Reproducible Health Research Pipelines

Leslie McIntosh Borrelli, PHD, MPH
Anthony Juehne, MPH
Cynthia Hudson Vitale, MLIS

What is the problem?

50-85%
of research
resources
wasted

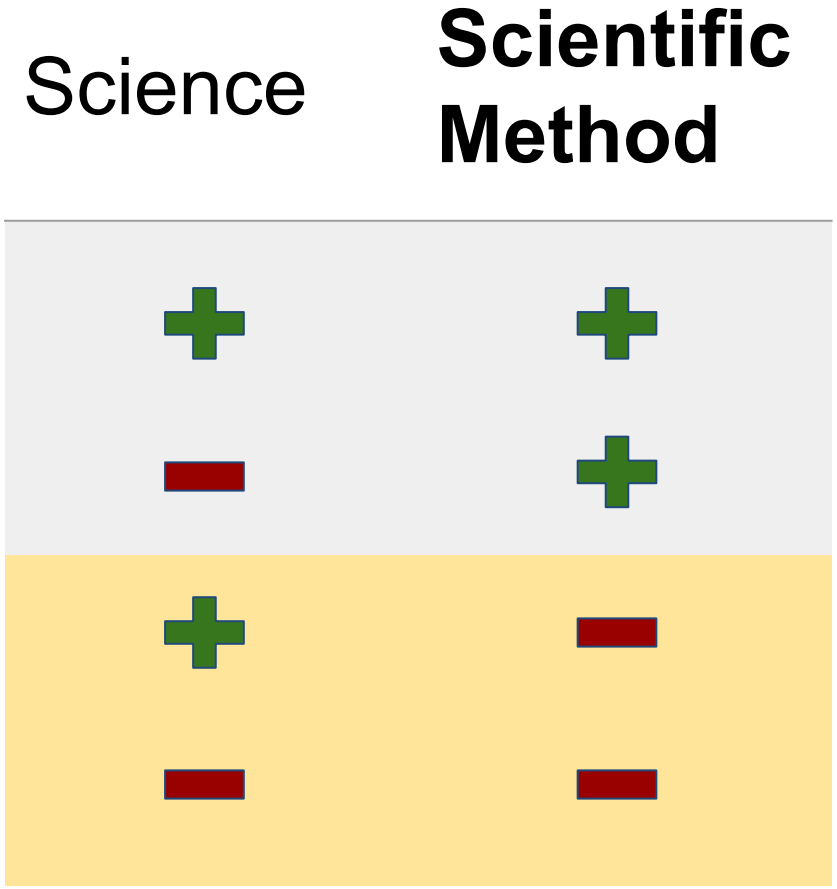
$28 Billion

Reproducibility

How are we doing this?

Scientific Method

RicochetConcept



Work-to-Date

RicochetOutputs

Publication

Repeat: A Framework to Assess Empirical Reproducibility in Biomedical Research

BioMed Central (accepted 2017)

Repos:

https://github.com/ripeta

https://osf.io/ppnwa/

International Engagement

Research Data Alliance P10 - Montreal

September 19-21, 2017

Evaluating Reproducibility: **Ricochet**byripeta

RicochetFramework

RicochetAutomator

RicochetScore

RicochetFramework

RicochetFramework

100+ unique variables

5 categories

- bibliographic
- database & data collection
- data mining & cleaning
- data analysis
- data sharing & documentation

RicochetFramework

Publication Overview and Bibliographic Information (21 items)	
Is the research hypothesis-driven or hypothesis-generating?	Hypothesis Driven Hypothesis Generating Unclear
Database and Data Collection (63 items)	
Publication states database(s) source(s) of data?	Yes/No
*Publication states database(s) source(s) of data in the following location:	Not Stated Supplementary materials Appendix Body of Text
Query methodology	Manual extraction Digital extraction through query interface Digital extraction through honest broker Not Applicable/Not Stated
*Does the shared query script for database contain comments and/or notations for ease of reproducibility?	Yes/No
Methods: Data Mining and Cleaning (19 items)	
Does the research involve natural language processing or text mining?	Yes/No
*Is the text mining software application proprietary or open? *If multiple applications were used, please select all options that apply.*	1. Proprietary 2. Mixed 3. Open

RicochetFramework

Methods: Data Analysis (15 items)	
Does the author state analysis methodology and process?	Yes/No
Does the author indicate the software used to develop the analysis code?	Yes/No
*Is the analysis software proprietary or open?	Proprietary Open
Data Sharing and Data Documentation (36 items)	
Is the finalized dataset shared?	Yes No
*Where is the finalized dataset shared?	Affiliated Research Center Website Author's Institution or Department Website Data Registry Journal or Publication's Website GitHub Other
Is there a clear process for requesting the data?	Yes No

Ricochet Framework has been tested for inter-rater reliability and face validity

RicochetExample

Predicting warfarin dosage in European–Americans and African–Americans using DNA samples linked to an electronic health record

Pharmacogenomics. 2012 Mar; 13(4): 407–418.
Published online 2012 Feb 13. doi: 10.2217/pgs.11.164

Hypothesis Stated = **Yes**

Here we test: whether the published associations between steady-state warfarin dose and variants in warfarin pharmacogenes could be replicated in BioVU; how implementing published pharmacogenomic algorithms affects dosing error; and if an improved algorithm for African–Americans can be generated using variants associated with stable dose in this population.

RicochetExample

Predicting warfarin dosage in European–Americans and African–Americans using DNA samples linked to an electronic health record

Cases were identified in BioVU, the Vanderbilt DNA Biobank, which accrues DNA samples extracted from blood remaining from routine clinical testing after it has been retained for 3 days and is scheduled to be discarded [27].

Data Source Stated = **Yes**

Data Source Cited = **Yes**

RicochetExample

Predicting warfarin dosage in European–Americans and African–Americans using DNA samples linked to an electronic health record

The R programming language was used for regression analyses, diagnostic-test calculations, and to implement and evaluate the algorithms (R Foundation for Statistical Computing, Vienna, Austria).

Software Stated = **Yes**
Software Cited = **Yes**
Software Version = **No**

RicochetExample

Predicting warfarin dosage in European–Americans and African–Americans using DNA samples linked to an electronic health record

Supplementary Table 4. Equations of novel *Expanded Genetic* algorithm.

Expanded Genetic:

5.9487517
- 0.0073436353 * race (AA=0,EA=1)
- 0.025161445 * age (in years)
+ 0.058138499 * sex (F=0,M=1)
+ 1.1848957 * bsa (kg/m^2)
+ 0.068020571 * smoking status (nonsmoker=0,smoker=1)
+ 0.058578086 * VTE indication (no=0,yes=1)
- 0.10646416 * atrial fibrillation indication (no=0,yes=1)
- 0.8142521 * amiodarone use (no=0,yes=1)
- 0.64877338 * *CYP2C9*2* (wt=0,heterozygote=1,homozygote=2)
- 1.0601067 * *CYP2C9*3* (wt=0,heterozygote=1,homozygote=2)
- 1.9737831 * *CYP2C9*6* (wt=0,heterozygote=1,homozygote=2)
- 1.0622944 * *CYP2C9*8* (wt=0,heterozygote=1,homozygote=2)
+ 0.24749973 * *CYP4F2* (wt=0,heterozygote=1,homozygote=2)
- 0.31996754 * *CALU* (wt=0,heterozygote=1,homozygote=2)
- 0.87262446 * *VKORC1* (wt=0,heterozygote=1,homozygote=2)
= log[weekly warfarin dose]

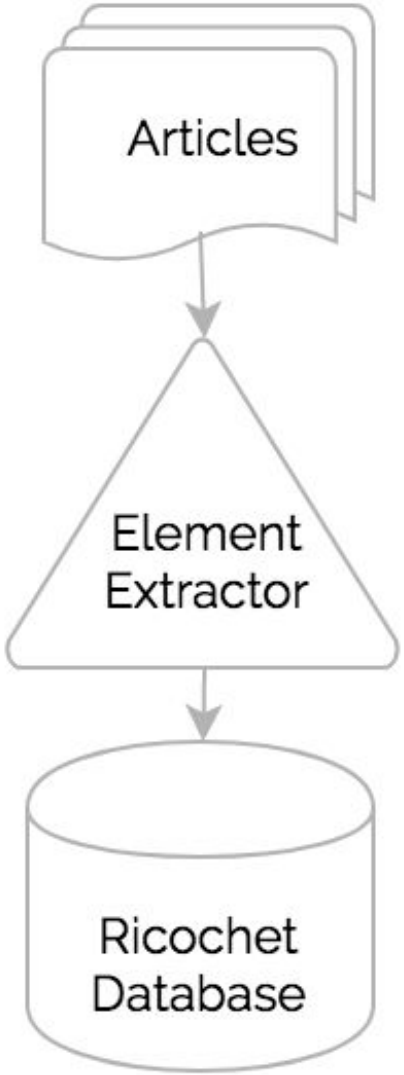
Analyses Code Present = **Partially**

RicochetResults

id	article doi	journal publication	publication date	state data sources	state database location	database cited	analysis software
14	10.1186/s13054-015-1180-6	Critical care (London,	1/15/16	1	0	0	SPSS
17	10.1016/j.urology.2012.11.002	Urology	1/30/13	1	0	0	SAS
24	10.1097/AJP.0b013e3181f06b06	Clinical Journal of Pai	1/27/11	1	0	0	STATA
25	10.1016/j.jmig.2011.01.009	Journal of Minimally I	3/16/11	1	0	0	SAS
26	10.1097/IOP.0b013e31828a92b0	Ophthalmic Plastic &	1/6/13	0	0		
27	10.1097/PSY.0b013e31821fbf9a	Psychosomatic Medic	6/23/11	0	0	0	
33	10.1016/j.clineuro.2014.02.016	Clinical Neurology an	2/25/14	0	0	0	SPSS
35	10.1016/j.jpainsymman.2015.02.022	Journal of Pain and S	4/1/15	0	0	0	
37	10.1148/radiol.14132418	Radiology	9/9/14	0	0	0	R
39	10.1111/apt.13505	Alimentary pharmacc	2/4/16	1	0	0	SAS
44	10.1016/j.burns.2013.12.002	Burns: Journal of the	1/14/14	1	0	0	IBM SPSS
45	10.1016/j.clineuro.2015.10.004	Clinical Neurology an	10/8/15	1	0	0	SPSS
46	10.1016/j.ygyno.2013.04.055	Gynecologic Oncolog	4/28/13	1	0	0	GraphPad
47	10.1016/j.jaapos.2014.06.006	Journal of American A	9/27/14	1	0	0	
49	10.1016/j.jstrokecerebrovasdis.2015.06.043	Journal of Stroke and	6/27/15	1	0	1	SPSS
50	10.1186/1472-6963-13-414	BMC health services r	10/14/13	1	0	1	SAS
54	10.1016/j.joms.2016.01.001	Journal of Oral and M	1/7/16	0	0	0	
55	PMC4384267	Pharmacy practice	3/15/15	1	1	0	
56	10.1186/1471-2318-14-75	BMC Geriatrics	6/17/14	1	0	0	STATA
58	10.1016/j.clinthera.2011.08.005	Clinical Therapeutics	9/3/11	1	0	0	SPSS
60	10.3399/bjgp15X686113	British Journal of Gen	8/1/15	1	0	1	STATA
63	10.1158/1055-9965.EPI-14-0487	Cancer Epidemiology,	7/30/14	1	0	0	SAS
66	10.1542/hpeds.2014-0085	Hospital Pediatrics	8/1/15	1	0	0	SPSS;SAS

RicochetAutomator

RicochetAutomator



RepeATAutomator **Prototype**	**Ricochet**Automator
Interface via Python command line	Easy to use web interface
Limited variables automatically detected	Extensive automation extraction
Designed for use by *ripeta* collaborators only	Designed for use by a diverse array of researchers and publishers
Single document analysis	Single or batch document analyses
No documentation for use as a library, Python-only API	Library/client design, well-documented JSON web API
Storage in proprietary clinical study database, usage requires API key	Storage in generic database, analysis requires no permissions
Exclusive focus on EHR studies	Ability to upload new data dictionaries and allow users to choose among many scientific domains

Django REST framework

Category List

Category List

OPTIONS GET

```
GET /api/v0/categories
```

```
HTTP 200 OK
Allow: GET, POST, HEAD, OPTIONS
Content-Type: application/json
Vary: Accept
```


```
[
    {
        "name": "bibliographic",
        "description": "Bibliographic Information",
        "order": 0
    },
    {
        "name": "data-collection",
        "description": "Database & Data Collection",
        "order": 1
    },
    {
        "name": "data-mining",
        "description": "Data Mining & Cleaning",
        "order": 2
    },
    {
        "name": "data-analysis",
        "description": "Data Analysis",
        "order": 3
    },
    {
        "name": "data-sharing",
        "description": "Data Sharing & Data Documentation",
        "order": 4
    }
]
```

Django REST framework

Extract

Extract

OPTIONS GET

Extract variables from a Paper. A single variable can be specified using its primary key, or all variables will be extracted.

TODO: multiple

```
GET /api/v0/extract/doi:def
```

```
HTTP 200 OK
Allow: GET, HEAD, OPTIONS
Content-Type: application/json
Vary: Accept

{
    "funding": [
        "the Canadian Institutes of Health Research",
        "This study was funded by a grant (CT62962) from the Canadian Institutes of Health Research (CIHR) and a CIHR/Wyeth Postdoctoral Fellowship Award to
    ],
    "grant_id": [
        "CT62962",
        "This study was funded by a grant (CT62962) from the Canadian Institutes of Health Research (CIHR) and a CIHR/Wyeth Postdoctoral Fellowship Award to
    ]
}
```

Ricochetbyripeta

-- analyzing research reproducibility

Upload a Research Paper (pdf)

Upload a Research Paper

It will be parsed and analyzed for research reproducibility.

RicochetFrontend

Upload a Research Paper

Unique id

doi:10.1186-1471-2318-14-75

A unique identifier for the paper, prefixed with doi:, isbn:, or pmid:

Title

Age and gender differences in the prevalence and patterns of multimorbidity in the older population

Authors

José Maŕi a Abad-Dĭ ez1, Amaia Calderó n-Larrañ aga, Antonio Poncel-Falcó , Beatriz Poblador-Plou, José Manuel Calderó n-Meza, Antoni Sicras-Mainar,

File

Browse... art%3A10.1186%2F1471-2318-14-75.pdf

***PDF file* is required.**

Upload Paper Cancel

RicochetFrontend

Research Papers

Title	Id		
Elevated Troponin Levels in Acute Stroke Patients Predict Long-term Mortality	doi:10.1016-j.jstrokecerebrovasdis.2015.06.043		Results
Differences between intentional and non-intentional burns in India: Implications for prevention	doi:10.1016-j.burns.2013.12.002		Results
Selective Versus Comprehensive Neck Dissection in Patients With T1 and T2 Oral Squamous Cell Carcinoma and cN0pN+ Neck	doi:10.1016-j.joms.2016.01.001		Results

Upload another research paper

API raw JSON for the first paper.

RicochetFrontend

Research Paper Results

Title: *Age and gender differences in the prevalence and patterns of multimorbidity in the older population*

Id: *doi:10.1186-1471-2318-14-75*

Category results

+	Bibliographic	See More
−	Database & Data Collection	See More
< >	Data Mining & Cleaning	See More
< >	Methods: Data Analysis	See More
−	Data Sharing & Data Documentation	See More

Download Ricochet Reproducibility Report

RicochetFrontend

Category results

+	Bibliographic	See More
−	Database & Data Collection	See More
< >	Data Mining & Cleaning	See More
< >	Methods: Data Analysis	See More
−	Data Sharing & Data Documentation	See More

Variables for *Bibliographic*

GOOD

	Question (variable)	Extracted Result	User Input *	
+	Article DOI (article_doi)	Available at http://dx.doi.org/10.1787/9789264122314-en.		+
+	Where did the funding for this study come from? (funding)	Acknowledgements The study was funded by Grant PI11/01126 from the Carlos III Health Institute.		+
+	What was the ID of the grant that funded this study? (grant_id)	Acknowledgements The study was funded by Grant PI11/01126 from the Carlos III Health Institute.		+

RicochetFrontend

Category results

+	Bibliographic	See More
−	Database & Data Collection	See More
< >	Data Mining & Cleaning	See More
< >	Methods: Data Analysis	See More
−	Data Sharing & Data Documentation	See More

Variables for *Methods: Data Analysis*

FAIR

	Question (variable)	Extracted Result	User Input *	
+	Software used for data analysis (analysis_sw)	STATA 12 was used for the statistical analysis and Excel 2007 for developing the graphs.		+
−	Does the author state analysis methodology and process? (analysis_process_clear)			+

* If you think the extractor missed some of the paper data, please enter it manually. Thanks!

Download Ricochet Reproducibility Report

RicochetFrontend

Category results

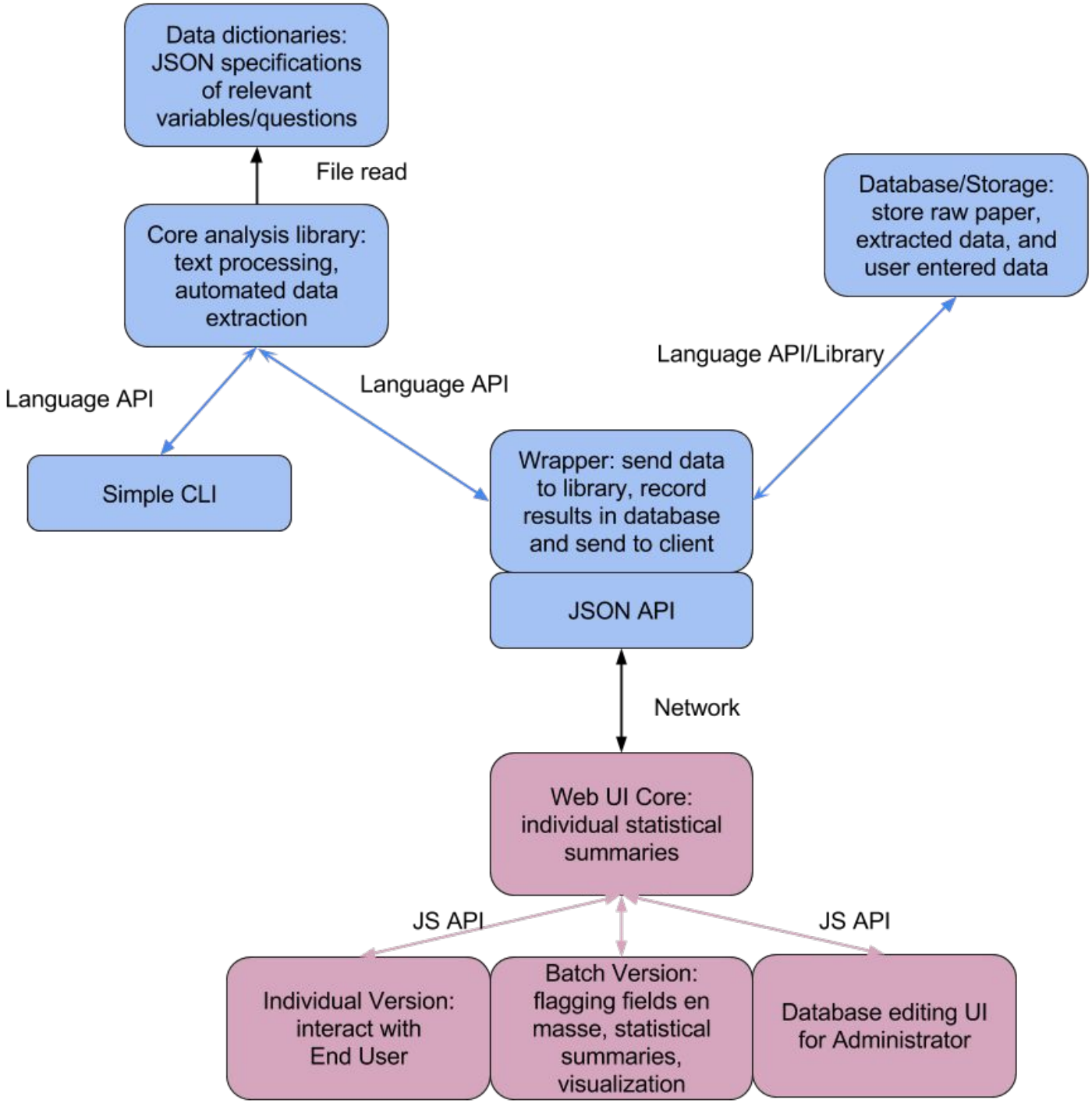
+	Bibliographic	See More
−	Database & Data Collection	See More
< >	Data Mining & Cleaning	See More
+	Methods: Data Analysis	See More
−	Data Sharing & Data Documentation	See More

Variables for *Methods: Data Analysis*

GOOD

	Question (variable)	Extracted Result	User Input *	
+	Software used for data analysis (analysis_sw)	STATA 12 was used for the statistical analysis and Excel 2007 for developing the graphs.		+
+	Does the author state analysis methodology and process? (analysis_process_clear)		This is an observational, retrospective, and multicentre study based on information gathered from the electronic health records (EHR) of primary care centres of two southern European regions of Spain: Aragon and Catalo- nia.	


Home
Paper List
About
Ricochetbyripeta
-- analyzing research reproducibility
Upload a Research Paper (pdf)
Upload a Research Paper
It will be parsed and analyzed for research reproducibility
DEMO:


Data dictionaries: JSON specifications of relevant variables/questions
File read
Core analysis library: text processing, automated data extraction
Database/Storage: store raw paper, extracted data, and user entered data
Language API/Library
Language API
Language API
Simple CLI
Wrapper: send data to library, record results in database and send to client
JSON API
Network
Web UI Core: individual statistical summaries
JS API
JS API
Individual Version: interact with End User
Batch Version: flagging fields en masse, statistical summaries, visualization
Database editing UI for Administrator

Mock-ups

View

Publisher

RicochetResults 2.0

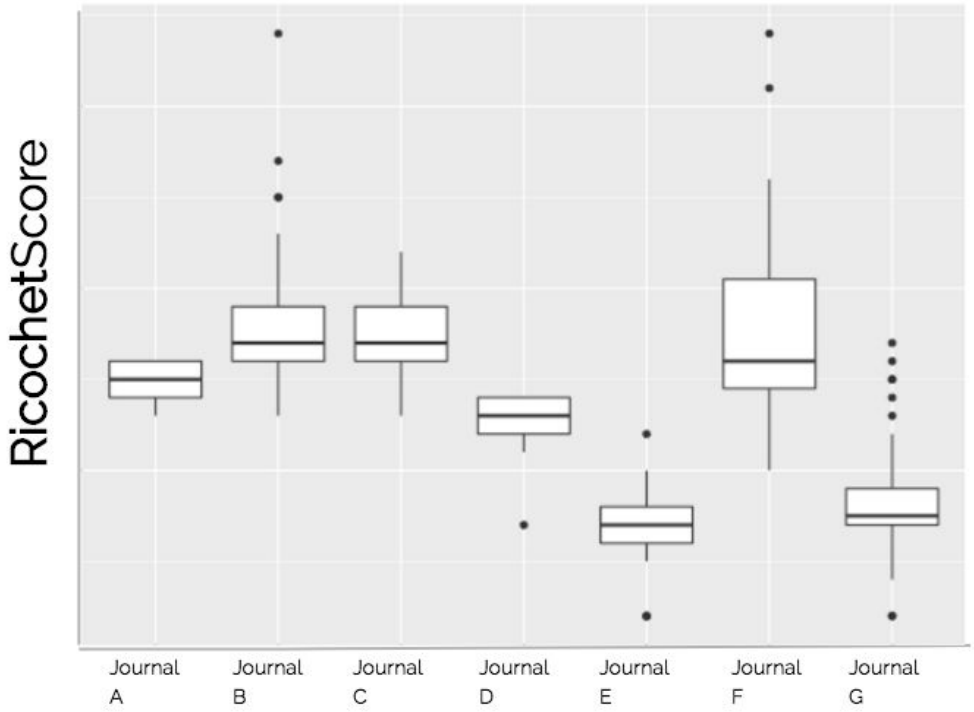
Sample Publisher
ricochetDashboard

April

Social Sciences & Humanities

- Arts & Humanities
- Business, Management & Accounting
- Decision Sciences
- Economics, Econometrics & Finance
- Psychology
- Social Sciences

Publisher	RicochetScore
Oxford University Press	3.6
PLOS	4.7
Sage Publications	3.9
Springer Publishing	4.2
Wlley	3.3



RicochetQuantSum

RicochetQuantSum

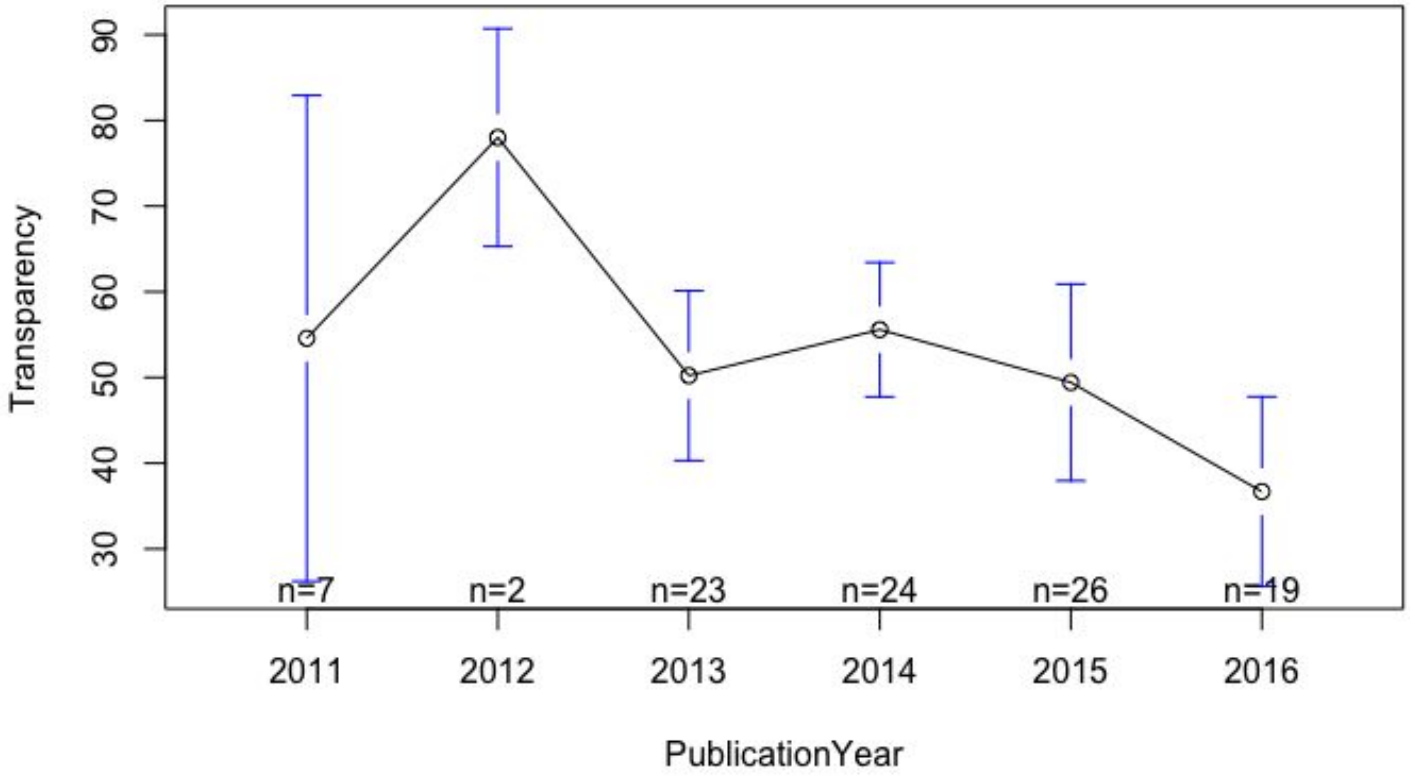
Cumulative summary of accessibility and transparency variables

Ordinal variables are ranked and weighted in relation to least to most transparent/accessible

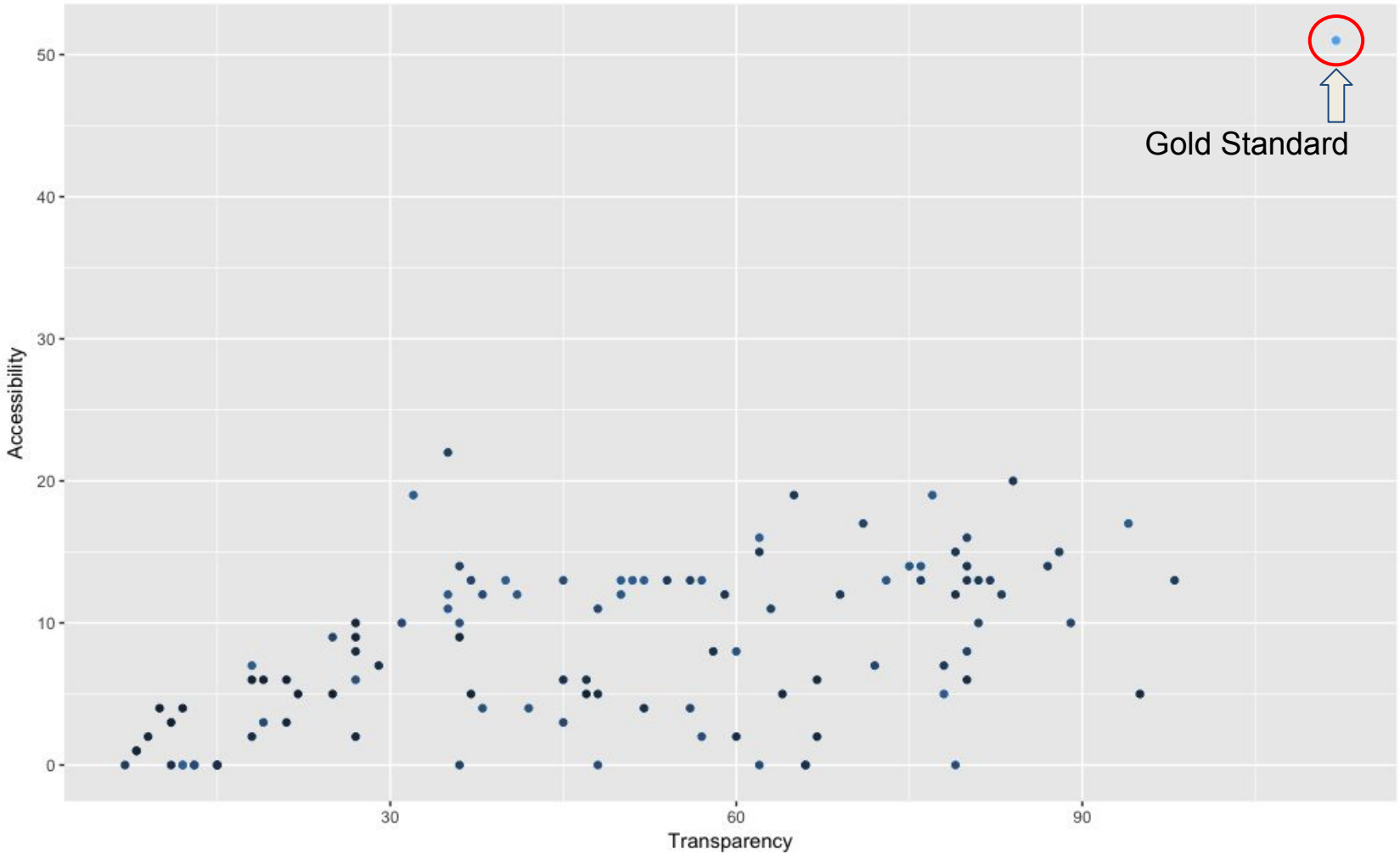
Quickly summarize and compare individual publications

Predict and compare trends of reporting behaviors within and across domains

Results


Transparency
30
40
50
60
70
80
90
n=7
n=2
n=23
n=24
n=26
n=19
2011
2012
2013
2014
2015
2016
PublicationYear

Results


Gold Standard
Accessibility
Transparency
0
10
20
30
40
50
30
60
90

McIntosh et al. BMC Medical Research Methodology _##### ######## ######## ###_
DOI 10.1186/s12874-017-0377-6

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Repeat: a framework to assess empirical reproducibility in biomedical research

Leslie D. McIntosh[1], Anthony Juehne[1*], Cynthia R. H. Vitale[2], Xiaoyan Liu[1], Rosalia Alcoser[1], J. Christian Lukas[1] and Bradley Evanoff[3]

OSF Pre-Print: https://osf.io/3sqbx/

Areas of Expansion

Biological Sciences

- Clinical Trials

https://drive.google.com/open?id=1ZeO_zOBGJQFEsv08JddQS9Aw0YP9KRjUZkQLVHAFX3I

- Oncology Health Outcomes

https://drive.google.com/open?id=0B8qs6MBUNJTpNTU1WHhxVEZldzg

- Pre-clinical
- Genomics

Grants & Proposals

How can you **get involved?**

- **Vet** our framework elements within an international community!

- **Test** our framework through adopting within your own research workflows

- Enter your own manuscripts to **refine** the following:
  - sensitivity
  - specificity
  - inter/intra rater reliability
  - international generalizability

Thank you!!

Github
https://github.com/ripeta

Open Science Framework
osf.io/ppnwa/