Northwestern Medicine®

Connecting R/R Markdown and Microsoft Word using StatTag for Collaborative Reproducibility

R-Ladies

September 26, 2019

Evanston, Illinois

Leah J. Welty, PhD

Disclosures

- Joint work with:
  - Luke V. Rasmussen, Lead Software Developer
  - Abigail S. Baldridge, Senior Statistical Analyst
  - Eric W. Whitley, Software Developer

- This project was supported, in part, by the National Institutes of Health's National Center for Advancing Translational Sciences, Grant Number UL1TR001422. *The content is solely the responsibility of the developers and does not necessarily represent the official views of the National Institutes of Health.*

Northwestern Medicine®

Reproducible Research: Background

Reproducible Research: Background

Origins lie in the inconvenience of *irreproducible* research

"In the mid-1980s, we realized that our laboratory's researchers often had difficulty reproducing their own computations without considerable agony."

MAKING SCIENTIFIC COMPUTATIONS REPRODUCIBLE

To verify a research paper's computational results, readers typically have to recreate them from scratch. ReDoc is a simple software filing system for authors that lets readers easily reproduce computational results using standardized rules and commands.

In the mid-1980s, we realized that our laboratory's researchers often had difficulty reproducing their own computations without considerable agony. We also noticed that junior students, who typically build on the work of more advanced students, frequently spent a great deal of time and effort just to reproduce their colleagues' computational results.

Reproducing computational research poses challenges in many environments. Indeed, the problem occurs wherever people use the traditional methods of scientific publication to describe computational research. For example, in a traditional article, the author simply outlines the relevant computations—the limitations of a paper medium prohibit complete documentation, which would ideally include experimental data, parameter values, and the author's programs. Readers who wish to use and verify the work must reimplement it, which is often a painful process. Even if readers have access to the author's source files (a feasible assumption given recent progress in electronic publishing) they can only recompute the results by invoking the various programs exactly as the author invoked them; such information is something that is usually undocumented and difficult to reconstruct.

To address these problems, we developed ReDoc, a system for reproducing scientific computations in electronic documents. Since implementing it in the early 1990s, ReDoc has become our principal means for organizing and transferring our laboratory's scientific computational research.

ReDocs are best defined operationally: After an author completes a ReDoc, readers can destroy all existing results—principally the illustrations—and rebuild them using the author's underlying programs and raw data. Using ReDoc, authors describe their computations and preserve their details in fully functional examples. Authors can also test their archived research software by occasionally removing and regenerating the document's results using ReDoc's standardized interface commands. ReDoc also lets authors develop automatic scripts to verify any document's completeness and reproducibility before its publication. Scientific journal publishers might also use ReDoc in the referee process to test the reproducibility of illustrations.

ReDoc benefits readers in several ways. Just as a driver wants to find the brake pedal in the

MATTHIAS SCHWAB, MARTIN KARRENBACH, AND JON CLAERBOUT
Stanford University

Northwestern Medicine®

Reproducible Research: Dynamic Documents

Synonymous with Dynamic Documents

Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis

Friedrich Leisch

Institut für Statistik und Wahrscheinlichkeitstheorie, Technische Universität Wien, Wiedner Hauptstraße 8-10/1071, A-1040 Wien, Austria

Abstract. Sweave combines typesetting with LaTeX and data anlysis with S into integrated statistical documents. When run through R or Splus, all data analysis output (tables, graphs, ...) is created on the fly and inserted into a final LaTeX document. Options control which parts of the original S code are shown to or hidden from the reader, respectively. Many S users are also LaTeX users, hence no new software has to be learned. The report can be automatically updated if data or analysis change, which allows for truly reproducible research.

Keywords. R, S, literate statistical practice, integrated statistical documents, reproducible research

1 Introduction

The traditional way of writing a report as part of a statistical data analysis project uses two separate steps: First, the data are analyzed using one's favorite statistical software package, and afterwards the results of the analysis (numbers, graphs, ...) are used as the basis for a written report. In larger projects the two steps may be repeated alternately, but the basic procedure remains the same. Many statistical software packages try to support this process by generating pre-formatted tables and graphics that can easily be integrated into a final report using copy-and-paste from the data analysis system to the word processor. The basic paradigm is to write the report around the results of the analysis.

Another approach for integration of data analysis and document writing is to embed the analysis itself into the document, which reverses the traditional paradigm. Over the last decade a number of systems have been developed that integrate analysis and documentation and allow for *literate statistical practice*, see Rossini (2001) for a survey.

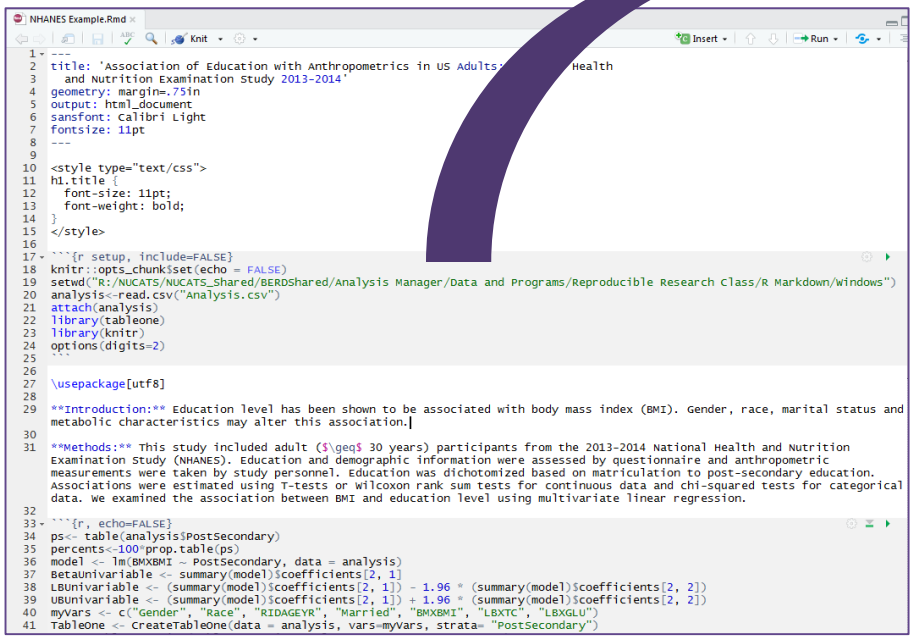
This new paradigm is probably most popular for creation of dynamic web pages and offers completely new possibilities for teaching statistics and delivering statistical methodology over the Internet. E.g., the ExploRe system (Härdle et al., 1999) provides means to embed statistical quantlets in web pages or electronic books to create interactive documents with direct access to a statistical data analysis package. Another example for a dynamic statistical analysis on a web page is given in Temple Lang (2001), by embedding R into netscape as a plugin. Report rendering is performed using XML and XSL.

We introduce a new system, called Sweave, which combines ideas from both worlds described above using literate programming tools. The purpose is to create dynamic reports, which can be updated automatically if data or analysis change, while using standard tools for both data analysis and word processing. Sweave is written in the S language, either the open source R (`http://www.R-project.org`) or the commercial Splus (`http://www.insightful.com`) can be used for statistical data analysis.

Sweave introduced in 2002. Reproducibility becomes synonymous with using dynamic documents that combine manuscript/report with code and data (e.g. Sweave, knitR).

Dynamic Documents: A Cornerstone of Reproducibility

Why we love them



Association of Education with Anthropometrics in US Adults: National Health and Nutrition Examination Study 2013-2014

Intro Education level has been shown to be associated with body mass index (BMI). ace, marital status and metabolic characteristics may alter this association.

Methods: dy included adult (≥ 30 years) participants from the 2013-2014 National Health and Examination Study (NHANES). Education and demographic information essed by questionnaire and anthropometric measurements were taken by stud nel. Education was dichotomized based on matriculation to post-secondary e sociations were estimated using T-tests or Wilcoxon rank sum tests for contin ata and chi-squared tests for categorical data. We examined the association betwe BMI and education level using multivariate linear regression.

Results: Among 4808 participants, 2649 (55.1%) self-reported any post-secondary education. Post-secondary education was associated with lower BMI (Beta: -0.62, 95% CI: -1.03 to -0.21). After adjusting for gender, race, age, marital status, fasting glucose and total cholesterol, post-secondary education was no longer significantly associated with BMI (Beta: -0.19, 95% CI: -0.79 to 0.41).

Table 1. Association of Education with Participant Characteristics among 2013-2014 NHANES Participants

	No Postsecondary	Postsecondary	p
n	2159	2649	
Gender = Male (%)	1079 (50.0)	1208 (45.6)	0.003
Race (%)			<0.001
Mexican American	456 (21.1)	173 (6.5)	
Non-Hispanic Asia	161 (7.5)	411 (15.5)	

- Numbers in document are updated automatically when data or models change.
- **Eliminates copying and pasting output.**
- Provides a link between a number in a manuscript and it's provenance.

Northwestern Medicine®

Reproducible Research: Sweave to R Markdown

Tools evolve: Sweave, knitR, R Markdown

```
\documentclass[a4paper]{article}

\title{Sweave Example 1}
\author{Friedrich Leisch}

\begin{document}

\maketitle

In this example we embed parts
\texttt{kruskal.test} help pag

<<>>=
data(airquality, package="data
library("stats")
kruskal.test(Ozone ~ Month, da
@
which shows that the location
distribution varies significan
include a boxplot of the data:

\begin{center}
<<fig=TRUE,echo=FALSE>>=
library("graphics")
boxplot(Ozone ~ Month, data =
@
\end{center}

\end{document}
```

R markdown example.Rmd

Knit Word | Run

```
---
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 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:

```{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. This is how I'm changing the data:

```{r newmean}
cars$speed[1]
cars$speed[1] <- 10
```

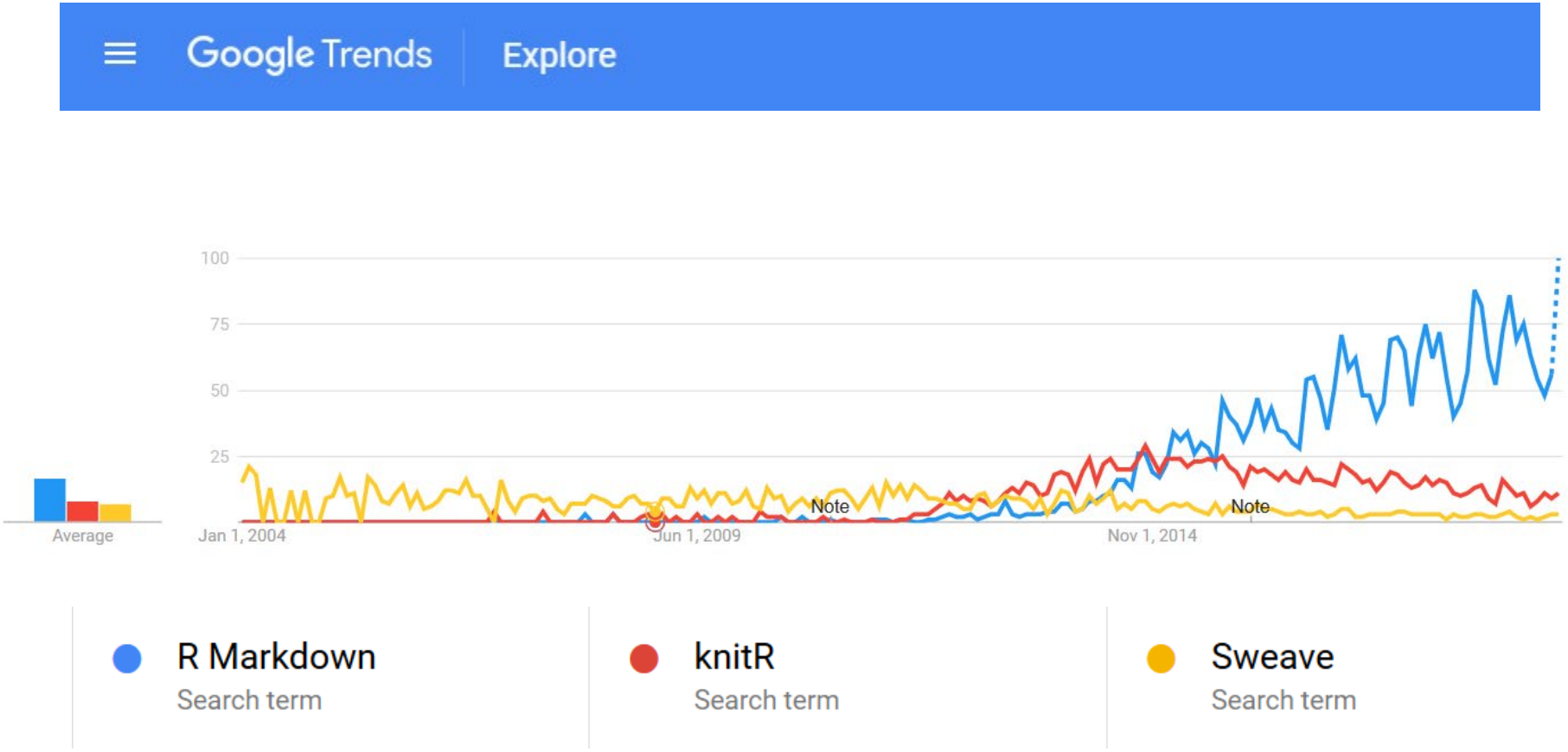
So now if I generate the mean speed, it is `r mean(cars$speed)`.

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]
```

Tools for Dynamic Documents and R



Reproducible Research: Incidents raise awareness

The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1309–1334
DOI: 10.1214/09-AOAS291

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

BY KEITH A. BAGGERLY[1] AND KEVIN R. CO

This article has been retracted

An Expression of Concern has been published for

Science 12 December 2014:
Vol. 346 no. 6215 pp. 1366-1369
DOI: 10.1126/science.1256151

REPORT

When contact changes minds: An experime for gay equality

Michael J. LaCour[1], Donald P. Green[2]

Author Affiliations

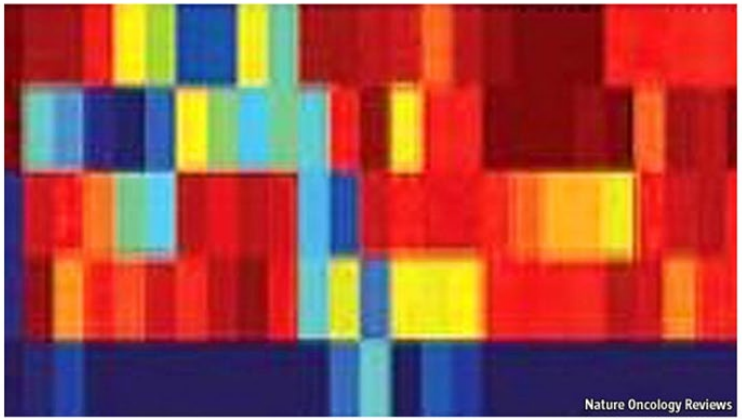
[1]Department of Political Science, University of California, Los Angele
[2]Department of Political Science, Columbia University, New York, NY

Misconduct in science

An array of errors

Investigations into a case of alleged scientific misconduct have revealed numerous holes in the oversight of science and scientific publishing

Sep 10th 2011 | From the print edition



ANIL POTTI, Joseph Nevins and their colleagues at Duke University in Durham, North Carolina, garnered widespread attention in 2006. They reported in the *New England Journal of Medicine* that they could predict the course of a patient's lung cancer using devices called expression arrays, which log the activity patterns of thousands of genes in a sample of tissue as a colourful picture (see above). A few months later, they wrote in *Nature Medicine* that they had developed a similar technique which used gene expression in laboratory cultures of cancer cells, known as cell lines, to predict which chemotherapy would be most effective for an individual patient suffering from lung, breast or ovarian cancer.

At the time, this work looked like a tremendous advance for personalised medicine—the idea that understanding the molecular specifics of an individual's illness will lead to a tailored

Northwestern Medicine®

Reproducible Research: NIH Statement

U.S. Department of Health & Human Services

NIH National Institutes of Health
Turning Discovery Into Health

Search NIH

NIH Employee Intranet | Staff Directory | En Español

Health Information | Grants & Funding | News & Events | Research & Training | Institutes at NIH | About NIH

Home » Research & Training

RIGOR AND REPRODUCIBILITY

Rigor and Reproducibility

- Principles and Guidelines
- Expanded Guidelines
- Application Instructions
- Training
- Funding Opportunities
- Meetings and Workshops
- Announcements
- Publications

Updated Application Instructions to Enhance Rigor and Reproducibility

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.

Johns Hopkins University students in a laboratory. *Johns Hopkins University*

In recent years, however, there has been a growing awareness of the need for rigorously designed published preclinical studies, to ensure that such studies can be reproduced. This webpage provides information about the efforts underway by NIH to enhance rigor and reproducibility in scientific research.

Email Updates

Sign up to receive email updates about rigor and reproducibility.

Sign up for updates

Related Links

Letter from Dr. Stephen I. Katz: An Update on the NIH Initiative to Enhance Research Rigor and Reproducibility

Contact Us

Please send email to NIHReprodEfforts@od.nih.gov.

Northwestern Medicine®

Reproducible Research: Beyond Biostatistics

Reproducible vs Replicable

Using the **same (raw) data** and information about the analysis methods and choices, we can recreate the results.

Scientific finding is verified or supported by **independent experiment or study**.

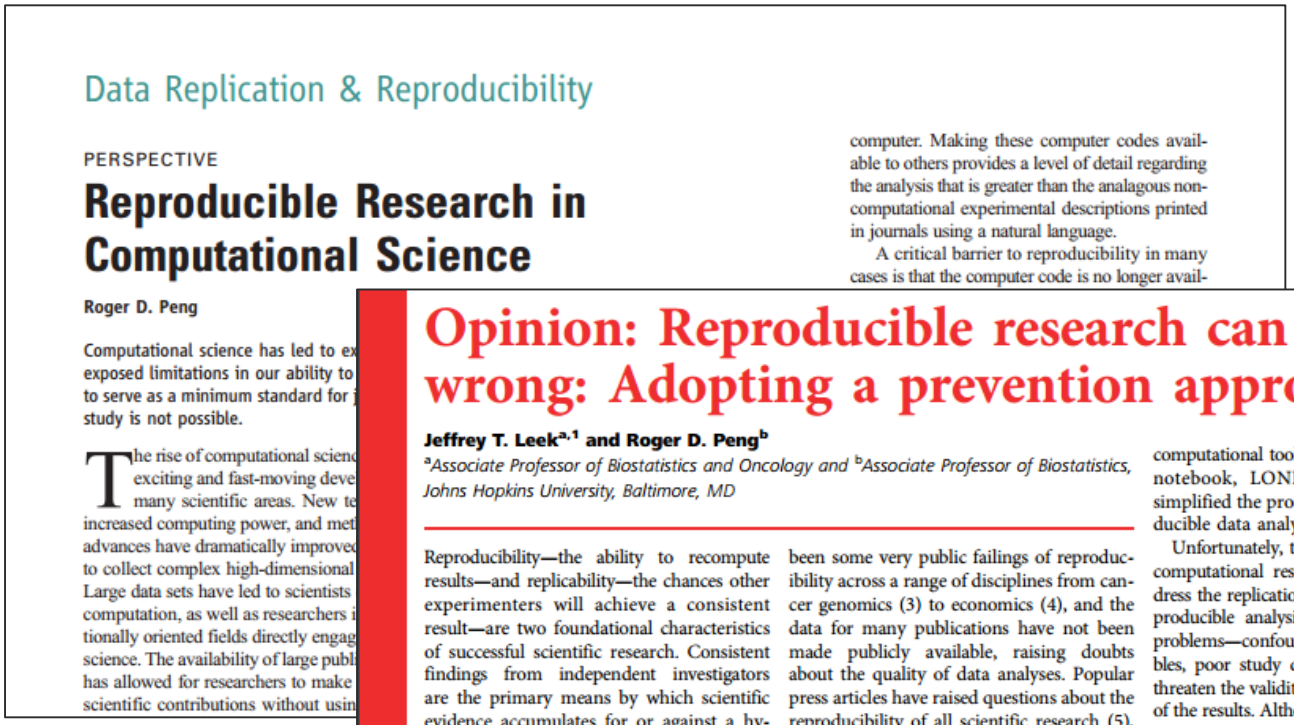
"We define reproducibility as the ability to re-compute data analytic results given an observed dataset and knowledge of the data analysis pipeline. The replicability of a study is the chance that an independent experiment targeting the same scientific question will produce a consistent result."

-- Leek and Peng 2015

Leek and Peng "Opinion: Reproducible Research can still be wrong: Adopting a prevention approach" PNAS, February 10, 2015, vol. 112, no. 6, 1645–1646

Reproducible Research: What can it really do?

Strengths and Limitations



Data Replication & Reproducibility

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

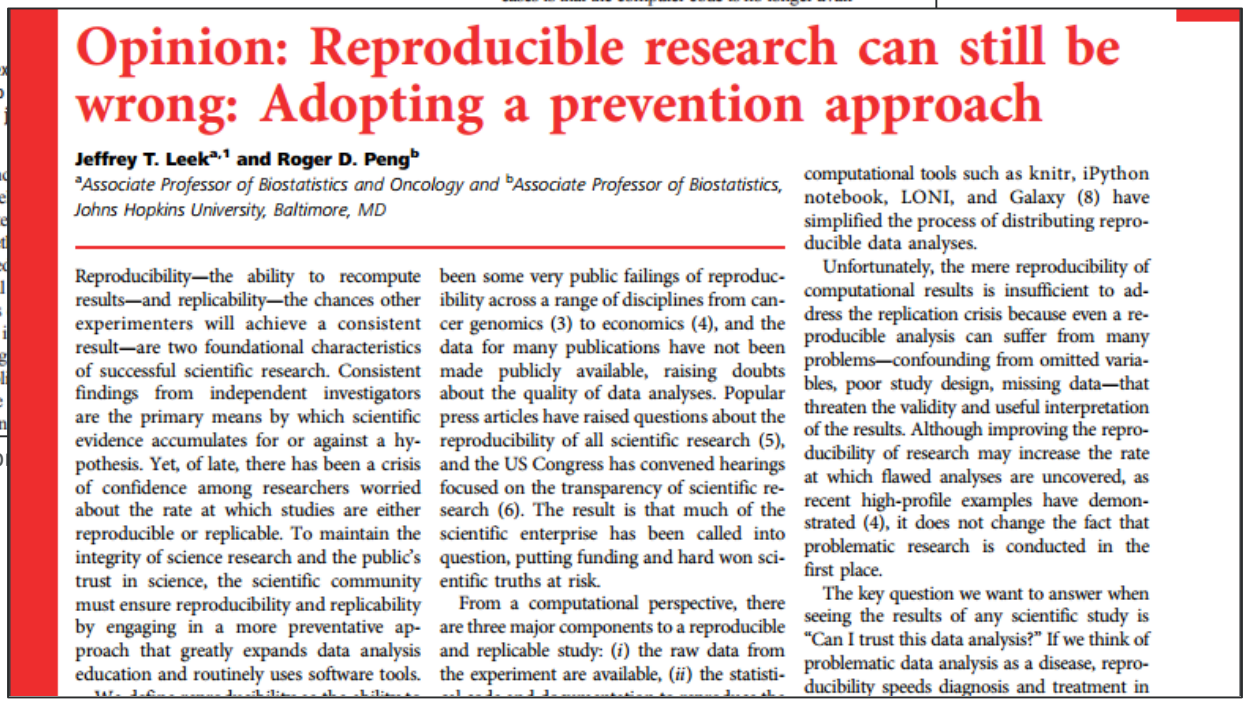
Computational science has led to ex
exposed limitations in our ability to
to serve as a minimum standard for
study is not possible.

The rise of computational scienc
exciting and fast-moving deve
many scientific areas. New te
increased computing power, and met
advances have dramatically improved
to collect complex high-dimensional
Large data sets have led to scientists
computation, as well as researchers i
tionally oriented fields directly engag
science. The availability of large publ
has allowed for researchers to make
scientific contributions without usin

computer. Making these computer codes available to others provides a level of detail regarding the analysis that is greater than the analagous non-computational experimental descriptions printed in journals using a natural language.

A critical barrier to reproducibility in many cases is that the computer code is no longer avail-

Peng, R "Rep



Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek[a,1] and Roger D. Peng[b]

[a]Associate Professor of Biostatistics and Oncology and [b]Associate Professor of Biostatistics, Johns Hopkins University, Baltimore, MD

Reproducibility—the ability to recompute results—and replicability—the chances other experimenters will achieve a consistent result—are two foundational characteristics of successful scientific research. Consistent findings from independent investigators are the primary means by which scientific evidence accumulates for or against a hypothesis. Yet, of late, there has been a crisis of confidence among researchers worried about the rate at which studies are either reproducible or replicable. To maintain the integrity of science research and the public's trust in science, the scientific community must ensure reproducibility and replicability by engaging in a more preventative approach that greatly expands data analysis education and routinely uses software tools.

been some very public failings of reproducibility across a range of disciplines from cancer genomics (3) to economics (4), and the data for many publications have not been made publicly available, raising doubts about the quality of data analyses. Popular press articles have raised questions about the reproducibility of all scientific research (5), and the US Congress has convened hearings focused on the transparency of scientific research (6). The result is that much of the scientific enterprise has been called into question, putting funding and hard won scientific truths at risk.

From a computational perspective, there are three major components to a reproducible and replicable study: (*i*) the raw data from the experiment are available, (*ii*) the statisti-

computational tools such as knitr, iPython notebook, LONI, and Galaxy (8) have simplified the process of distributing reproducible data analyses.

Unfortunately, the mere reproducibility of computational results is insufficient to address the replication crisis because even a reproducible analysis can suffer from many problems—confounding from omitted variables, poor study design, missing data—that threaten the validity and useful interpretation of the results. Although improving the reproducibility of research may increase the rate at which flawed analyses are uncovered, as recent high-profile examples have demonstrated (4), it does not change the fact that problematic research is conducted in the first place.

The key question we want to answer when seeing the results of any scientific study is "Can I trust this data analysis?" If we think of problematic data analysis as a disease, reproducibility speeds diagnosis and treatment in

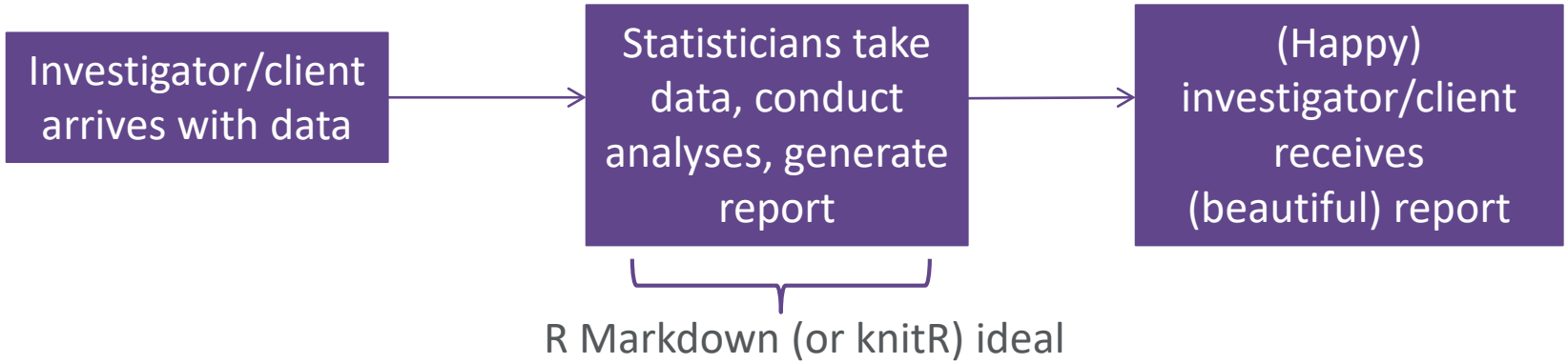
Leek and Peng "Opinion: Reproducible Research can still be wrong: Adopting a prevention approach" PNAS, February 10, 2015, vol. 112, no. 6, 1645–1646

Northwestern Medicine®

Dynamic Documents: Challenges for Collaboration

Reproducible Research: Role of the Statistician

Challenges with Collaborations and Consultations



Reproducible +

Investigator/client comes back in a year with questions, we still know what we did

Investigator/client comes back with updated data, no problem to update

Reproducible –

Reproducibility only exists (to our knowledge) in the middle step

What if we have greater responsibilities and involvement?

MS Word is Ubiquitous for Manuscript Preparation

Clinical and Basic Sciences

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."

R Markdown and Microsoft Word

.Rmd file can generate a .rtf (.docx) file

NHANES Example.Rmd

Knit

```
---
title: 'Association of Education with Anthropometr
  and Nutrition Examination Study 2013-2014'
geometry: margin=.75in
output: html_document
sansfont: Calibri Light
fontsize: 11pt
---

<style type="text/css">
h1.title {
  font-size: 11pt;
  font-weight: bold;
}
</style>

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE)
setwd("R:/NUCATS/NUCATS_Shared/BERDShared/Analysis
analysis<-read.csv("Analysis.csv")
attach(analysis)
library(tableone)
library(knitr)
options(digits=2)
```

\usepackage[utf8]

Introduction: Education level has been shown t
metabolic characteristics may alter this associati

Methods: This study included adult ($\geq$ 30
Examination Study (NHANES). Education and demograp
measurements were taken by study personnel. Educat
Associations were estimated using T-tests or Wilc
data. We examined the association between BMI and

```{r, echo=FALSE}
ps<- table(analysis$PostSecondary)
percents<-100*prop.table(ps)
model <- lm(BMXBMI ~ PostSecondary, data = analys
BetaUnivariable <- summary(model)$coefficients[2,
LBUnivariable <- (summary(model)$coefficients[2, 1
UBUnivariable <- (summary(model)$coefficients[2, 1
myVars <- c("Gender", "Race", "RIDAGEYR", "Marriec
TableOne <- CreateTableOne(data = analysis, vars=m
```

Association of Education with Anthropometrics in US Adults: National Health and Nutrition Examination Study 2013-2014

Introduction: Education level has been shown to be associated with body mass index (BMI). Gender, race, marital status and metabolic characteristics may alter this association.

Methods: This study included adult (≥ 30 years) participants from the 2013-2014 National Health and Nutrition Examination Study (NHANES). Education and demographic information were assessed by questionnaire and anthropometric measurements were taken by study personnel. Education was dichotomized based on matriculation to post-secondary education. Associations were estimated using T-tests or Wilcoxon rank sum tests for continuous data and chi-squared tests for categorical data. We examined the association between BMI and education level using multivariate linear regression.

Results: Among 4808 participants, 2649 (55.1%) self-reported any post-secondary education. Post-secondary education was associated with lower BMI (Beta: -0.62, 95% CI: -1.03 to -0.21). After adjusting for gender, race, age, marital status, fasting glucose and total cholesterol, post-secondary education was no longer significantly associated with BMI (Beta: -0.19, 95% CI: -0.79 to 0.41).

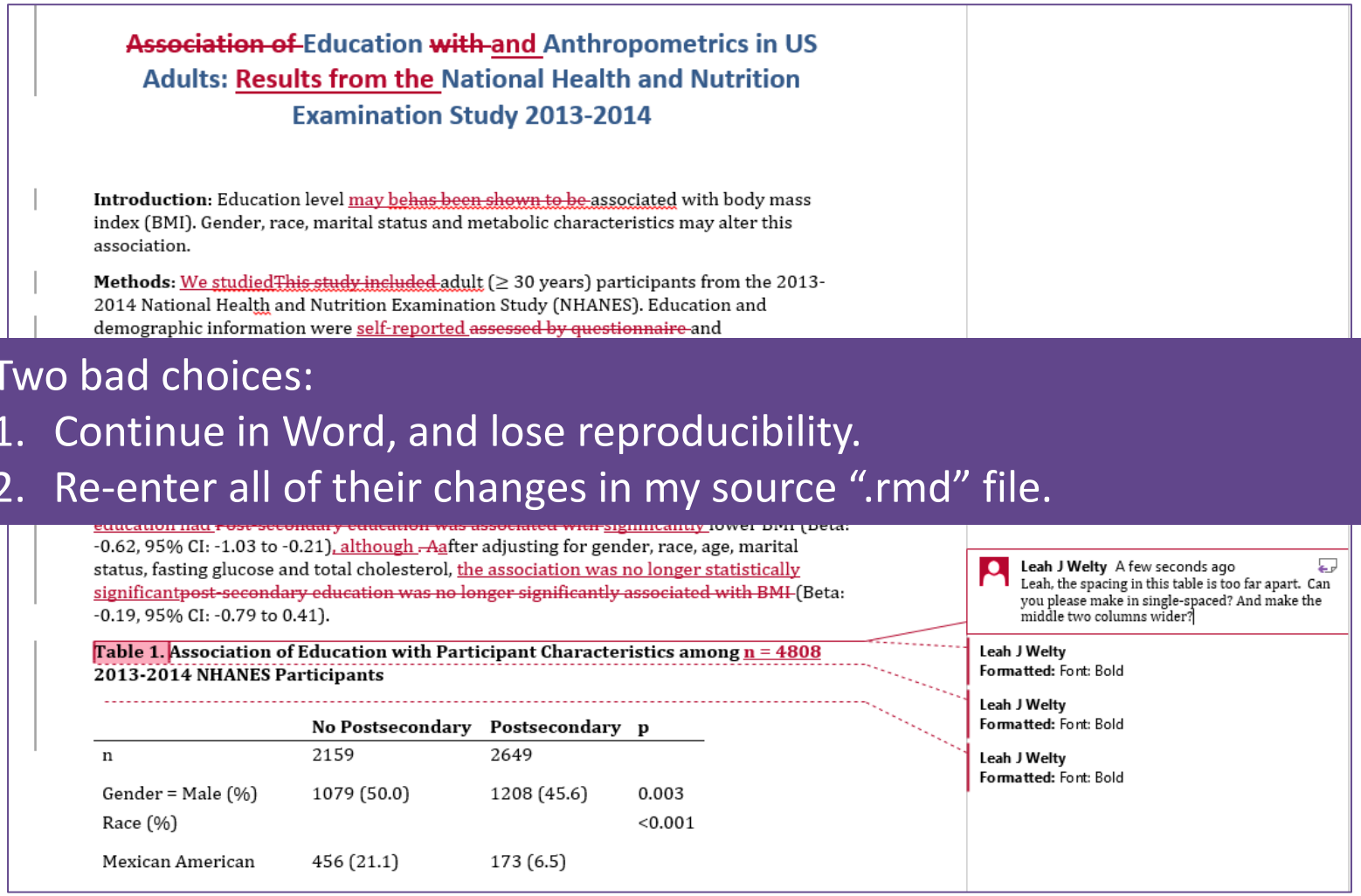
Table 1. Association of Education with Participant Characteristics among 2013-2014 NHANES Participants

	No Postsecondary	Postsecondary	p
n	2159	2649	
Gender = Male (%)	1079 (50.0)	1208 (45.6)	0.003
Race (%)			<0.001
Mexican American	456 (21.1)	173 (6.5)	
Non-Hispanic Asia	161 (7.5)	411 (15.5)	

Northwestern Medicine®
```

Practical Limitations

I send my collaborators the dynamic document, they send back:



~~Association of~~ Education ~~with~~ and Anthropometrics in US Adults: Results from the National Health and Nutrition Examination Study 2013-2014

Introduction: Education level may ~~be~~has been shown to be associated with body mass index (BMI). Gender, race, marital status and metabolic characteristics may alter this association.

Methods: We studied~~This study included~~ adult (≥ 30 years) participants from the 2013-2014 National Health and Nutrition Examination Study (NHANES). Education and demographic information were self-reported ~~assessed by questionnaire~~ and

Two bad choices:

1. Continue in Word, and lose reproducibility.
2. Re-enter all of their changes in my source ".rmd" file.

~~education had Post-secondary education was associated with significantly~~ lower BMI (Beta: -0.62, 95% CI: -1.03 to -0.21), although ~~,~~ Aafter adjusting for gender, race, age, marital status, fasting glucose and total cholesterol, the association was no longer statistically significant~~post-secondary education was no longer significantly associated with BMI~~ (Beta: -0.19, 95% CI: -0.79 to 0.41).

Table 1. Association of Education with Participant Characteristics among n = 4808 2013-2014 NHANES Participants

	No Postsecondary	Postsecondary	p
n	2159	2649	
Gender = Male (%)	1079 (50.0)	1208 (45.6)	0.003
Race (%)			<0.001
Mexican American	456 (21.1)	173 (6.5)	

Leah J Welty A few seconds ago
Leah, the spacing in this table is too far apart. Can you please make in single-spaced? And make the middle two columns wider?

Leah J Welty
Formatted: Font: Bold

Leah J Welty
Formatted: Font: Bold

Leah J Welty
Formatted: Font: Bold

Northwestern Medicine®

Practical Limitations

Collaborators and plain text editors

Markdown and knitR require drafting documents in a plain text editor:

```
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)`.

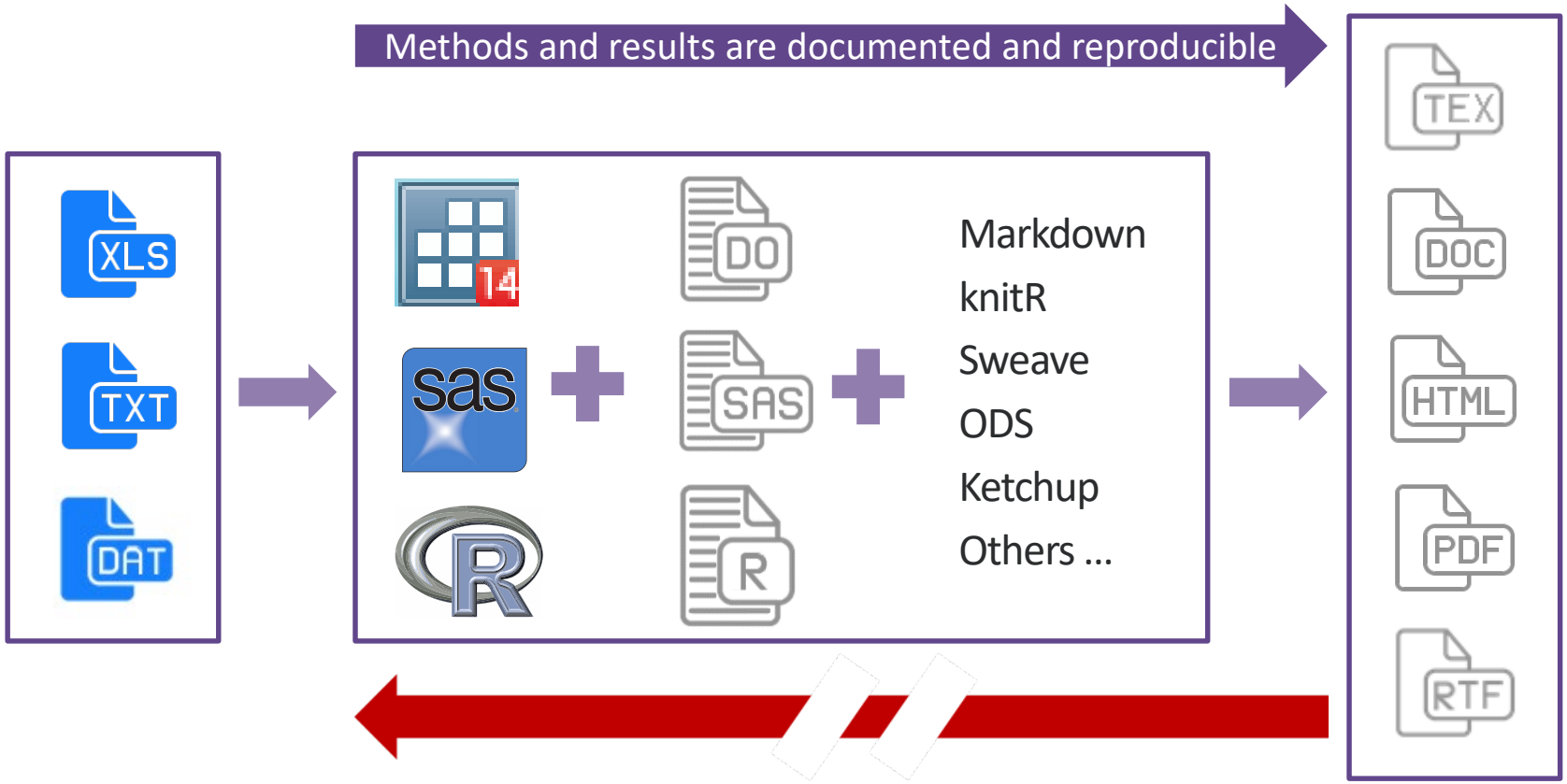
That's pretty nice, because if I change something about the data, then that number can be automatically updated.

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

Are your collaborators willing to work this way?
My collaborators (primarily doctors and social scientists) are not.

Dynamic Documents: Existing Tools

Limitations of Existing Tools for Dynamic Documents



Why not R Markdown?

Markdown is GREAT, but not a universal solution for reproducibility

1. Collaborators may be unwilling to work in plain text.

2. Many (clinical) journals prefer/require submissions in MS Word.

3. Markdown can generate an RTF, but any changes to it must be re-entered in the source file.

4. Weaving tools (Markdown, Sweave, knitR) work with one software program at a time (e.g. R).

5. It can be messy to have statistical code and manuscript text all in the same place.

Northwestern Medicine®

StatTag: Reproducible Research using Microsoft Word

Reproducible Research using Microsoft Word

Searching for tools in 2012

J Mesirov "Accessible Reproducible Research" *Science* 2010

- Collaboration with Microsoft for Gene Pattern Software
- Very limited use

Nolan, Peng and Lang 2011 "Enhanced Dynamic Documents for Reproducible Research"

- RWordXML: cut and paste R code in to Word document
- R package to process it and create a pdf or html file

Stata Automation Report

- Word plug-in
- Only for Stata
- Worked only in some circumstances

Finding a team and funding

Solving a pain-point takes resources

StatTag: A complementary approach

Incorporate existing strengths, pragmatic

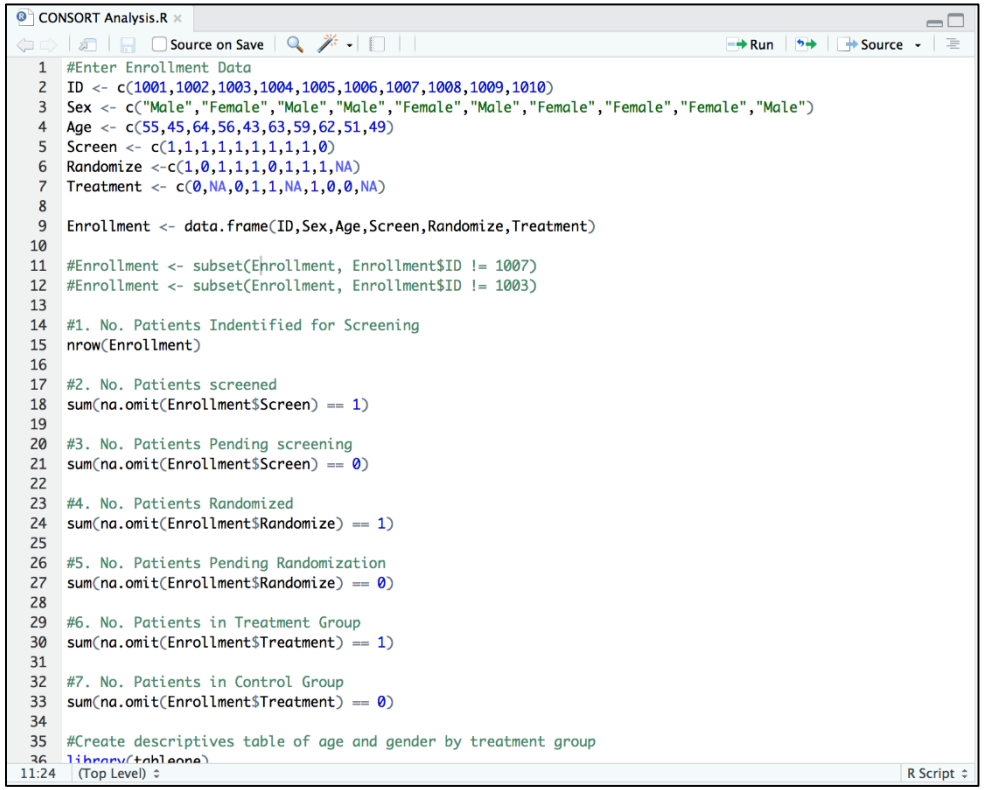
- StatTag creates a link between a statistical code file and a Word document
- Embeds output (values, tables, figures, verbatim) in document
- Can work separately on the code and the Word document but retain link
- Software agnostic: connects R, R Markdown (or Stata, SAS) to Word document
- Can connect multiple code files to the same document

Northwestern Medicine®

StatTag: How it Works

StatTag

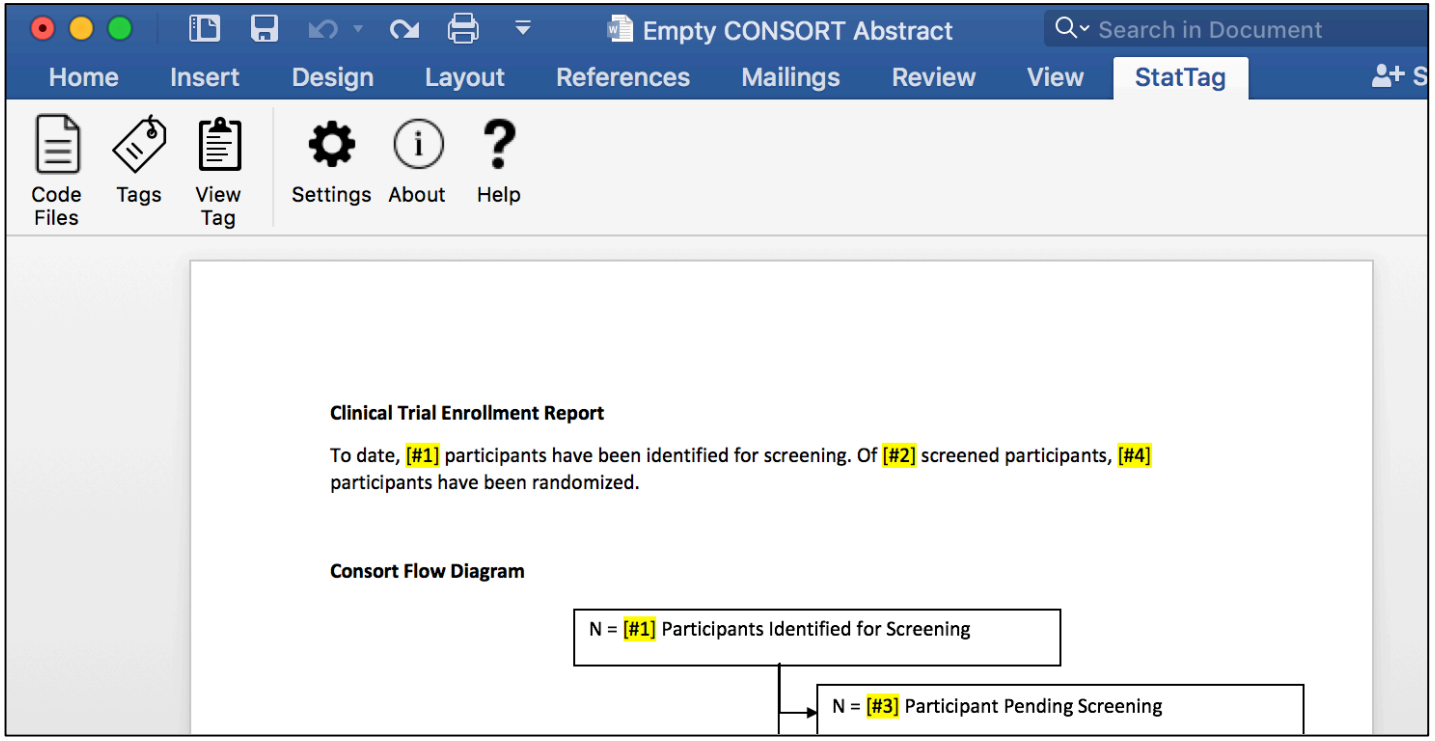
Step 1: Write your R/R Markdown code.



Northwestern Medicine®

StatTag: How it Works

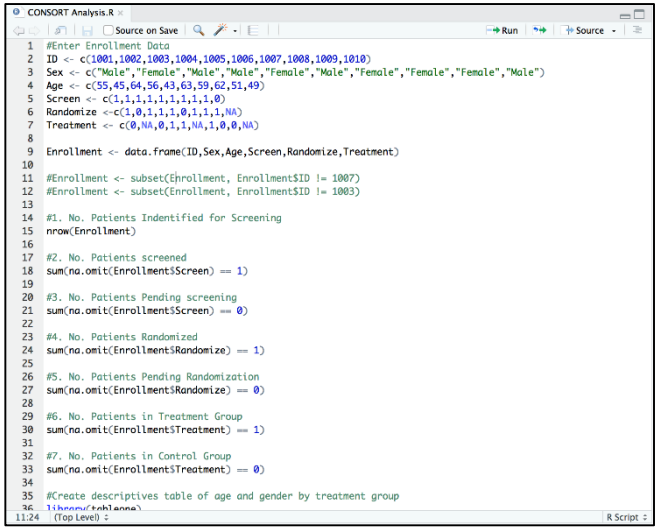
Step 2: Open Word and write some or all of your text.

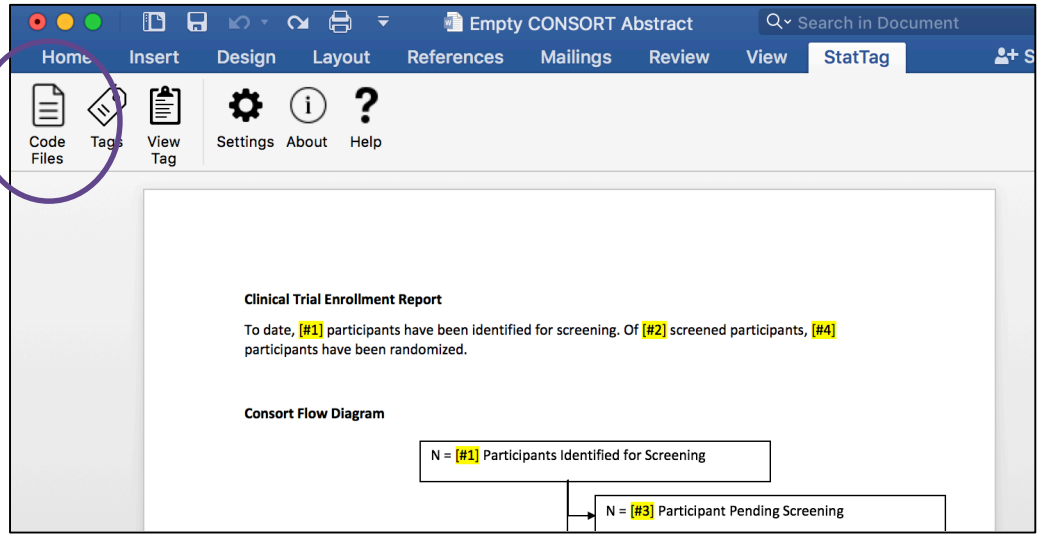


Northwestern Medicine®

StatTag: How it Works

Step 3: Use StatTag to connect your code file to the Word document.

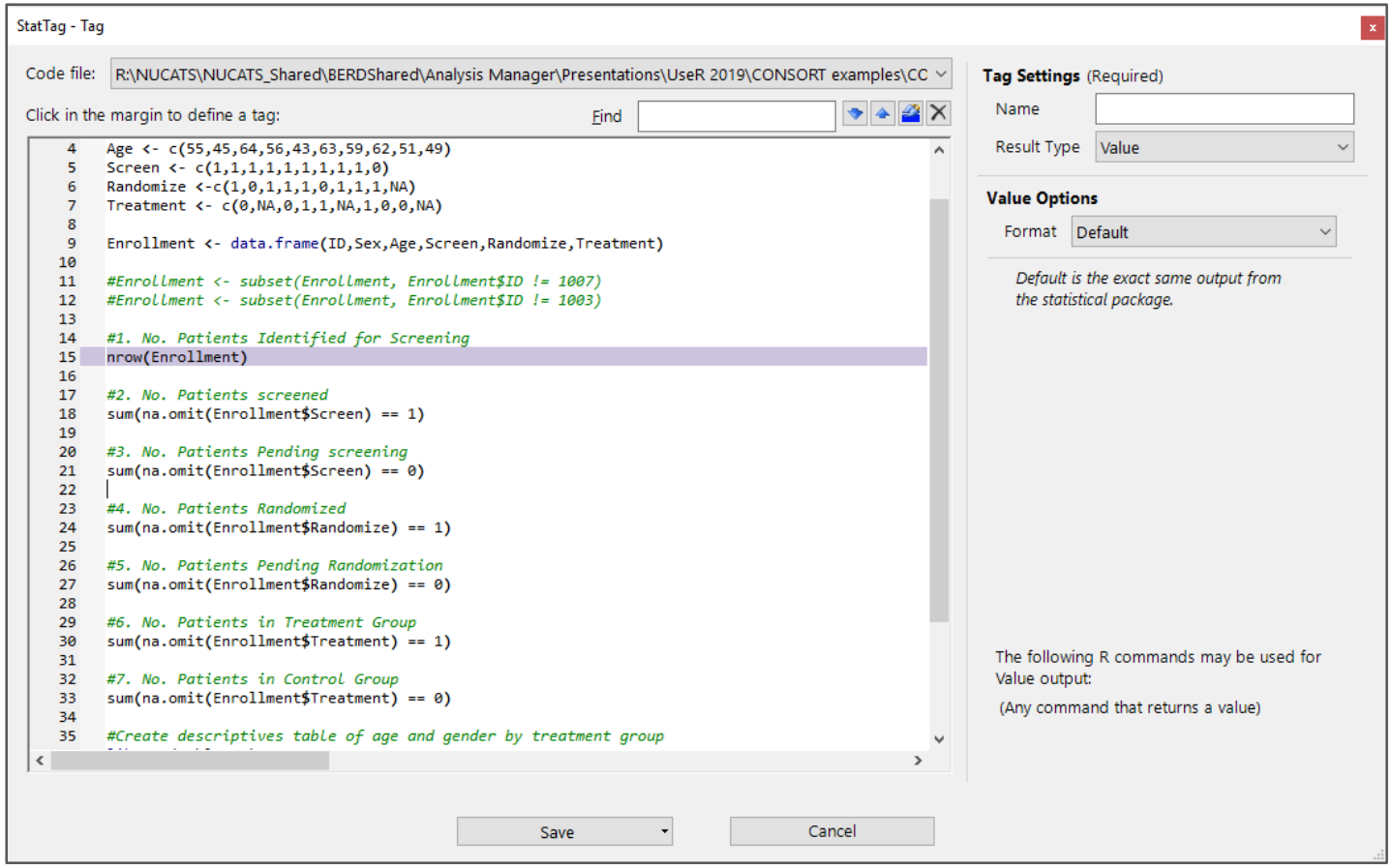




Northwestern Medicine®

StatTag: How it Works

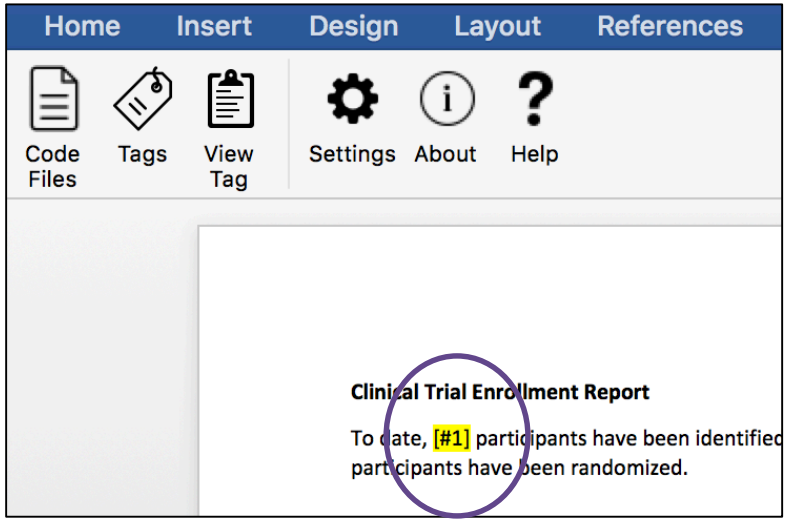
Step 4: Use StatTag to identify "tags" – portions of output that you want to insert – using a pop-up from Word.

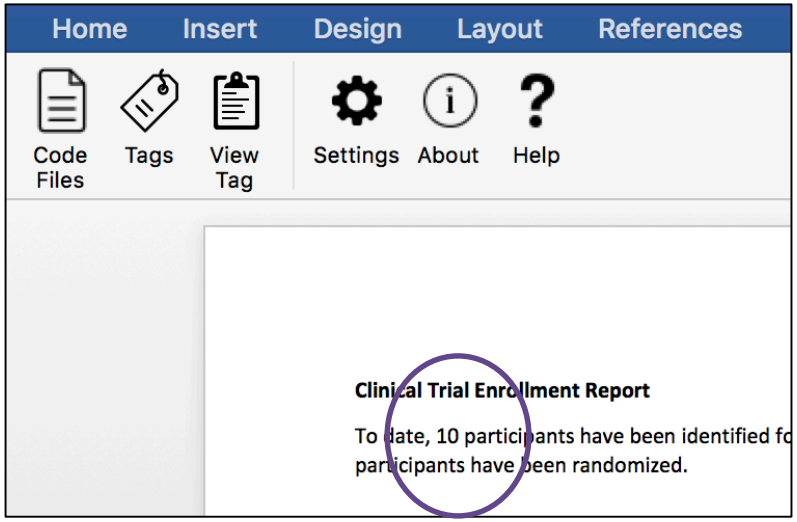


Northwestern Medicine®

StatTag: How it Works

Step 5: Use StatTag to insert “tags” in your document. StatTag will run the code in the background and insert the results in to Word.



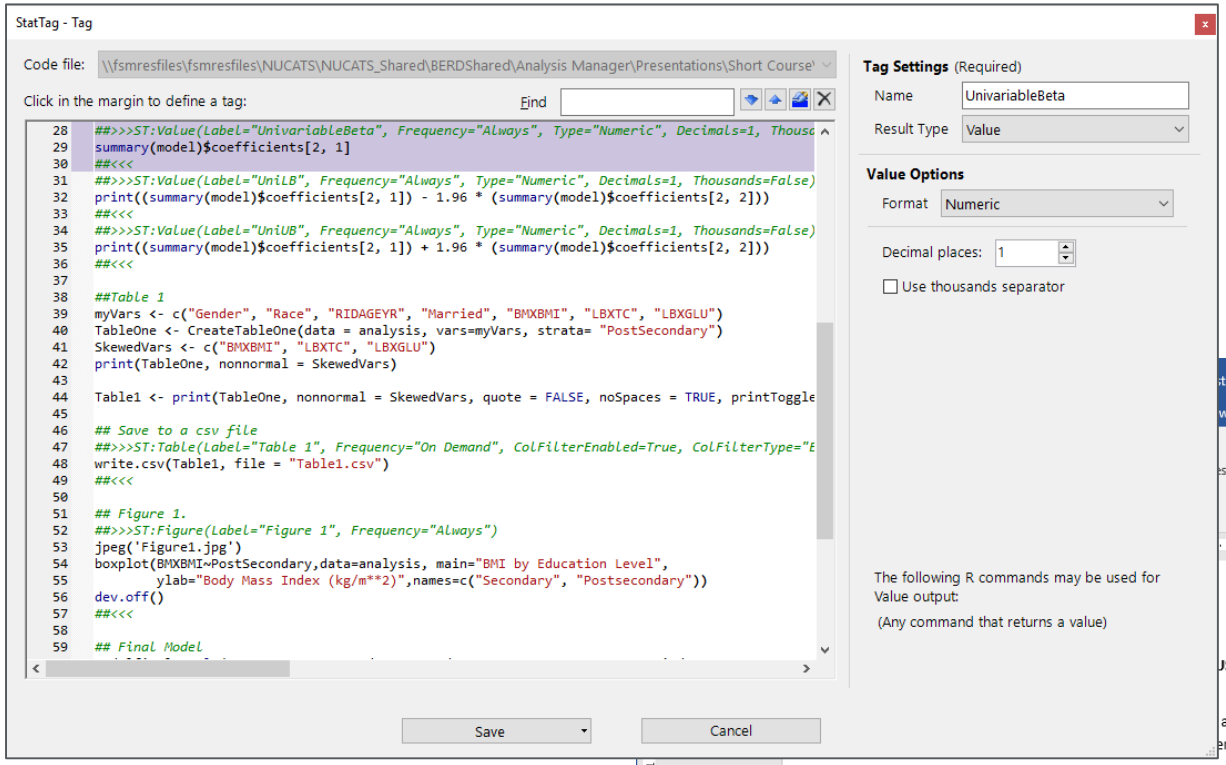


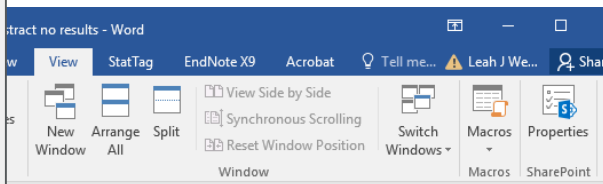
StatTag Demo video at Open StatTag on YouTube:

https://www.youtube.com/watch?v=K3QwG4LB9a4

StatTag: How it Works

Double-clicking a tag brings up associated code





US Adults: National Health and Nutrition Examination Study

associated with body mass index (BMI). Gender, race,
er this association.

Methods: This study included a 20 years) participants from the 2013-2014 National Health and Nutrition Examination Study (NHANES). d demographic information were assessed by questionnaire and anthropometric measurement by study personnel. Education was dichotomized based on matriculation to post-second sociations were estimated using T-tests or Wilcoxon rank sum tests for continuous data and chi-s for categorical data. We examined the association between BMI and education level using multivariat ession.

Results: Among 4808 participants, 2649 self-reported any post-secondary education. Post-secondary education was associated with lower BMI (Beta: -0.6, 95% CI: -1.0 to -0.2). After adjusting for gender, race, age, marital status, fasting glucose and total cholesterol, post-secondary education was no longer significantly associated with BMI (Beta:-0.2, 95% CI: -0.8 to 0.4).

Table 1. Association of Education with Participant Characteristics among 2013-2014 NHANES Participants

Characteristic	No Post-Secondary n=2159	Post-Secondary n=2649	P-Value
Gender = Male (%)	1079 (50.0)	1208 (45.6)	0.003
Race (%)			<0.001
Mexican American	456 (21.1)	173 (6.5)	

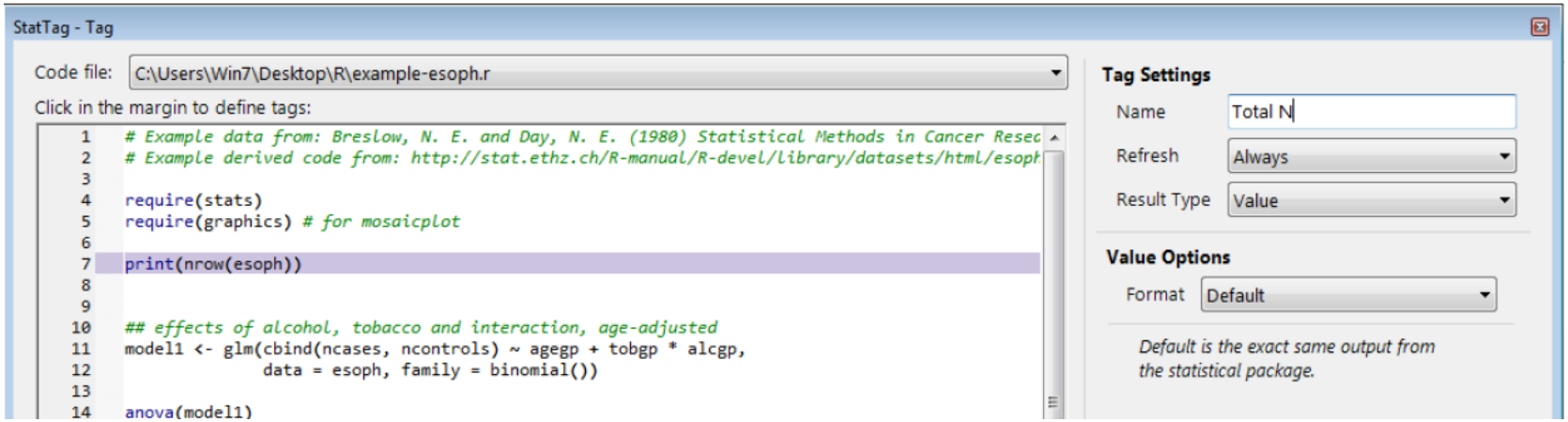
Northwestern Medicine®

StatTag: How it works

Inserting tags directly in code, or using StatTag dialog box

```
**>>>ST:Value(Label=" ", Frequency="", Type="")
[R, SAS, or Stata code]
**<<<

**>>>ST:Table(Label="", Frequency="", Type="", AllowInvalid=True,
Decimals=0, Thousands=False)
[R, SAS, or Stata code]
**<<<
```



Northwestern Medicine®

StatTag: How it Works

Recognizes different key words to identify results.

Type			 or 
Numeric Values	display	%put	Any command that returns a value
Tables	matrix list, any command that returns a .xls or .csv file	ODS CSV	Any command that returns a data frame, matrix, vector or list, any command that returns a .xls or .csv file
Figures	graph export	ODS PDF	pdf, win.metafile, png, jpeg, bmp, postscript
Verbatim	Any Code	Any Code	Any Code

StatTag: How it Works

Sharing Word documents (and code) with collaborators

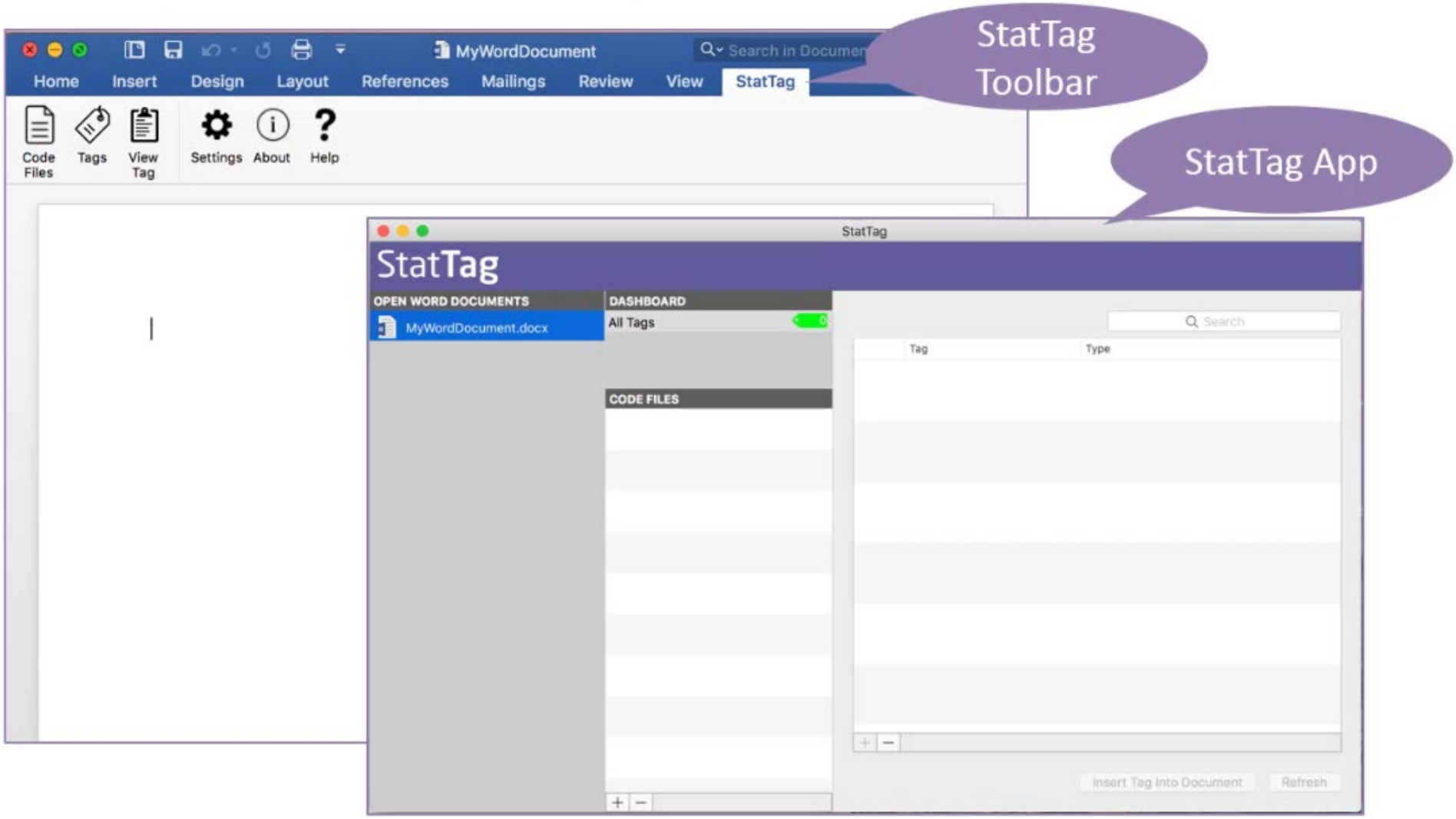
StatTag

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®

StatTag: How it Works (Mac)

App, rather than plug-in

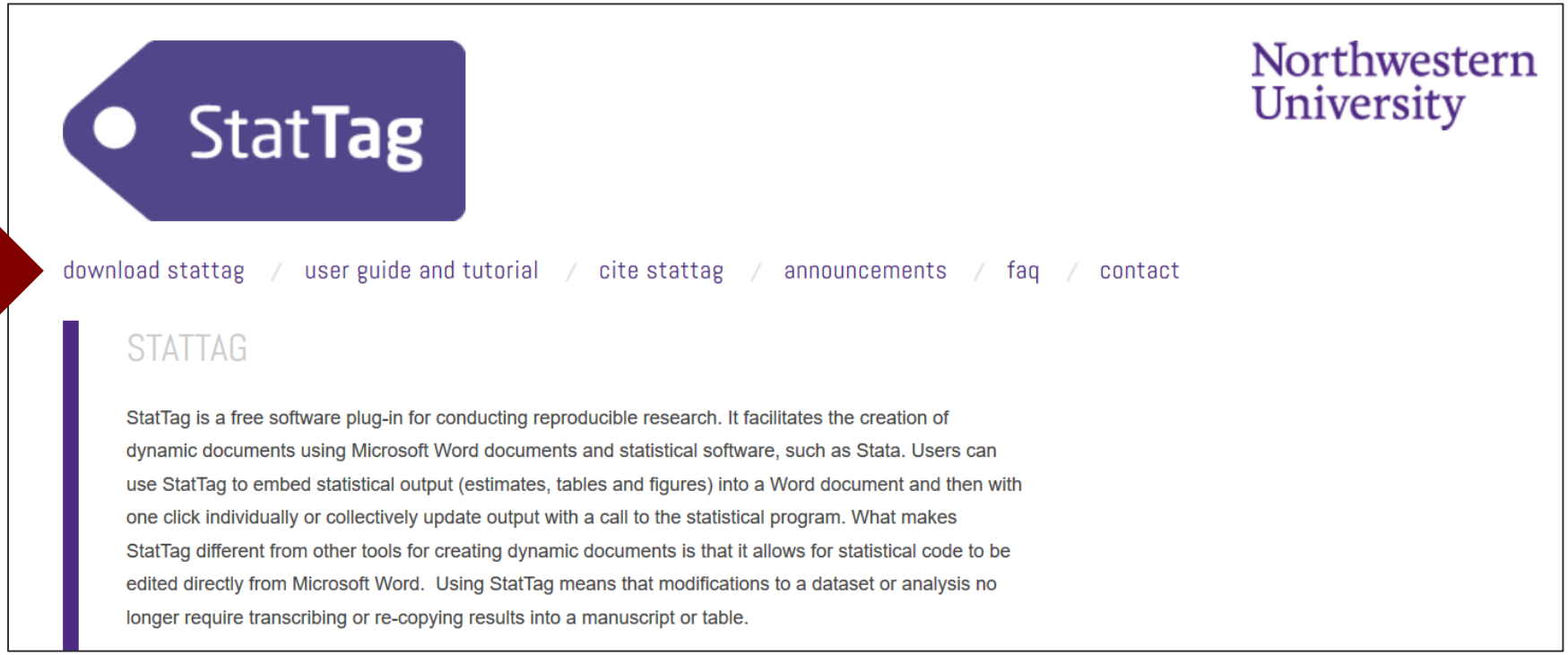


Northwestern Medicine®

Getting StatTag

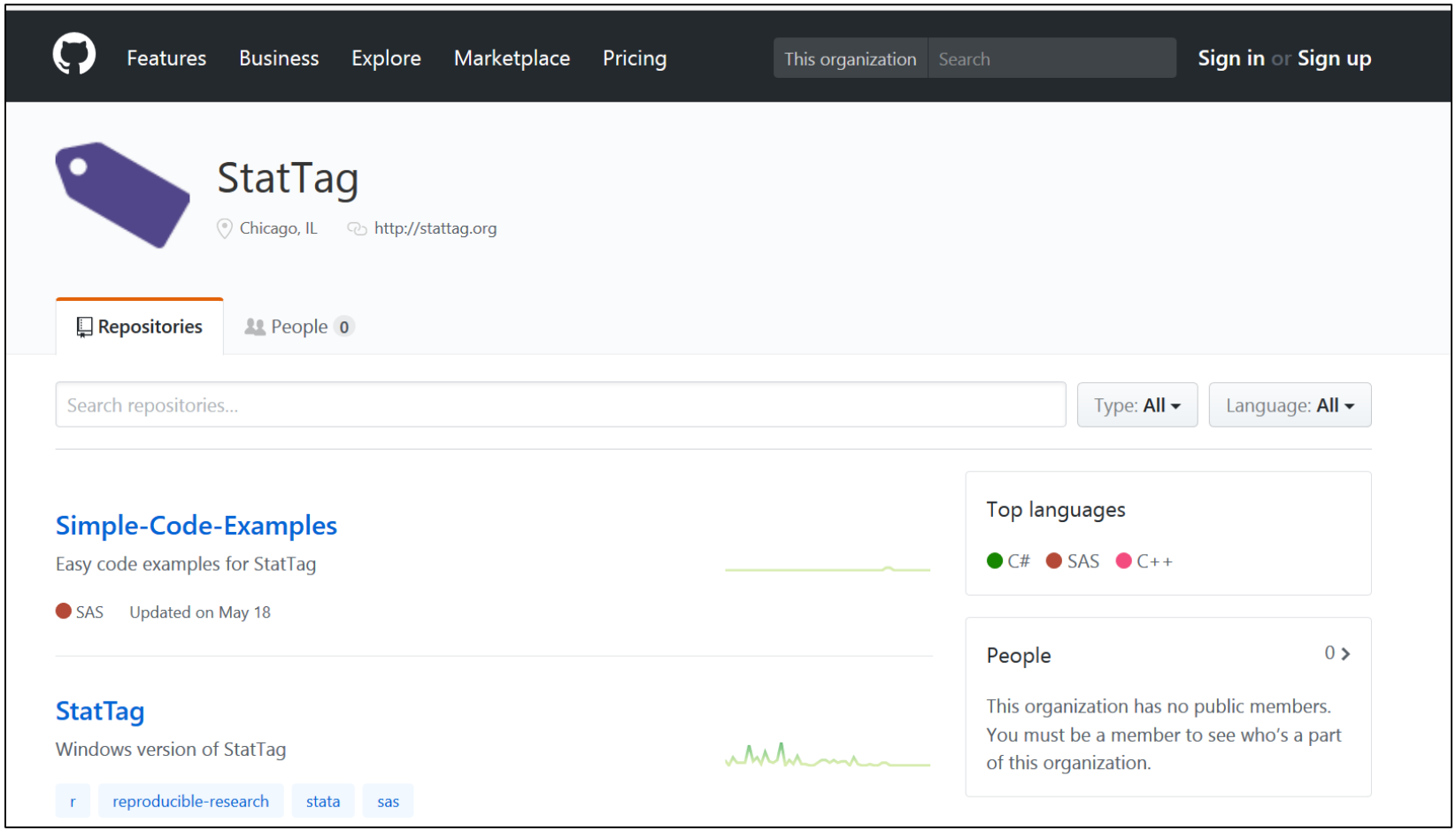
Freely available, open source, and evolving

stattag.org



(Really) Getting StatTag

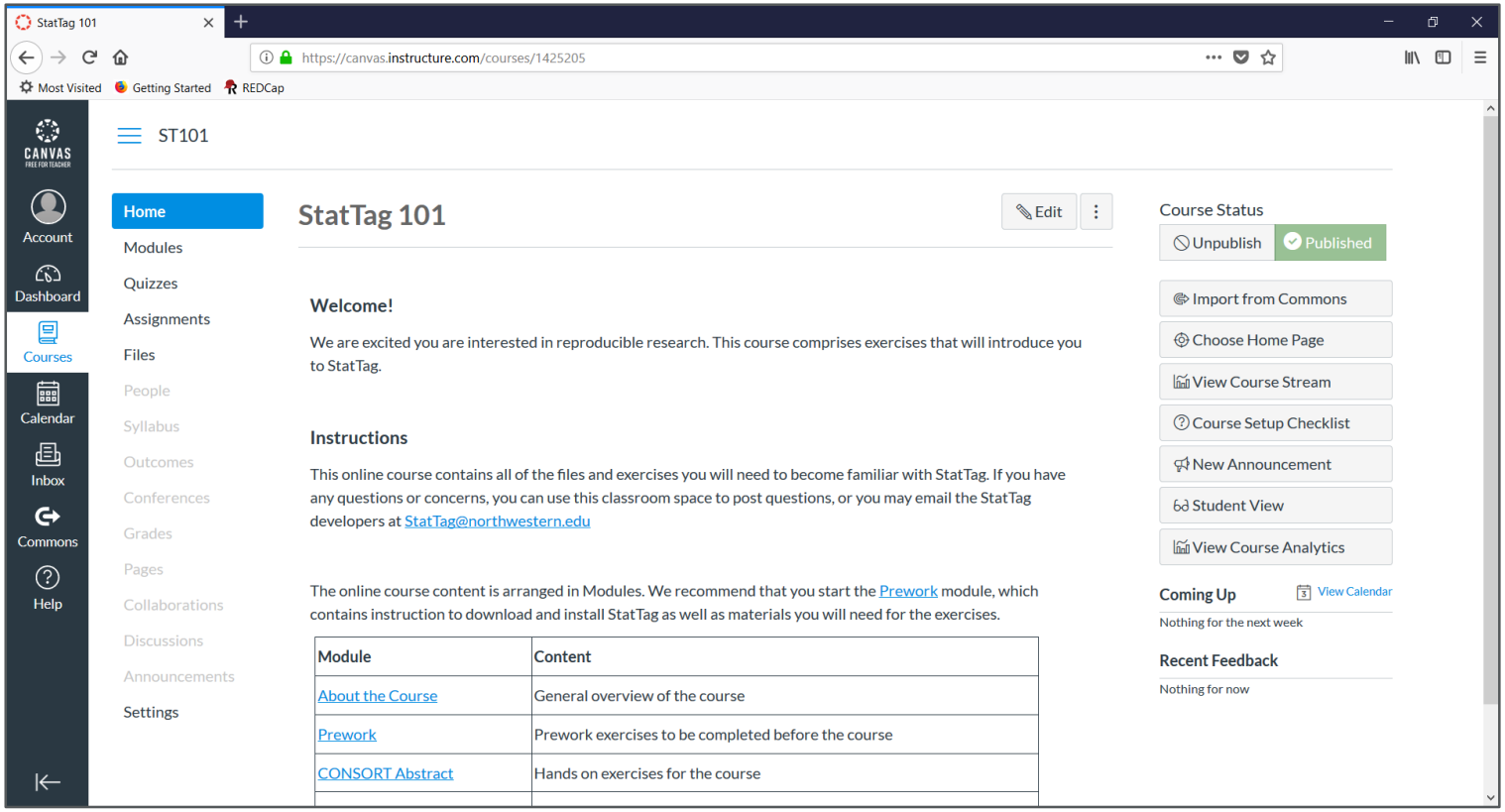
GitHub repo at github.com/StatTag



Northwestern Medicine®

Learning StatTag

Open course on Instructure Canvas



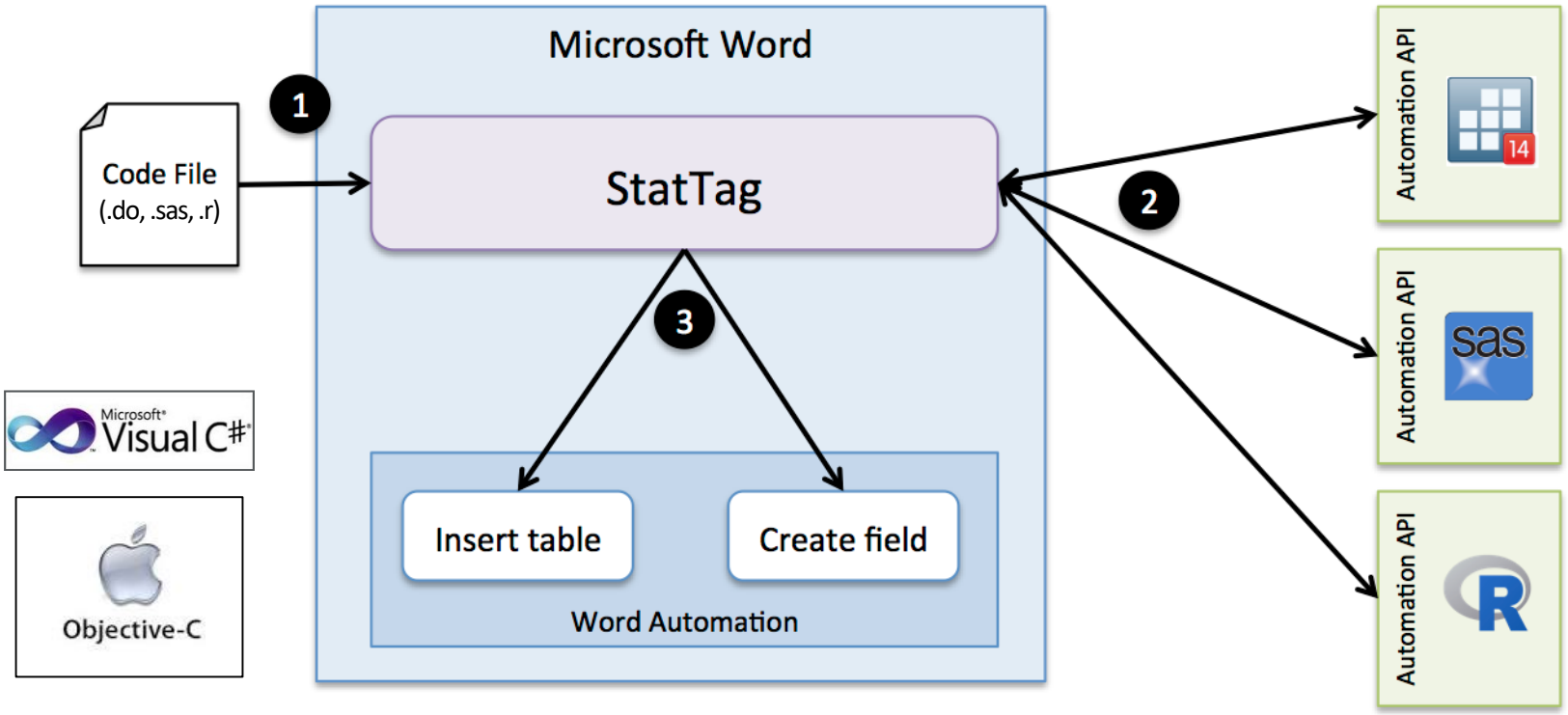
Online course available at: https://canvas.instructure.com/enroll/DPCLGC

Or enter code DPCLGC at https://canvas.instructure.com/register

Northwestern Medicine®

StatTag: Architecture

How we built it



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).

Future directions

Tech savvy tools to adopt for collaborative reproducibility

- Connecting to even more
  - Python
  - Jupyter notebook

- Version Control
  - Github
  - Subversion, cvs, etc.
  - Great tools, but what about WYSISYG or non-plain text files?

- "Containers" (e.g. docker)
  - Facilitate packaging the code, data, packages, libraries and software program/environment together

Citations and Acknowledgements

Thank you!

- 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 being developed with funding through a Clinical Translational Sciences Award (CTSA) to Northwestern University.

Northwestern Medicine | NUCATS Clinical and Translational Sciences Institute

CTSA Clinical & Translational Science Awards

Northwestern Medicine

StatTag

Questions?

- Visit stattag.org for more information
- Get involved! Contact us at stattag@northwestern.edu
  - Questions
  - Comments and suggestions
  - Volunteer to test new releases – ***especially if you are also a Python user!***

Northwestern Medicine®