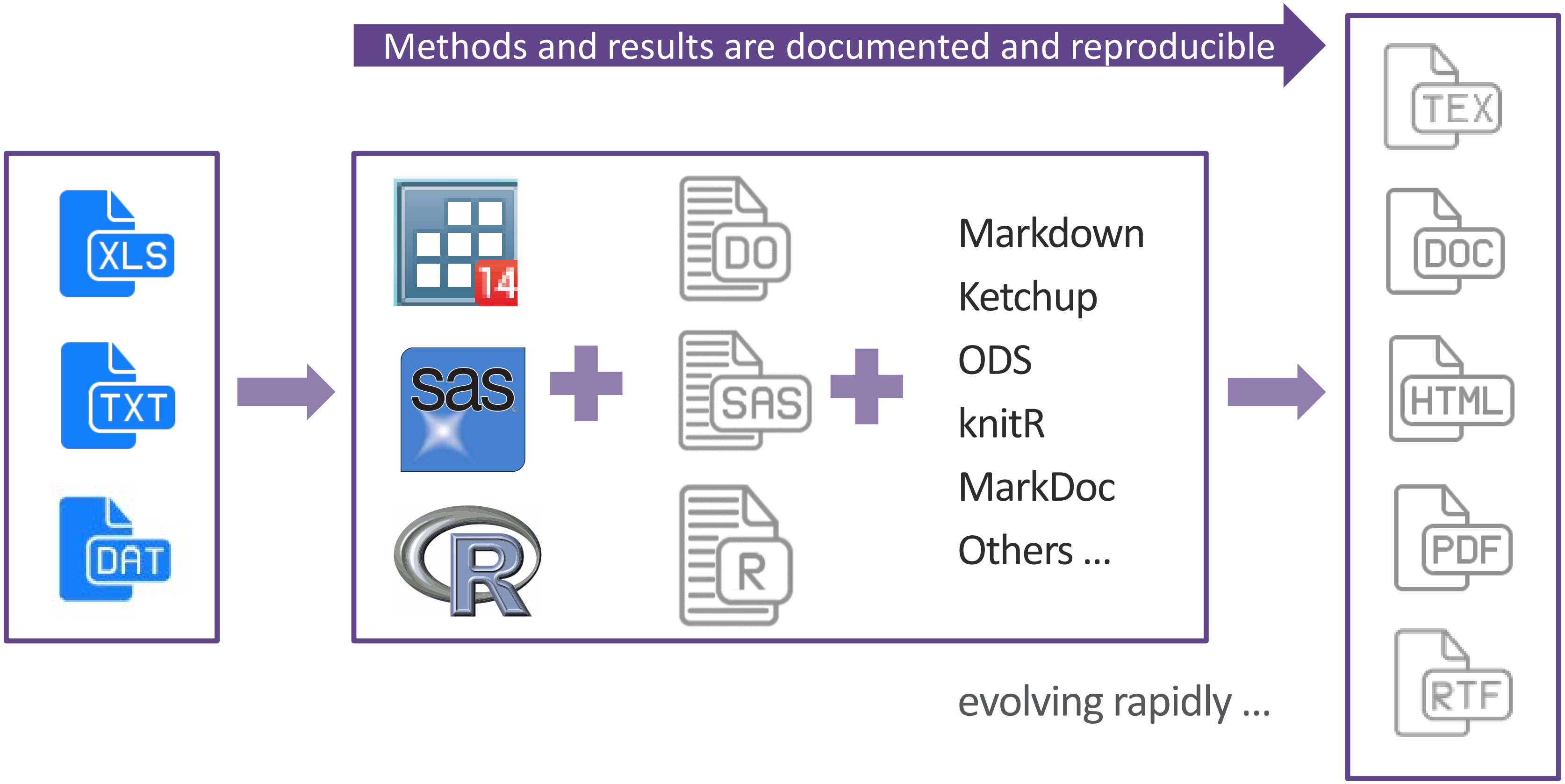
Characteristics	*Control (N=56)*	*Intervention (N=64)*	*p-value*
Sex- no. (%)			
Male	29 (51.79%)	31 (48.44%)	0.714
Female	27 (48.21%)	33 (51.56%)	
Age Group- no (%)			
30-45	23 (41.07%)	17 (26.56%)	0.130
45-59	19 (33.93%)	21 (32.81%)	

- Rather than results being hard coded in a manuscript, they can be updated automatically when data or models change.

Northwestern Medicine®

Tools for Reproducible Research: Manuscript Prep

Existing tools for Dynamic Documents



Methods and results are documented and reproducible

Markdown

Ketchup

ODS

knitR

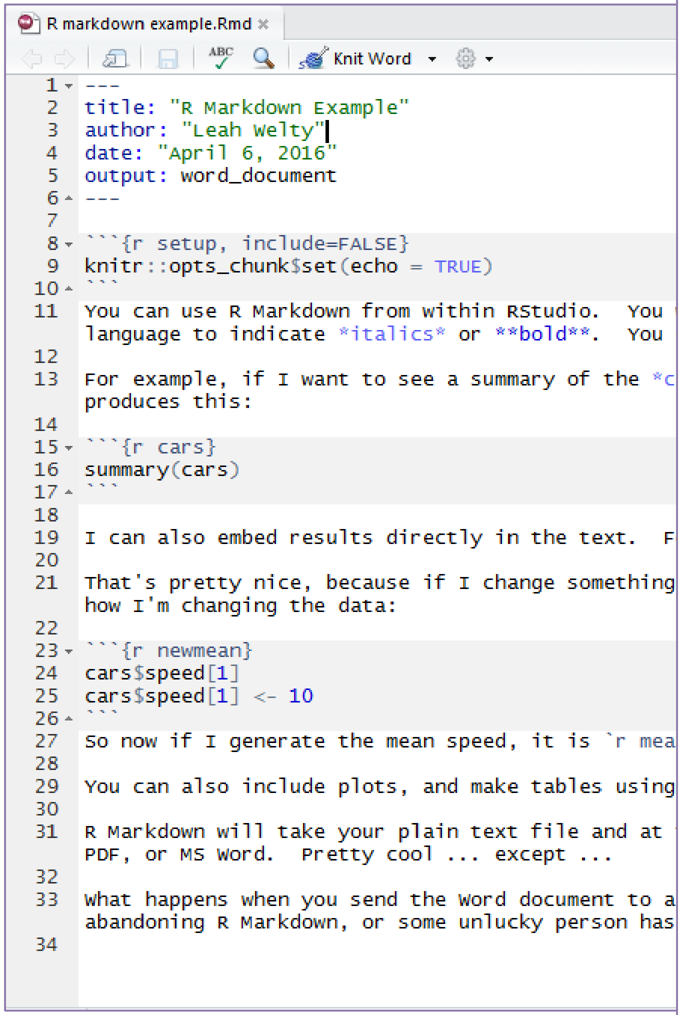
MarkDoc

Others ...

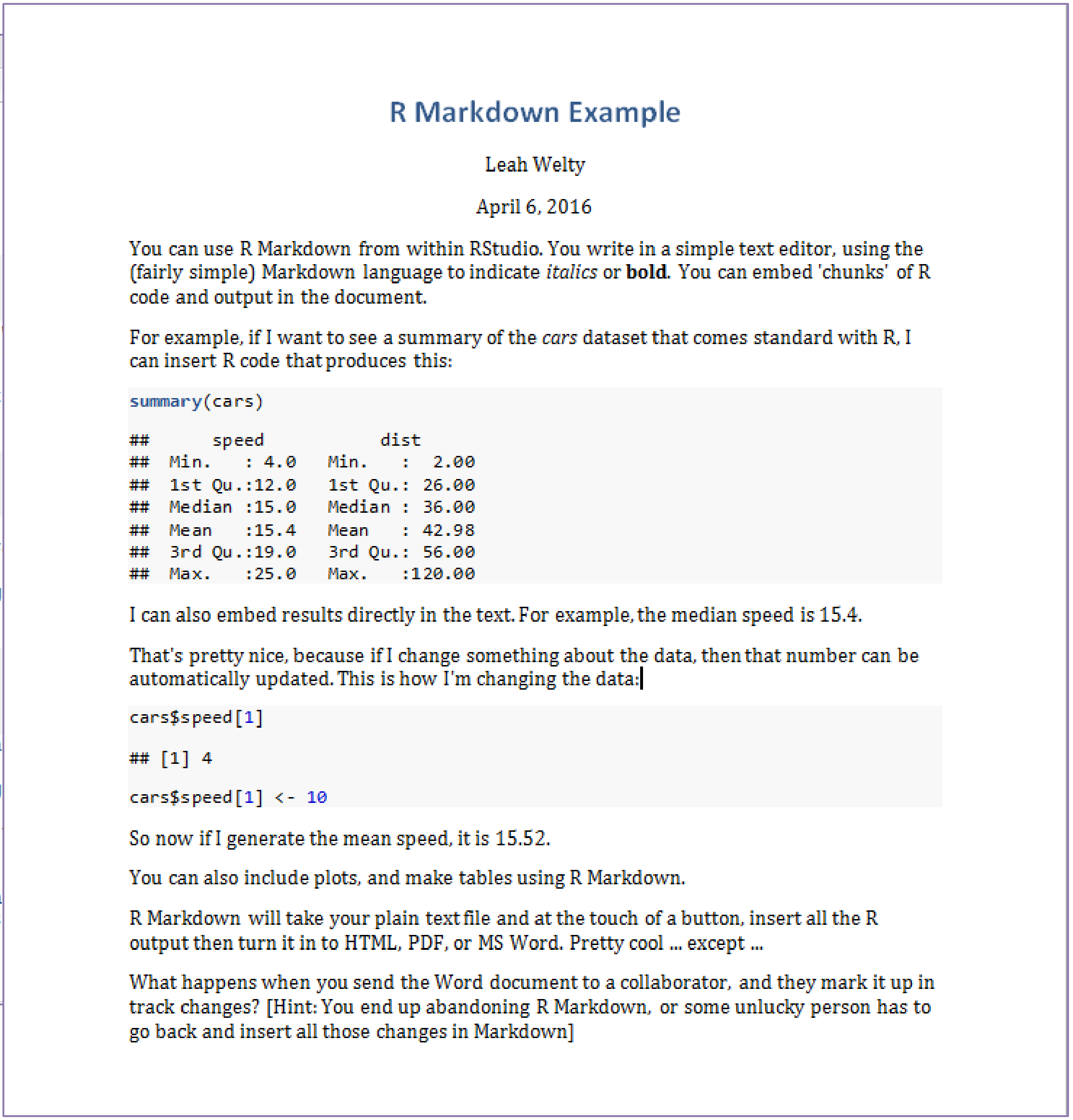
evolving rapidly ...

Tools for Reproducible Research: Manuscript Prep

Dynamic Documents with R Markdown



~~~
---
title: "R Markdown Example"
author: "Leah Welty"
date: "April 6, 2016"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
You can use R Markdown from within RStudio.  You
language to indicate *italics* or **bold**.  You

For example, if I want to see a summary of the *c
produces this:

```{r cars}
summary(cars)
```

I can also embed results directly in the text.  F

That's pretty nice, because if I change something
how I'm changing the data:

```{r newmean}
cars$speed[1]
cars$speed[1] <- 10
```
So now if I generate the mean speed, it is `r mea

You can also include plots, and make tables using

R Markdown will take your plain text file and at
PDF, or MS Word.  Pretty cool ... except ...

What happens when you send the Word document to a
abandoning R Markdown, or some unlucky person has

~~~



R Markdown Example

Leah Welty

April 6, 2016

You can use R Markdown from within RStudio. You write in a simple text editor, using the (fairly simple) Markdown language to indicate *italics* or **bold**. You can embed 'chunks' of R code and output in the document.

For example, if I want to see a summary of the *cars* dataset that comes standard with R, I can insert R code that produces this:

```
summary(cars)

##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

I can also embed results directly in the text. For example, the median speed is 15.4.

That's pretty nice, because if I change something about the data, then that number can be automatically updated. This is how I'm changing the data:

```
cars$speed[1]

## [1] 4

cars$speed[1] <- 10
```

So now if I generate the mean speed, it is 15.52.

You can also include plots, and make tables using R Markdown.

R Markdown will take your plain text file and at the touch of a button, insert all the R output then turn it in to HTML, PDF, or MS Word. Pretty cool ... except ...

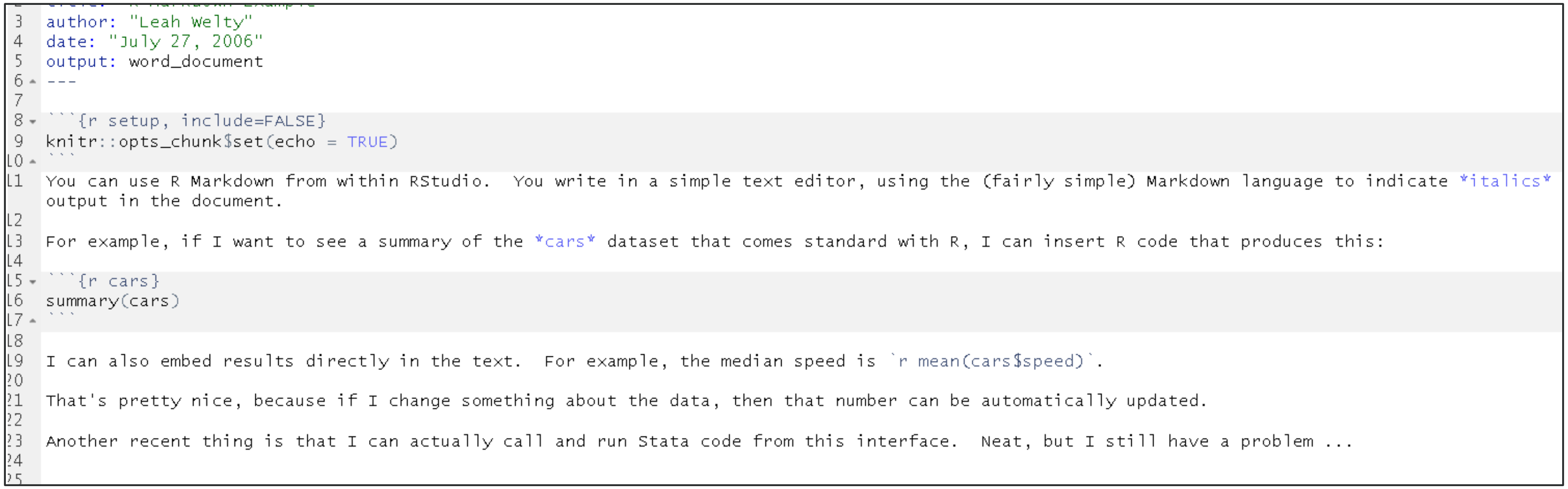
What happens when you send the Word document to a collaborator, and they mark it up in track changes? [Hint: You end up abandoning R Markdown, or some unlucky person has to go back and insert all those changes in Markdown]

Northwestern Medicine®

Tools for Reproducible Research: Manuscript Prep

The Problem with Dynamic Documents: Text Files and Collaborators

Current tools require writing within a text editor. For example, a Markdown document looks something like this:



~~~
title: "R Markdown Example"
author: "Leah Welty"
date: "July 27, 2006"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~
You can use R Markdown from within RStudio. You write in a simple text editor, using the (fairly simple) Markdown language to indicate *italics* output in the document.

For example, if I want to see a summary of the *cars* dataset that comes standard with R, I can insert R code that produces this:

```{r cars}
summary(cars)
```

I can also embed results directly in the text. For example, the median speed is `r mean(cars$speed)`.

That's pretty nice, because if I change something about the data, then that number can be automatically updated.

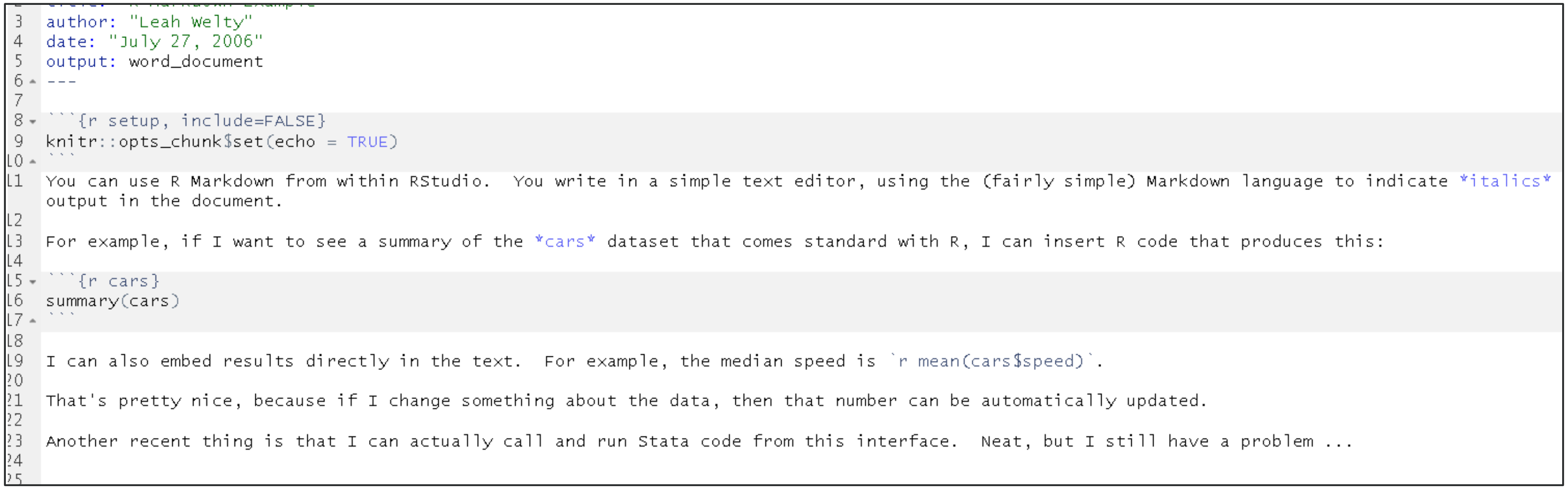
Another recent thing is that I can actually call and run Stata code from this interface. Neat, but I still have a problem ...

```
```

Tools for Reproducible Research: Manuscript Prep

The Problem with Dynamic Documents: Text Files and Collaborators

Current tools require writing within a text editor. For example, a Markdown document looks something like this:



~~~
author: "Leah Welty"
date: "July 27, 2006"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~
You can use R Markdown from within RStudio. You write in a simple text editor, using the (fairly simple) Markdown language to indicate *italics*
output in the document.

For example, if I want to see a summary of the *cars* dataset that comes standard with R, I can insert R code that produces this:

```{r cars}
summary(cars)
```

I can also embed results directly in the text. For example, the median speed is `r mean(cars$speed)`.

That's pretty nice, because if I change something about the data, then that number can be automatically updated.

Another recent thing is that I can actually call and run Stata code from this interface. Neat, but I still have a problem ...
```

Do you have non-technical collaborators who are willing to work this way?  My collaborators (primarily doctors and social scientists) are not.

Northwestern Medicine®
```

Tools for Reproducible Research: Manuscript Prep

A Problem for Dynamic Documents: Track Changes

I create a dynamic document, generate the Word file and send it to collaborators.

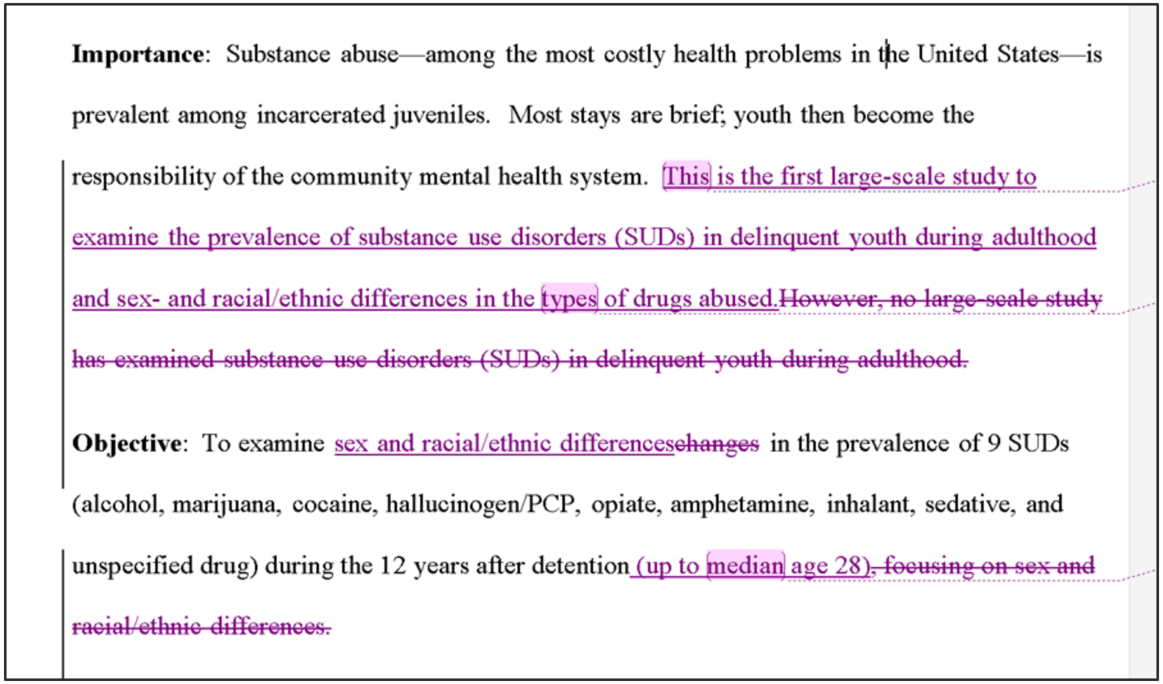
They send back:

Tools for Reproducible Research: Manuscript Prep

A Problem for Dynamic Documents: Track Changes

I create a dynamic document, generate the Word file and send it to collaborators.

They send back:



Importance: Substance abuse—among the most costly health problems in the United States—is prevalent among incarcerated juveniles. Most stays are brief; youth then become the responsibility of the community mental health system. This is the first large-scale study to examine the prevalence of substance use disorders (SUDs) in delinquent youth during adulthood and sex- and racial/ethnic differences in the types of drugs abused.~~However, no large-scale study has examined substance use disorders (SUDs) in delinquent youth during adulthood.~~

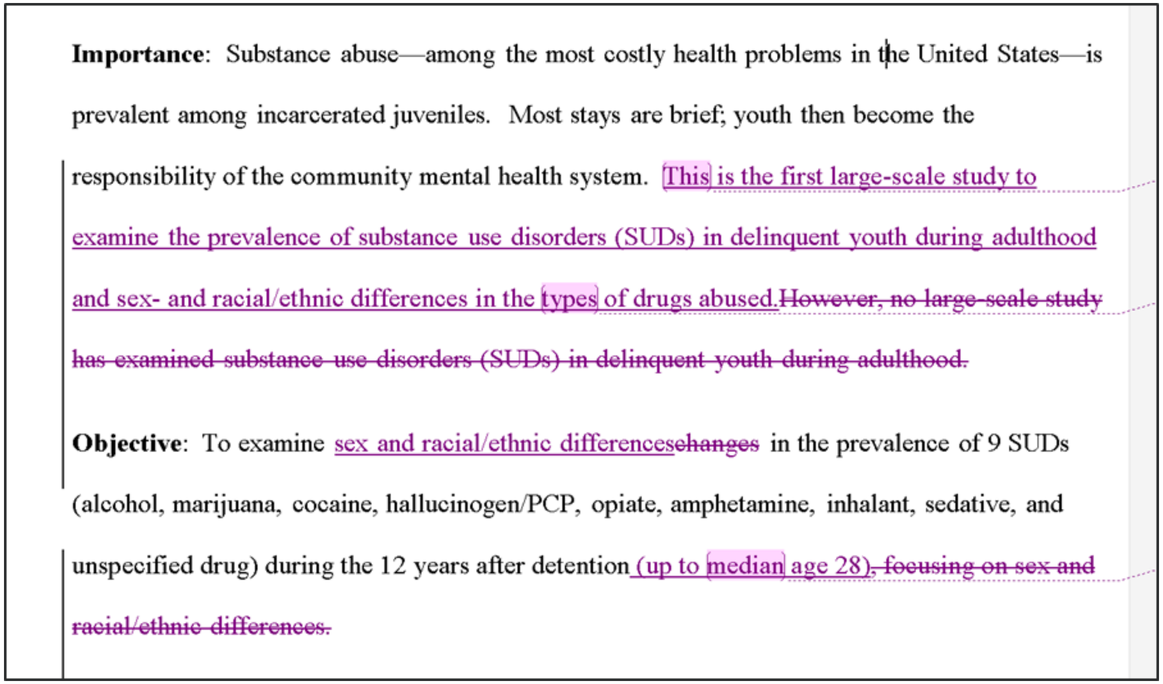
Objective: To examine sex and racial/ethnic differences~~changes~~ in the prevalence of 9 SUDs (alcohol, marijuana, cocaine, hallucinogen/PCP, opiate, amphetamine, inhalant, sedative, and unspecified drug) during the 12 years after detention (up to median age 28)~~, focusing on sex and racial/ethnic differences.~~

Tools for Reproducible Research: Manuscript Prep

A Problem for Dynamic Documents: Track Changes

I create a dynamic document, generate the Word file and send it to collaborators.

They send back:



Importance: Substance abuse—among the most costly health problems in the United States—is prevalent among incarcerated juveniles. Most stays are brief; youth then become the responsibility of the community mental health system. This is the first large-scale study to examine the prevalence of substance use disorders (SUDs) in delinquent youth during adulthood and sex- and racial/ethnic differences in the types of drugs abused.~~However, no large-scale study has examined substance use disorders (SUDs) in delinquent youth during adulthood.~~

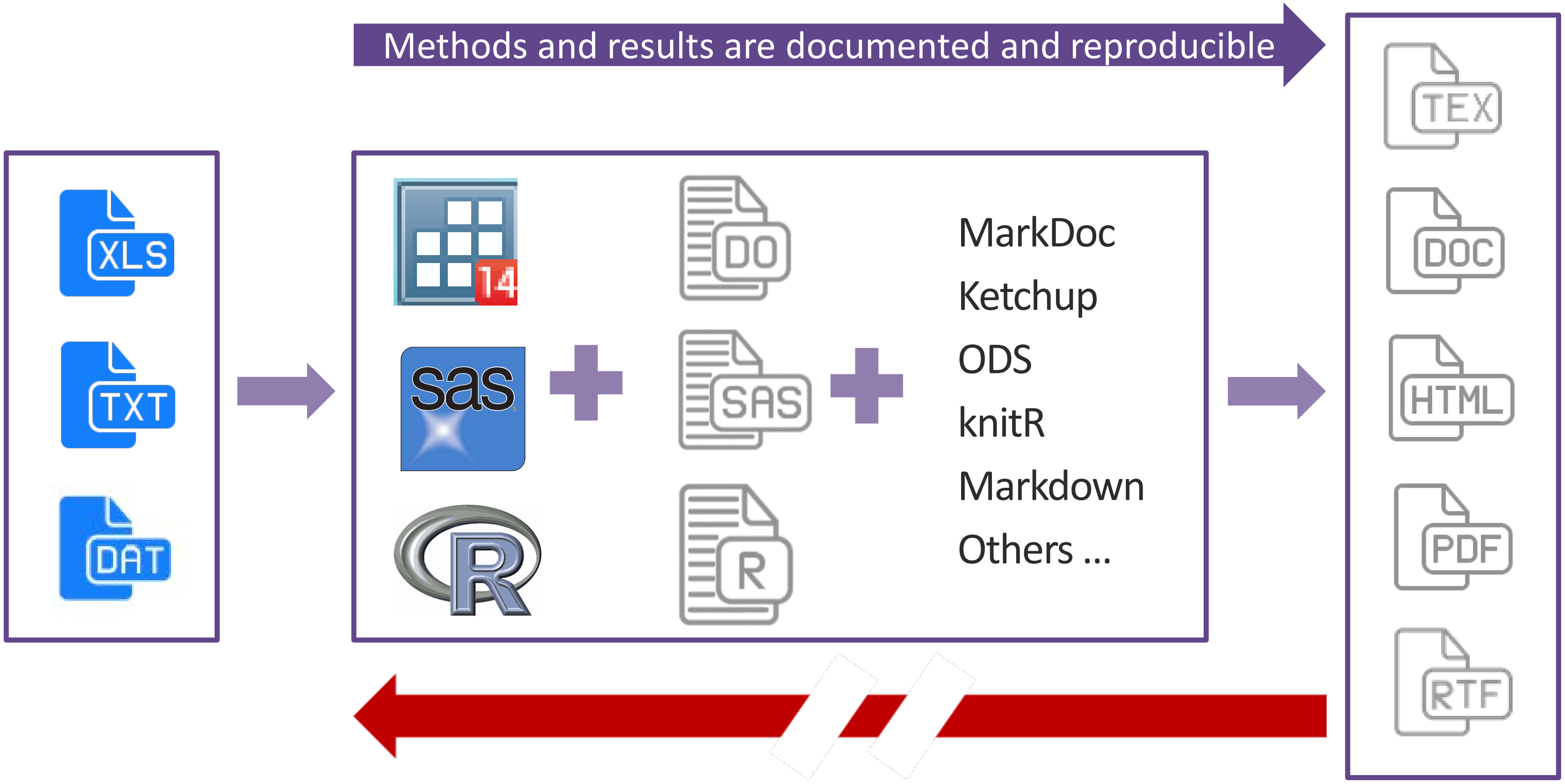
Objective: To examine sex and racial/ethnic differences~~changes~~ in the prevalence of 9 SUDs (alcohol, marijuana, cocaine, hallucinogen/PCP, opiate, amphetamine, inhalant, sedative, and unspecified drug) during the 12 years after detention (up to median age 28)~~, focusing on sex and racial/ethnic differences.~~

I have two (bad) choices:

1. Continue in Word, and loose the dynamic nature of the document.
2. Re-enter all of their changes in my source file.

Tools for Reproducible Research: Manuscript Prep

Limitations of Existing Tools for Dynamic Documents



Methods and results are documented and reproducible

MarkDoc
Ketchup
ODS
knitR
Markdown
Others …

Tools for Reproducible Research: Manuscript Prep

A Problem for Dynamic Documents: MS Word is Ubiquitous

The NEW ENGLAND JOURNAL of MEDICINE

*“All text…should be in one double-spaced electronic document (preferably a **Word Doc**)”*

JAMA The Journal of the American Medical Association

*“For submission and review, please submit the manuscript as a **Word document**. Do not submit your manuscript in PDF format.”*

Science AAAS

“*Science* **prefers to receive files in Word’s .docx** format.”

Reproducible research using Microsoft Word?

Overview

What makes StatTag different than other programs?

- StatTag is a free plug-in for Microsoft Word (Windows) and app (Mac)
  - Connects Stata, SAS (Windows only) or R code and Word document
  - You and your collaborators can work from the same Word document without breaking links between the code and data
  - Can work separately on code and the Word document

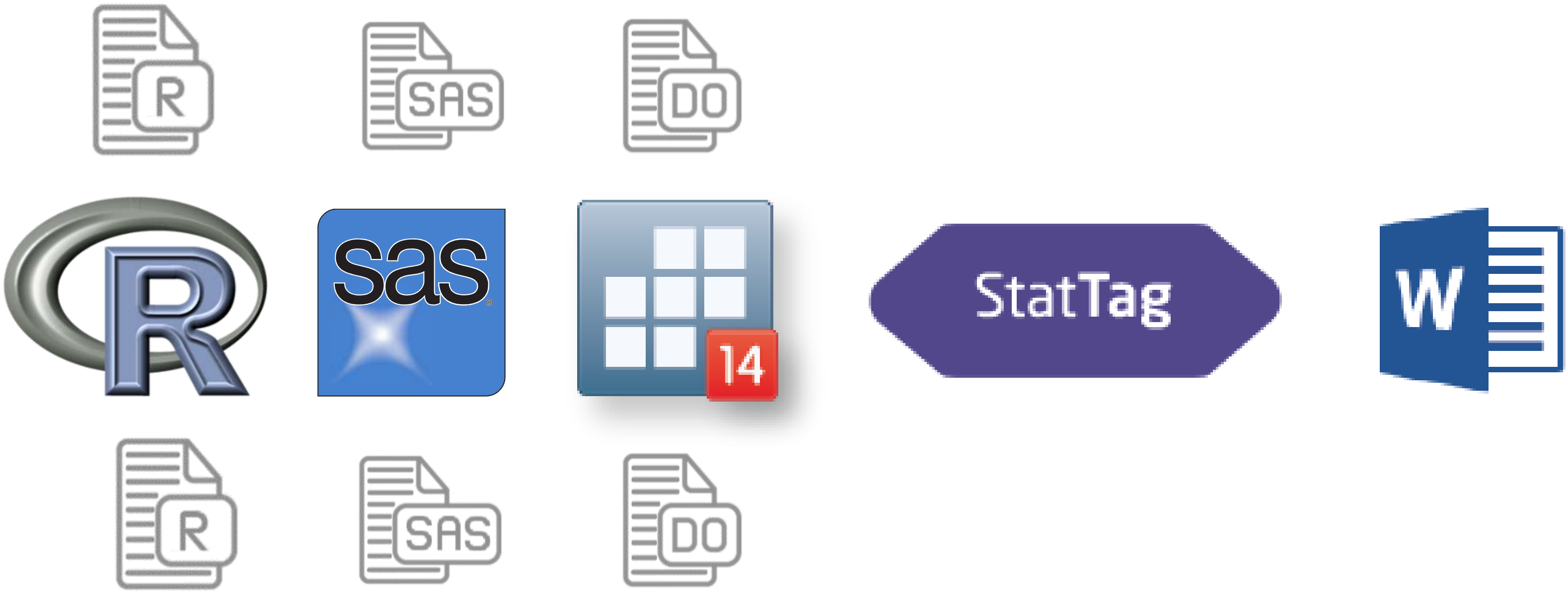
- User-friendly, easy learning curve, still evolving
  - StatTag menu consistent with Word layout
  - EndNote:Citations as StatTag:Results

[See stattag.org for videos]

Software Agnostic

StatTag works with multiple code files of different types

- StatTag can connect to multiple .do, .sas and .r files.



- This is a departure from other tools that work with one specific program (e.g., R Markdown, SAS ODS)

Document Sharing

How does StatTag work when sharing a Word document with collaborators?

StatTag

If I have…	I can…		
	Review/edit manuscript text	View code associated with a tag	Insert or update a tag
Microsoft Word	✓	✗	✗
+ StatTag and Stata/SAS/R code	✓	✓	✗
+ Stata/SAS/R code and Data	✓	✓	✓

Northwestern Medicine®

Windows and Mac Versions

Not just for R/SAS/Stata on Windows

StatTag

- **The first releases were for Windows and Stata/SAS/R.**
- A Mac version of StatTag for Stata and R is available in beta.

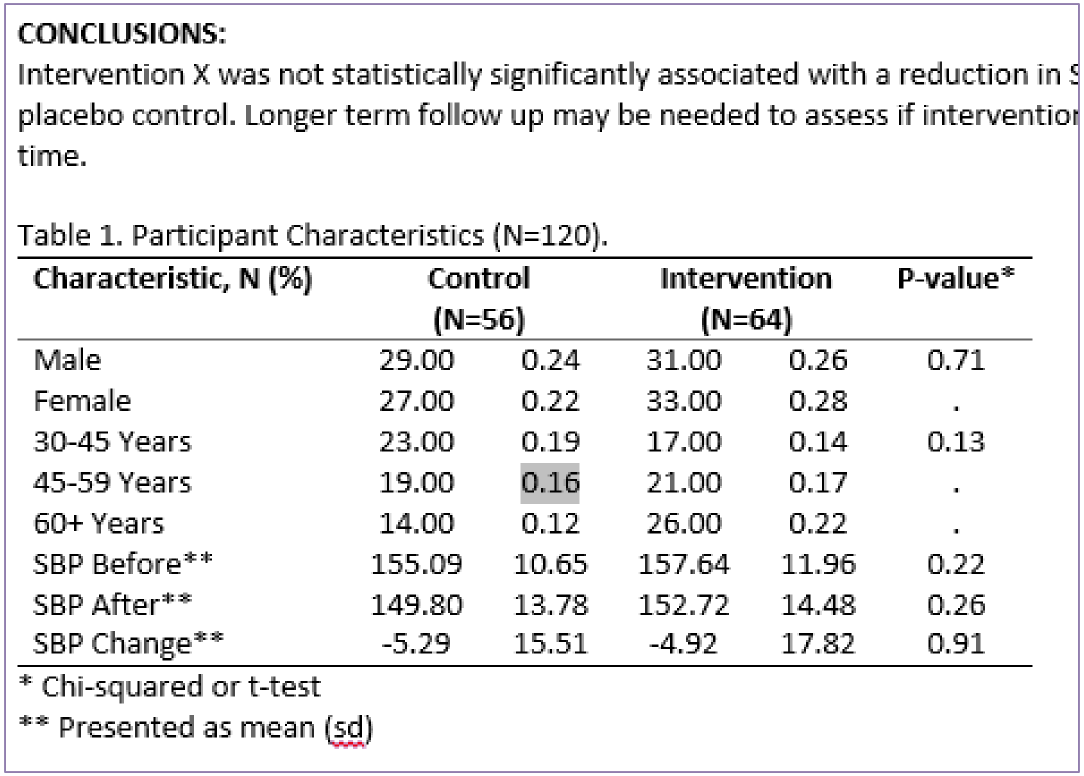
	Stata	SAS	R
Windows	✓ (July 2016)	✓ (Sept 2016)	✓ (May 2017)
Mac	✓ (July 2017) *Beta*	X	✓ (July 2017) *Beta*

Northwestern Medicine®

Tag highlighting

StatTag identifies tags in a document

- Inserted tags are highlighted when they are clicked on.
- Future versions will include a “highlight all tags” function to quickly find any inserted tags in a document.



CONCLUSIONS:
Intervention X was not statistically significantly associated with a reduction in S
placebo control. Longer term follow up may be needed to assess if interventio
time.

Table 1. Participant Characteristics (N=120).

Characteristic, N (%)	Control (N=56)		Intervention (N=64)		P-value*
Male	29.00	0.24	31.00	0.26	0.71
Female	27.00	0.22	33.00	0.28	.
30-45 Years	23.00	0.19	17.00	0.14	0.13
45-59 Years	19.00	0.16	21.00	0.17	.
60+ Years	14.00	0.12	26.00	0.22	.
SBP Before**	155.09	10.65	157.64	11.96	0.22
SBP After**	149.80	13.78	152.72	14.48	0.26
SBP Change**	-5.29	15.51	-4.92	17.82	0.91

* Chi-squared or t-test

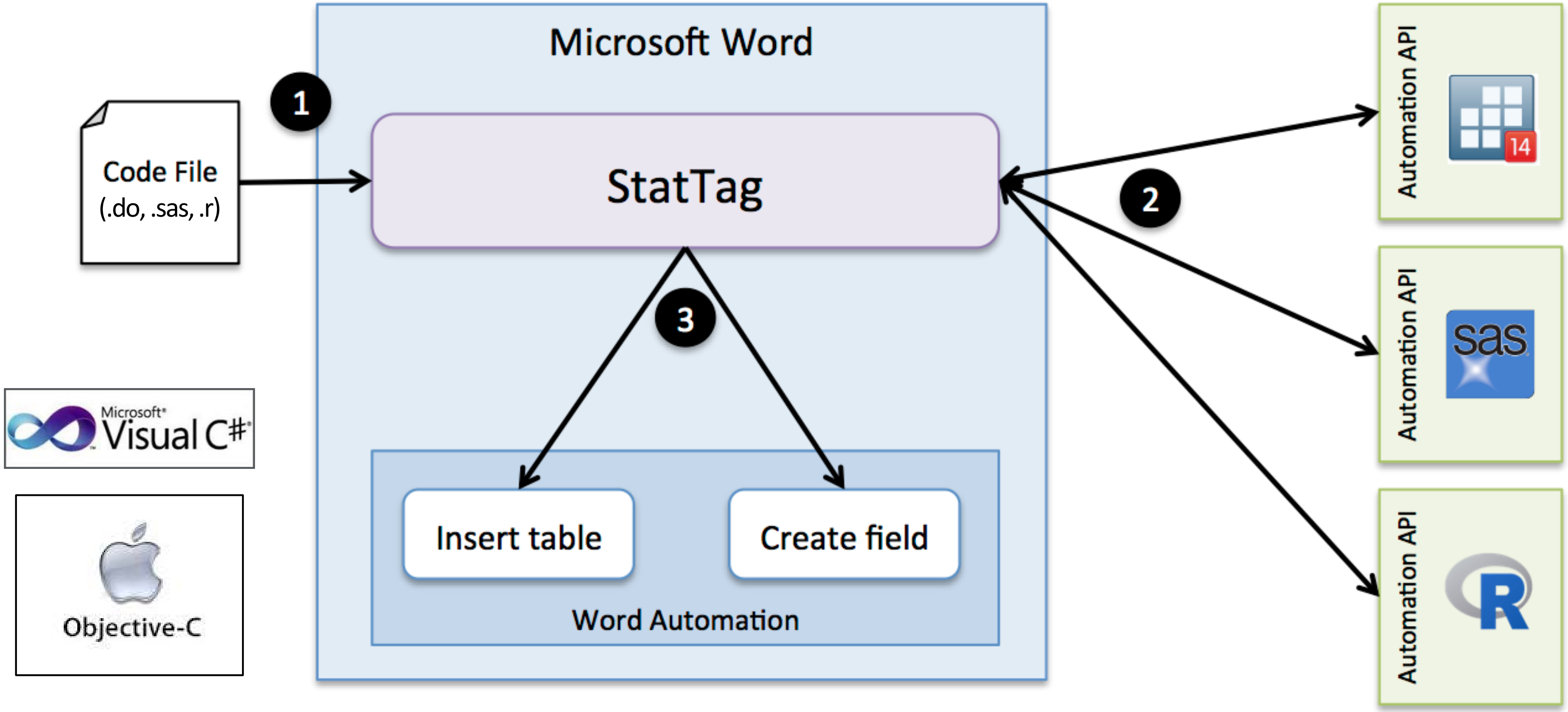
** Presented as mean (sd)

Data Security

Concerns over PHI, PII

- StatTag *doesn't* store a copy of the data.
- StatTag will eventually store a *read-only* copy of code files

Architecture

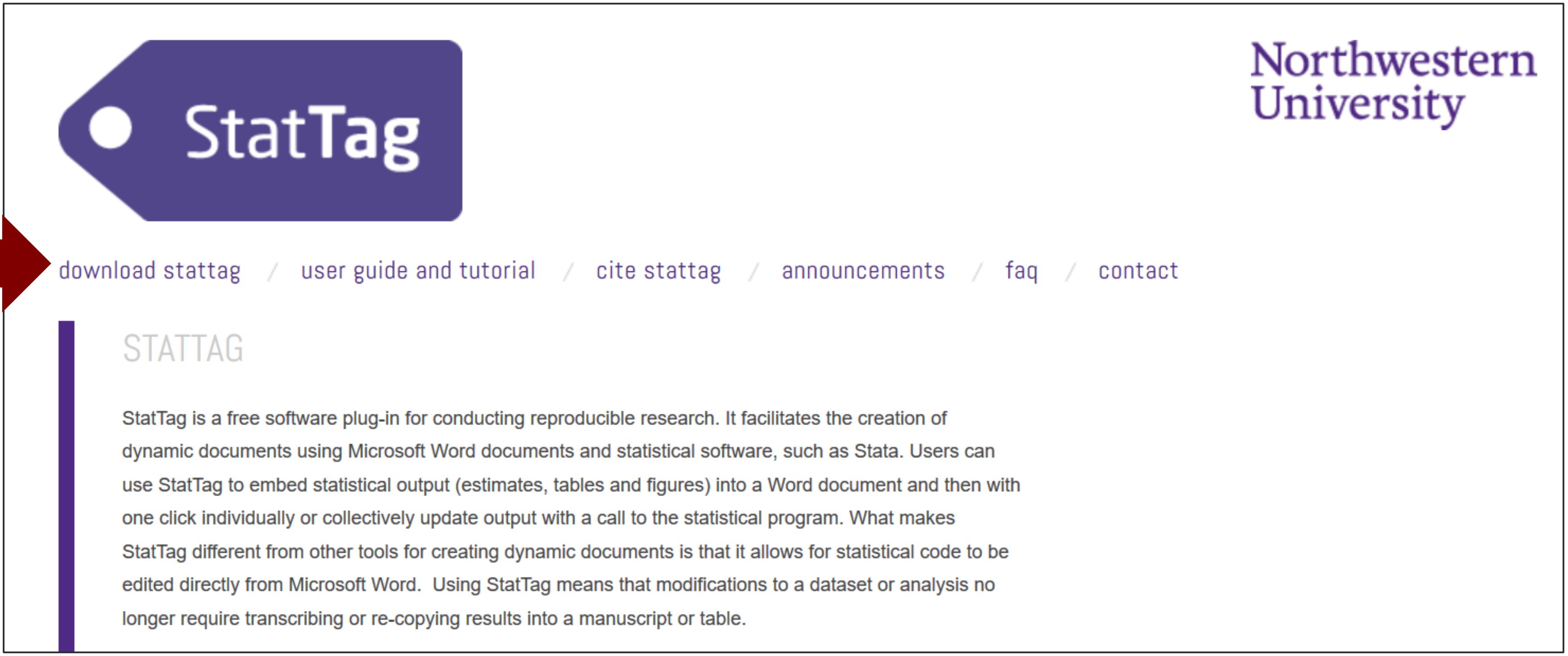


1. Read the code file & parse out the tags
2. Send commands to the stat program and get individual results
3. Use Word automation to add results to the document (using native Word formatting for tables and fields).

Getting StatTag

Freely available at stattag.org

stattag.org



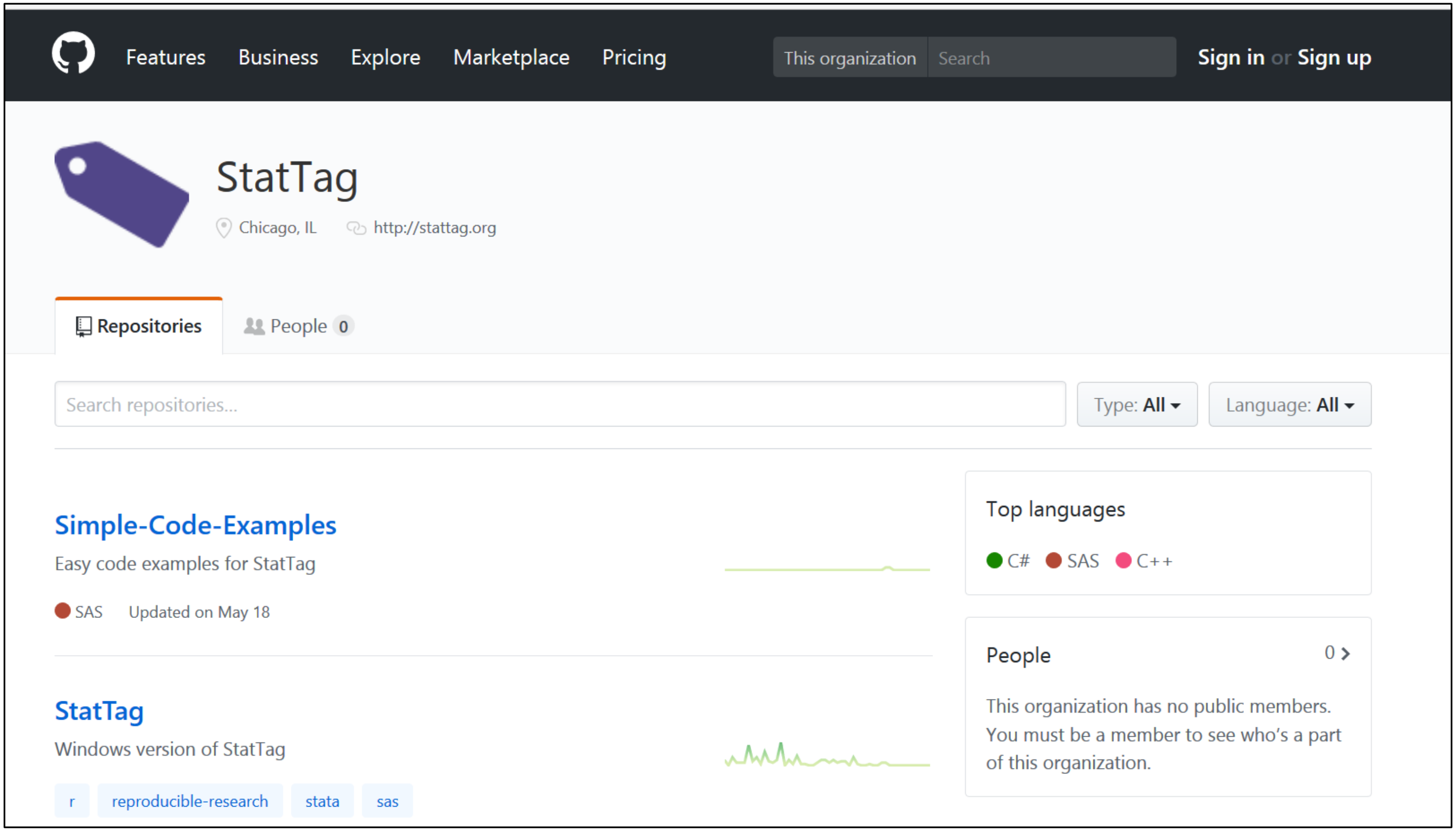
StatTag
Northwestern University
download stattag / user guide and tutorial / cite stattag / announcements / faq / contact

STATTAG

StatTag is a free software plug-in for conducting reproducible research. It facilitates the creation of dynamic documents using Microsoft Word documents and statistical software, such as Stata. Users can use StatTag to embed statistical output (estimates, tables and figures) into a Word document and then with one click individually or collectively update output with a call to the statistical program. What makes StatTag different from other tools for creating dynamic documents is that it allows for statistical code to be edited directly from Microsoft Word. Using StatTag means that modifications to a dataset or analysis no longer require transcribing or re-copying results into a manuscript or table.

(Really) Getting StatTag

GitHub repo at github.com/StatTag



StatTag

Chicago, IL   http://stattag.org

Simple-Code-Examples

Easy code examples for StatTag

● SAS   Updated on May 18

StatTag

Windows version of StatTag

r   reproducible-research   stata   sas

Top languages

● C# ● SAS ● C++

People   0 ›

This organization has no public members. You must be a member to see who's a part of this organization.

Planned Enhancements and Future Work

StatTag

- Compatibility with native code editors
- Improved table construction and formatting
- Tag highlighting and inventory
- Possible extensions to Matlab, SQL, other programs
- **Feedback – suggestions, complaints, comments – are very welcome!**

  stattag@northwestern.edu

Citation and Acknowledgements

- We ask that anyone who uses StatTag to please cite:
  - Welty, L.J., Rasmussen, L.V., Baldridge, A.S, and Whitley E. (2016). *StatTag*. Chicago, Illinois, United States: Galter Health Sciences Library. doi:10.18131/G3K76

- StatTag is distributed under the MIT License

- StatTag was developed with funding through a Clinical Translational Sciences Award (CTSA) to Northwestern University. Tracking the impact of the award is a key metric in demonstrating effectiveness.

CTSA Clinical & Translational Science Awards

Northwestern Medicine | NUCATS Clinical and Translational Sciences Institute

Acknowledgements (continued)

StatTag

- StatTag was inspired in part by the Stata Automation Report project: Lo Magno, G.L. (2013). Sar: Automatic generation of statistical reports using Stata and Microsoft Word for Windows. *The Stata Journal*, 13(1); 39-64.

- StatTag makes use of the following open source projects:

  Scintilla - http://www.scintilla.org/

  ScintillaNET - https://github.com/jacobslusser/ScintillaNET

  Json.NET - http://www.newtonsoft.com/json

  Use of these projects does not imply endorsement of StatTag by the respective project owners, or endorsement of the use of these projects by Northwestern University.

References

- Peng, R “Reproducible Research in Computational Science” (Science) 2 December 2011, vol. 334
- Leek and Peng “Opinion: Reproducible Research can still be wrong: Adopting a prevention approach” PNAS, February 10, 2015, vol. 112, no. 6, 1645–1646
- Matthias Schwab, Martin Karrenback, and Jon Claerbout “Making scientific computations reproducible” (2000) *Computing in Science and Engineering*, 2, pp. 61 – 67.
- Roger D. Peng. “Reproducible research and *Biostatistics.”* (2009) *Biostatistics,* pp. 405-408.
- Paul Thompson and Andrew Burnett. “Reproducible Research” CORE Issues in Professional and Research Ethics, Volume 1, Paper 6, 2012. Accessed from http://nationalethicscenter.org/content/article/175
- Jonathan Buckheit and David Donoho. “WaveLab and Reproducible Research.” (1995) Technical Report No. 474, Department of Statistics, Stanford Univeristy. Accessed from http://statistics.stanford.edu/~ckirby/techreports/NSF/EFS%20NSF%20474.pdf, February 2013.
- Babel: Introduction. OrgMode; Available from: http://orgmode.org/worg/org-contrib/babel/intro.html.
- Buckheit JB, Donoho DL. WaveLab and Reproducible Research. In: Antoniadis A, Oppenheim G, editors. Wavelets and Statistics. Springer New York; 1995. p. 55-81.
- Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. Ann Intern Med. 2007;146(6):450-3.
- Thompson PA, Burnett A. Reproducible Research. CORE Issues [serial on the Internet]. 2012; Vol. 1, Paper 6: Available from: https://nationalethicscenter.org/content/article/175.
- Ware J. Reproducible Research Standards and Definitions. CTSPedia; 2010; Available from: http://www.ctspedia.org/do/view/CTSpedia/ReproducibleResearchStandards.
- Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. Annals of Applied Statistics. 2009;3(4):1309-34.
- Baggerly KA, Coombes KR. What information should be required to support clinical "omics“ publications? Clin Chem. 2011;57(5):688-90.
- Coombes KR, Wang J, Baggerly KA. Microarrays: Retracing steps [4]. Nat Med. 2007;13(11):1276-7.
- Donoho DL. An invitation to reproducible computational research. Biostatistics. 2010;11(3):385-8.
- Gentleman R. Reproducible research: a bioinformatics case study. Stat Appl Genet Mol Biol. 2005;4:Article2.
- Announcement: Reducing our Irreproducibility. Nature. 2013;496:398 (Editorial).
- Code share: Papers in Nature journals should make computer code accessible where possible. Nature. 2014;514:536 (Editorial).
- Diggle PJ, Zeger SL. Embracing the concept of reproducible research. Biostatistics (Oxford, England). 2010;11(3):375.
- Godlee F, Groves T. The new BMJ policy on sharing data from drug and device trials. BMJ (Clinical research ed). 2012;345:e7888.
- Groves T, Godlee F. Open science and reproducible research. BMJ (Clinical research ed). 2012;344:e4383.
- McNutt M. Journals unite for reproducibility. Science (New York, NY). 2014;346(6210):679.

Thank You!

StatTag

C 118
C 117
StatTag: A Tool for Connecting Statistical Output to Microsoft Word
StatTag