CompImpact: Computational Tools for Assessing the Social Impact of CSS Research

Daniel J. Hicks

Contents

1 Executive Summary

This report presents two quantitative studies of the social impact of high-throughput toxicology [HTT] research conducted under the auspices of EPA's Chemical Safety for Sustainability [CSS] national research program. Both assessment approaches rely on data-intensive, computational tools; the first examines the Twitter profile to CSS research, while the second applies text mining tools to analyze 15 years of reporting by Bloomberg BNA. Besides reporting some preliminary findings, this report also examines data- and model-based challenges to deploying these tools, and offers suggestions for how EPA can improve its ability to use these tools in the future. This report has the following major findings and recommendations:

- About one-third of CSS publications received any attention on Twitter. Papers that receive tweets reach a large but highly variable audience, on the order of 100-10,000 accounts per paper.

- Several Twitter accounts produced either 5 or more tweets of CSS papers, shared CSS papers with an estimated total of over potential 10,000 views, or both.

- Tweets often occur within a short period of time after a paper is published, two weeks or less. However, there is also a long right tail, with a few papers receiving tweets years after they are published.

- Bloomberg BNA's coverage of ToxCast and related topics has increased since 2005. This change is due in part, but not entirely, to changes in the length and number of articles on ToxCast-related topics.

- Coverage is highly variable over time. Months with especially high levels of coverage typically include multiple stories on general near- and medium-term regulatory uses of CSS research.

- Coverage of CSS research was almost always more positive than negative, and almost always more trusting than fearful. These patterns did not change over time.

- A major challenge for studying the impact of CSS publications is the initial task of assembling a suitable list of publications. In the future, modifications to STICS could be used to collect DOIs for published papers, which would greatly facilitate bibliometric studies.

- A major challenge for studying CSS-related media coverage using computational tools is that these tools are highly dependent on a large number of assumptions and analytical decisions, and that the resulting analyses are not necessarily robust. In the future, computational media analysis should be carefully tailed to specific analytical questions. Exploratory media analysis can be used to develop precise hypotheses or comunications strategies, which can then be tested using confirmatory methods.

2 Introduction

On July 5 1945, two months after the unconditional surrender of German forces and one month before the detonation of nuclear weapons over Hiroshima and Nagasaki, Director of the White House Office of Scientific Research and Development Vannevar Bush transmitted to President Truman a report on government support for scientific research. *Science, the Endless Frontier* provided the intellectual foundation for the National Science Foundation: basic scientific research, provisioned with resources and directed by scientific curiosity alone, would produce technological advances to improve health, ensure national security, and drive economic growth (Bush 1945).

In 1997, NSF's National Science Board introduced a significant change to the merit criteria used to review grant proposals. Scientific research would be evaluated not just on its intellectual merit, but also by its "broader impacts." Rather than simply assuming, as Bush had done, that basic scientific research would (somehow, automatically) produce new technology would (somehow, automatically) produce social benefits, NSF directed prospective grantees to include in their proposals explicit discussions of what social benefits would result from their research, and how these benefits would be realized. (For critical discussion of the limited success of the broader impacts criterion, see J. Holbrook 2005; J. B. Holbrook and Hrotic 2013.)

In an important sense, researchers at EPA have never fit within the simple division between basic and applied science. EPA research has always been dedicated to the ultimate aim of protecting human health and the environment, and in this sense does not share the pure curiosity, science-for-science's-sake of "basic research." At the same time, EPA researchers have been and continue to be important innovators, often working at the leading scientific edge; they have not merely applied pre-existing scientific tools.

Despite this overarching aim on social impacts, EPA researchers have placed relatively little emphasis on developing empirical methods to measure and assess the social impacts of their work. In part this may be because, for many areas of EPA research, there are well-established pathways for deploying research products in regulatory contexts. The social impact of these research areas can be measured simply in terms of the impacts of the downstream regulation. Conventional air quality research or chemical risk assessments might be good examples of this "normal regulatory science." However, other research areas — such as research that is actively developing new tools for public health or high-throughput toxicology — do not yet have established regulatory connections or simple regulatory metrics. The social impact of a screening prioritization decision for a potential endocrine disruptor, for example, is itself difficult to define. Assessing the social impact of these research areas requires more subtle tools.

This report presents two quantitative studies of the social impact of high-throughput toxicology [HTT] research conducted under the auspices of EPA's Chemical Safety for Sustainability [CSS] national research program. Both assessment approaches rely on data-intensive, computational tools; the first examines the Twitter profile to CSS research, while the second applies text mining tools to analyze 15 years of reporting by Bloomberg BNA. Besides reporting some preliminary findings, this report also examines data- and model-based challenges to deploying these tools, and offers suggestions for how EPA can improve its ability to use these tools in the future.

This report is fully reproducible; indeed, the figures and tables below are generated automatically by a collection of Python and R scripts. The complete source code for this report, along with the data sources required to reproduce it, are open source and available online at https://github.com/dhicks/comp-impact. Details on how to reproduce the report are given in its source code.

3 Study 1: Altmetrics

The term "altmetrics" has two distinct meanings. Small-a "altmetrics" is a portmanteau of "alternative bibliometrics," that is, alternatives to such metrics as paper citation counts, journal impact factors, and researcher H-indices. These standard bibliometrics are designed to measure the impact of research within the scholarly community, and thus are generally not useful for assessing social impacts. Consequently, there is a substantial amount of active work on developing altmetrics that are relevant to social impact (Wolf et al. 2013).

Large-a "Altmetrics" refers to "Altmetric.com," a UK-based company that develops an integrated set of small-a altmetrics and makes them available programmatically using a web-based API [application programming interface] (http://api.altmetric.com/). Large-a Altmetrics focuses on social media references to research publications, such as tweets and blog posts. The analysis explored here uses Altmetrics' tweets data to examine the social media impact of CSS publications.

3.1 Data

Bibliometrics, like other opportunistic uses of independently-cultivated data, depends on stable identifiers that can track individual research targets — such as research publications — across datasets from multiple sources. DOI, or the digital object identifier, has emerged as a major standard identifier for research publications. Other identifiers, such as PubMed's internal identifier, are common, but not as widely-used as DOI (Kraker, Enkhbayar, and Lex 2015). Altmetrics' API accepts queries in terms of both DOIs and PubMed IDs.

The first step in any bibliometrics analysis, then, is determining the DOI for each publication of interest. For an analysis of CSS publications, ideally, these DOIs could be identified automatically along with the publications of interest, using a database that (a) includes DOIs in the publication metadata, (b) permits a search by EPA research program, and (c) can export machine-readable search results. As far as I have been able to tell, EPA does not have any general publications database that satisfies all three desiderata. The "public-facing" version of Science Inventory (https://cfpub.epa.gov/si/) does not appear to satisfy any of these three requirements. The "internal application" version of Science Inventory (https://cfext.epa.gov/si/SciInv/stmProtoLogin.cfm) requires a separate registration for access; when I attempted to register a new account, the system generated errors that could not be resolved.

STICS [Science and Technical Information Clearance System] is designed to support the clearance and approval of research products before they are submitted to a journal for publication. STICS satisfies criteria (b) and (c), and thus can export a list of all research products associated with CSS. However, presumably because it is designed for use only

pre-publication, STICS does not include DOIs. After examining the metadata outputs from STICS, I decided that the most efficient way to identify DOIs corresponding to STICS records would be to search for matching titles in Scopus, a large database of research publications similar to Web of Science or PubMed. (For a comparative analysis of Scopus, Web of Science, PubMed, and Google Scholar for bibliometrics projects, see Mingers and Leydesdorff 2015.) Python scripts were prepared to conduct both "quoted" and "unquoted" searches for each research product title. A quoted search matches the title as a complete phrase; for example, a quoted search for the title "Recent Work in High-Throughput Toxicology" would not match "Recent Work for High-Throughput Toxicology." An unquoted search matches the individual terms; for example, "Recent Work in High-Throughput Toxicology" *would* match with "Recent Work for High-Throughput Toxicology." An unquoted search is useful for catching publications for which the title had been changed slightly during the review process, or for handling encoding errors (such the title stored in STICS as `A Framework for &quot;Fit for Purpose&quot; Dose Response Assessment`); however, an unquoted search is obviously more likely to return incorrect matches.

The results of these STICS-derived searches were compared with two manually-curated databases of publications: an EndNote database sporadically updated by CSS staff and a database of NCCT-related publications curated by Monica Linnenbrink. Both of the latter two databases include DOIs, but neither is intended to be a comprehensive collection of all CSS publications. In particular, the NCCT database includes several publications that predate the creation of the CSS national research program in 2011 or that describe work by non-EPA researchers using NCCT-developed tools; these publications were excluded from analysis. Otherwise, the combined results from all four database/searches were used in the analysis in the next section.

Manual inspection of the results identified several apparent false matches; publications with no EPA-affiliated authors were discarded.

Table 1 gives an overview of the number of DOIs included in each database/search. Combining the results of all four database/searches yields a total of 459 distinct DOIs. Both of the two STICS searches include a majority of these DOIs; the EndNote database contains somewhat fewer than half; and the NCCT database contains just over 10% of all DOIs.

Figure 1 shows the distribution of every individual paper across the four database/searches, as lit cells in a heatmap and as lines across parallel coordinates. These plots indicate that the two STICS searches include almost exactly the same publications, and that the EndNote and NCCT databases include a fair number of publications that were not identified using STICS searches. This may be because of changes to the titles between finishing clearance and final publication. Tables 2 and 3 make these same points quantitatively.

3.2 Methods

Given the set of DOIs for the target papers, a Python script queries the Altmetrics API, retrieving data on every tweet that references one of the target papers. The analysis then considers the number of tweets per paper, the number of CSS-related tweets per Twitter account, the estimated reach per paper and over time, and the delay (time between publication and first tweet) and lifespan (time between first and last tweet) for each Tweeted paper.

Table 1: Publications with DOIs found in each database/search

EndNote	NCCT	STICS (Quoted)	STICS (Unquoted)
204	47	350	350

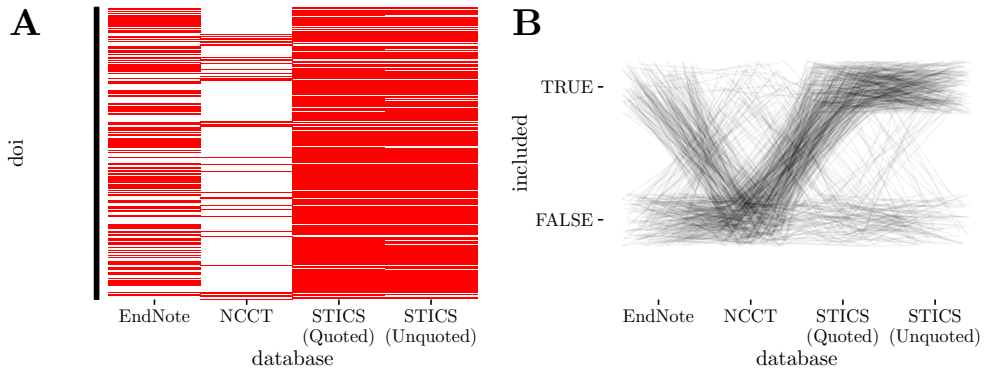

Figure 1: Distribution of individual papers across the four database/searches. **A**: Each paper is represented by a single row; a red cell indicates that the given paper is included in the given database/search. **B**: Each paper is represented by an unbroken line across the parallel coordinates. Y-axis indicates whether the given paper is included in the given database/search.

Table 2: Distribution of papers across the four database/searches

STICS (Quoted)	STICS (Unquoted)	EndNote	NCCT	n
			X	36
		X		54
		X	X	2
	X			11
	X	X		6
X				10
X		X		7
X	X			195
X	X		X	3
X	X	X		129
X	X	X	X	6

Table 3: Concordance between quoted and unquoted STICS search results

STICS (Quoted)	STICS (Unquoted)	n
		92
	X	17
X		17
X	X	333

3.3 Results

A total of 459 DOIs for CSS publications were identified. Altmetrics returned responses for 219 of these DOIs, of which 167 (76%) had 1 or more tweets, for a total of 483 tweets. 7 papers had 10 or more tweets; see figure 2A and table 4. There is considerable variation in the number of tweets over time; see figure 2B. Tweets are made in 28 different countries, though 59% have no country information; see figure 3.

The 483 tweets were made by 302 distinct accounts. 247 accounts (91%) made only a single tweet of a CSS paper, while 19 accounts made 5 or more tweets; see figure 4 and table 5. Many of these relatively high-tweeting accounts promote recent publications in a scientific topic, in a journal, or of potential interest to members of a professional society.

3.3.1 Reach

In marketing, *reach* refers to the size of the potential audience. The Altmetrics API provides the number of followers at the time of each individual tweet for at least some accounts. These follower counts can be used to estimate reach at the paper level and over time. (Note that the followers for two given tweets can overlap, and follower counts appear to be unavailable for a substantial number of accounts, so these estimates are susceptible to errors that are difficult to correct.) Aggregated both ways, reach varies over more than two orders of magnitude. See figure 5.

3.3.2 Delay and Lifespan

Delay can be defined as the time between publication date (as recorded in the Altmetrics metadata) and the first tweet. Delay had a mean of 70 days and median of 5.7 days, with an interquartile range of 42 days. 52% of papers had a delay of less than 7 days; notably, 21% of papers had a negative delay, indicating that the paper was first tweeted before it was officially published (according to Altmetric's records). No relationship was identified between delay and the total number of tweets received by a paper. See figure 6 and table 6.

Similarly, *lifespan* can be defined as the time between the first and last tweet. Lifespan had a mean of 65 days and median of 4.4 *hours*, with an interquartile range of 57 days. 64% of papers had a lifespan of less than 7 days. Note that 72 papers received only a single tweet, and thus had a lifespan of 0; excluding these papers, mean lifespan is 115 days and median lifespan is 27 days. See figure 7, and tables 6 and 7.

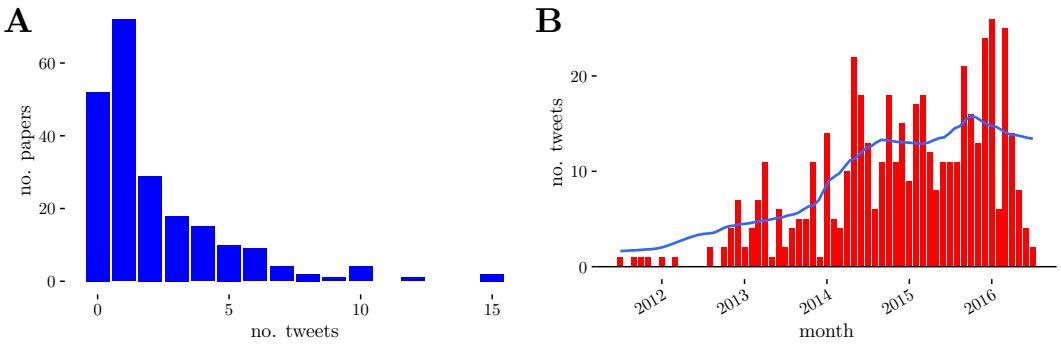

Figure 2: Number of tweets. **A**: Histogram of tweets per paper. **B**: Number of tweets per month. Blue line is a moving average.

Table 4: Highly-tweeted papers

DOI	title	no. tweets
10.1038/nbt.2914	Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms	15
10.1016/j.envint.2015.12.008	Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring	15
10.1093/toxsci/kfv195	Mining the Archives: A Cross-Platform Analysis of Gene Expression Profiles in Archival Formalin-Fixed Paraffin-Embedded Tissues	12
10.1016/j.ijheh.2015.02.002	Characterization of silver nanoparticles in selected consumer products and its relevance for predicting children's potential exposures	10
10.1093/toxsci/kfw002	High-throughput screening of chemical effects on steroidogenesis using H295R human adrenocortical carcinoma cells	10
10.1021/es503583j	High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals.	10
10.1093/toxsci/kfw034	Tiered High-Throughput Screening Approach to Identify Thyroperoxidase Inhibitors within the ToxCast Phase I and II Chemical Libraries	10

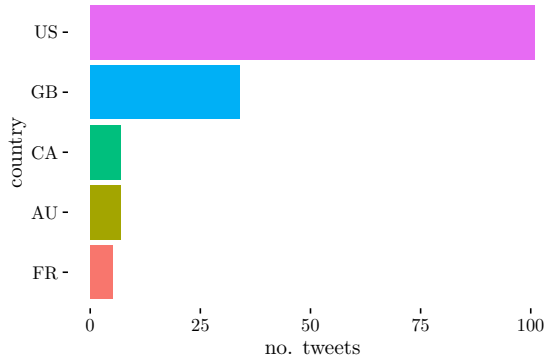

Figure 3: Number of tweets per country, countries with at least 5 tweets

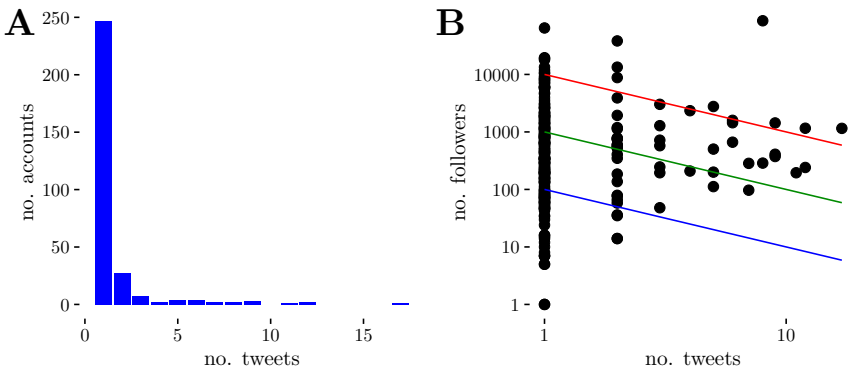

Figure 4: Accounts tweeting CSS papers. **A**: Histogram of tweets per account. **B**: Scatterplot of tweets and followers per account; note log scales. Colored lines are curves where total reach (no. followers $\times$ no. tweets) $= k$, for $k = 10^4$ (red), 10^3 (green), and 10^2 (blue).

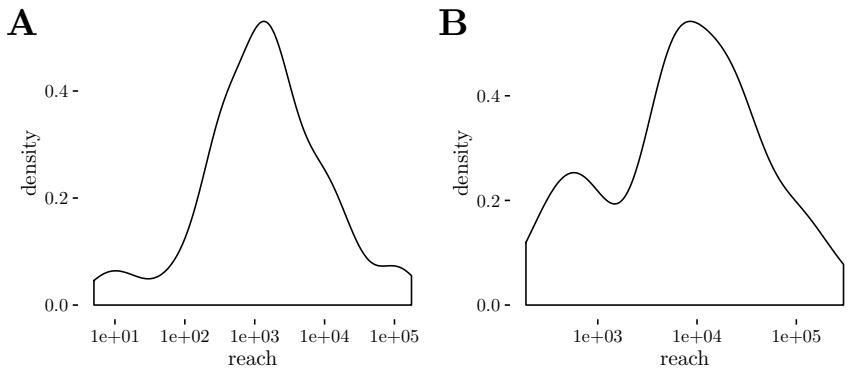

Figure 5: Kernel density estimates of estimated reach. Note log scales. **A**: Reach per paper. **B**: Reach per month.

Table 5: Twitter accounts producing more than 5 tweets of CSS publications or total reach (no. followers × no. tweets) greater than 10,000

account	no. tweets	no. followers	total reach	location
@EPAresearch	8	85728	685824	
@SciReports	2	38219	76438	GB
@NatureBiotech	1	64236	64236	
@BlackPhysicists	2	13373	26746	
@ecotoxicology	17	1153	19601	
@onlyorganic	1	19520	19520	
@F1000	1	18992	18992	
@AmSciMag	1	17966	17966	US
@rlanzara	2	8809	17618	US
@SOToxicology	12	1159	13908	US
@EnvSciTech	5	2767	13835	US
@StemCellMarket	1	13395	13395	US
@Loose_Lab_Rat	9	1432	12888	
@pant_prateek	1	11795	11795	US
@NanotechWeek	1	10703	10703	US
@nature_brains	1	10394	10394	
@LRIG_ORG	6	1591	9546	
@USGS_MN	6	1445	8670	US
@LabRobot	6	663	3978	
@limnologia	9	409	3681	
@acsdchas	9	374	3366	
@ForecomBio	12	241	2892	GB
@Immunol_papers	5	503	2515	
@cornelllabsafe	8	287	2296	
@ToxAndBeyond	11	195	2145	
@etc_editor	7	284	1988	
@ToxSci	5	200	1000	
@labsustain	7	97	679	
@ASRS	5	112	560	
@ChemConnector	6			

Table 6: Weekly cumulative percentiles for lifespan and delay

week	delay (cum. %)	lifespan (cum. %)
0	21	43
2	61	69
4	70	72
6	74	74
8	80	75
10	82	75
12	82	79
14	84	80

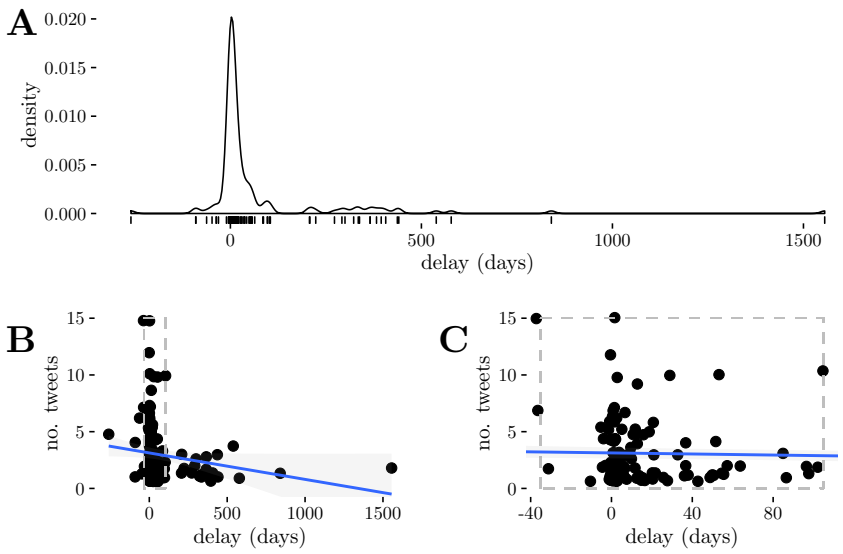

Figure 6: Delay. **A**: Kernel density estimate of delay. **B** and **C**: Scatterplot of the total number of tweets vs. delay. Grey rectangles in the two plots correspond. Blue line is a linear regression for the entire dataset.

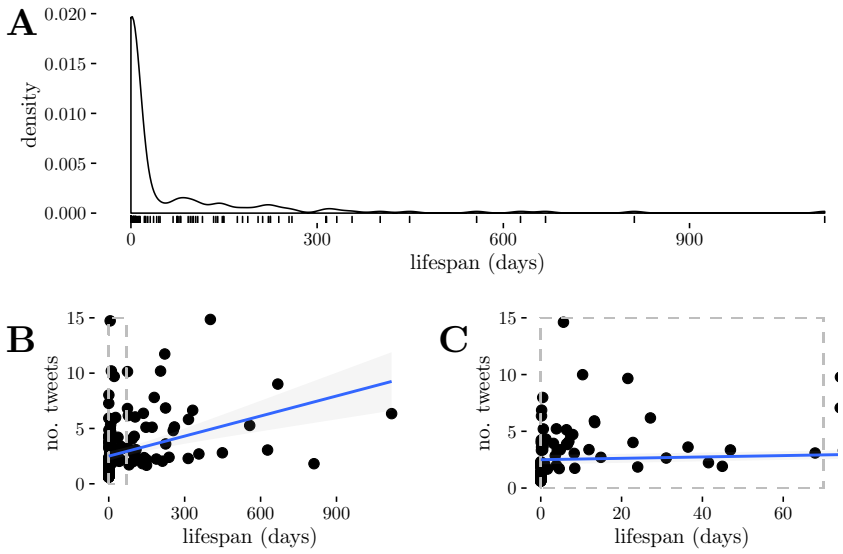

Figure 7: Lifespan, time between first and last tweet. **A**: Kernel density estimate of lifespan **B** and **C**: Scatterplot of the total number of tweets vs. lifespan. Grey rectangles in the two plots correspond. Blue line is a linear regression for the entire dataset.

Table 7: Papers with lifespan greater than 365 days

DOI	title	published	last tweet	no. tweets
10.1016/j.neuro.2011.11.005	Developmental thyroid hormone disruption: Prevalence, environmental contaminants and neurodevelopmental consequences	December 6, 2011	February 27, 2014	3
10.1371/journal.pcbi.1002996	A Computational Model Predicting Disruption of Blood Vessel Development	April 4, 2013	April 26, 2016	6
10.1021/es403094q	Uptake of Perfluoroalkyl Acids into Edible Crops via Land Applied Biosolids: Field and Greenhouse Studies	November 13, 2013	April 5, 2016	2
10.1021/tx400310w	Development of a Thyroperoxidase Inhibition Assay for High-Throughput Screening	January 6, 2014	October 4, 2015	3
10.1650/condor-13-045.1	Landscape and regional context differentially affect nest parasitism and nest predation for Wood Thrush in central Virginia, USA	May 1, 2014	March 11, 2016	9
10.1038/nbt.2914	Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms	May 18, 2014	June 25, 2015	15
10.1021/es502976y	Toxicity, Bioaccumulation and Biotransformation of Silver Nanoparticles in Marine Organisms.	November 6, 2014	June 3, 2016	5

There is evidence of a statistically significant ($p = 3.54 \times 10^{-6}$) relationship between lifespan and the number of tweets a paper receives, but this relationship is weak ($b = 0.006$ additional tweets per additional day of lifespan; $R^2 = 0.12$).

Finally, there is no evidence of a relationship between delay and lifespan. See figure 8.

3.4 Discussion

Data issues — specifically, acquiring a comprehensive list of all and only CSS publications — presented a major challenge for this analysis. If EPA chooses to explore the use of altmetrics in the future, it may be worthwhile to expand the capabilities of STICS (or a similar system) to track publication DOIs. For example, after a publication has completed clearance, STICS could send quarterly or biannual reminders to authors, encouraging them to provide final publication metadata (at a minimum, accepted journal and DOI) for their research products.

Only about one-third of all CSS papers (36%) received any tweets at all, and many of these papers received only a single tweet. The distribution of tweets, delay, and lifespan was highly skewed with a long right tail; that is, while many papers receive a few tweets for a short period around the time they are published, a few papers have an especially high number of tweets, a long delay before receiving their first tweet, or are tweeted about for a long period of time. Only a single paper received more than 15 tweets. At the same time, these tweets have a large reach, on the order of 100-10,000 people per paper.

J. Britt Holbrook (personal communication) has proposed that researchers could use altmetrics to identify future opportunities to increase the social impact of their work. Tables 4, 7, and 5 may be especially useful here. Tables 4 and 7 may indicate which topics or CSS project areas tend to attract relatively broad attention on social media; strategically tweeting about these topics or project areas in the future might be an effective way to increase the social media presence of CSS research. Similarly, table 5 indicates Twitter accounts/users who tend to pay attention to CSS research. CSS communications staff could engage with these accounts/users, encouraging them to tweet about CSS research more in the future.

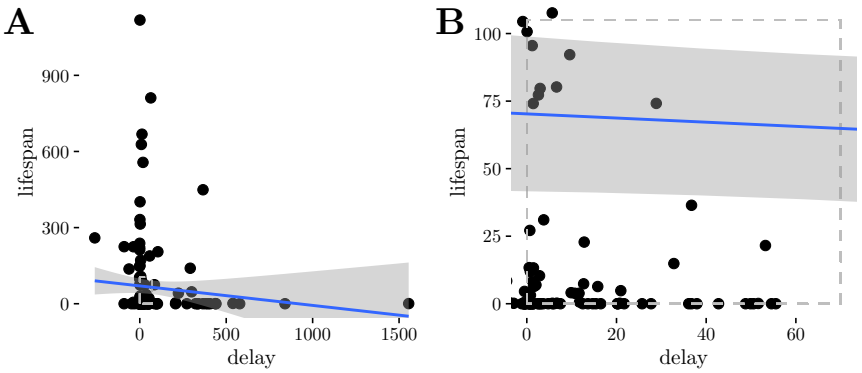

Figure 8: Scatterplot of lifespan vs. delay. Grey rectangles in A and B correspond. Blue line is a linear regression of the entire dataset.

4 Study 2: Bloomberg BNA

Media analysis provides another approach to assessing the impact of a public-interest scientific research program. Traditional methods of media analysis require manual review of articles and other documents. While these methods can provide a rich and nuanced understanding of the way scientific research is represented to the public, and changes in this representation over time, they may be too resource-intensive in some cases. In such cases, text-mining tools — computational tools for quantitative analysis of large bodies of text — may be used to produce quick — but potentially "thin" or un-nuanced — analyses and suggest refined targets for analysis using classical techniques

This section applies text mining tools to analyze a large dataset of trade media coverage of regulatory toxicology. This study addresses two questions:

1. How has coverage of the ToxCast program, CSS, and high-throughput toxicology [HTT] more generally, changed over time?

2. When HTT is covered, how is it represented?

4.1 Data

With assistantance from Bloomberg BNA staff, BNA's web API was used to retrieve all articles using the word "toxicology" published since 2000. Article metadata were retrieved in XML files; custom R scripts were used to parse and combine these XML files, then retrieve text and byline for each article. Manual inspection of the resulting article set found several hundred items that were short summaries of *Federal Register* notices, announcements of public comment periods, lists of events, and other article types that were judged to be irrelevant; all of these article types were removed.

Because BNA's own database was queried directly using the API, the dataset can be considered complete and error-free. One remaining potential source of data error is the inclusion of repeated or duplicate articles. Manual review of titles and a programmatic check indicated no duplicate articles.

4.2 Methods

4.2.1 "ToxCast" Scores

To address research question 1, the vocabulary used in the articles was first normalized. Individual terms were converted to lowercase; punctuation was removed (except for interword hyphens, as in "high-throughput"), along with numbers, special symbols, and stopwords (words so common that they can confound text mining, such as "the").

To examine coverage of CSS within this corpus, the analysis focused on the (normalized) term "toxcast" and associated terms, meaning terms that tend to appear in the same documents as "toxcast," as well as sets of the 10, 100, and 1000 terms most strongly associated with "toxcast". "Scores," occurrence frequencies across each of the four sets of terms, were calculated for

each document, as both raw totals and normalized by document length. Finally, scores were aggregated by month and linear trends over time were examined. For the analysis presented here, distance calculations include February 2000. For an analysis of the robustness of these calculations, see the appendix.

4.2.2 Sentiment Analysis

Sentiment analysis uses manually-prepared reference lists of emotionally-laden terms to estimate the emotional valence of texts. In the particular technique used here, each text is assigned an emotional valence score — as "positive" or "fearful," say — based on the occurrence of individual words that human curators have judged to be "positive" or "fearful" (Mohammad and Turney 2013). While this technique is obviously useful for addressing research question 2, concerning the way HTT is represented in the trade media, its results should not be over-interpreted. Specifically, the results depend heavily on the content of the reference lists; do not take into account sentence structure; and are easily confounded by irony and other complex rhetorical constructions.

This sentiment analysis tool was applied to every article in the corpus with a non-zero toxcast100 score. The sentiment analysis tool estimates scores for ten emotional valences: "anger," "anticipation," "disgust," "fear," "joy," "sadness," "surprise," "trust," "negative," and "positive." Both raw (total frequency) and normalized (per 1,000 words in the article) sentiment scores were calculated.

4.3 Results

Publication date, title, byline, and full text were obtained for 1,580 articles. After normalization, the vocabulary included 26,298 distinct terms and a total of 802,385 tokens (word-instances). See figure 9. The large spike in total length in February 2000 is due to five long, high-level EPA policy documents apparently republished by BNA. Consequently, articles from February 2000 are generally excluded from the analyses below. After excluding February 2000, the corpus contained 735,167 tokens.

4.3.1 "ToxCast" Scores

Table 8 shows the 100 terms most closely associated with "toxcast" in the dataset. Figure 10 shows monthly total ToxCast scores for the term "toxcast" by itself and the 10, 100, and 1000 terms most closely associated with "toxcast," excluding February 2000. In the remainder of this report, these sets of terms and scores are referred to as "toxcast10," "toxcast100," and "toxcast1000."

The first instance of "toxcast" was on March 1, 2006. The plot suggests that, after January 2005, the term "toxcast" occurred more frequently but roughly constantly. The toxcast10 set shows a steady increase over the entire period 2000-2016; notably, toxcast100 and toxcast1000 show decreasing trends prior to January 2005, then increasing trends after this point. Table 9 confirms this visual impression, finding statistically significant differences in trends before and after January 2005 for the toxcast100 and toxcast1000 sets of terms.

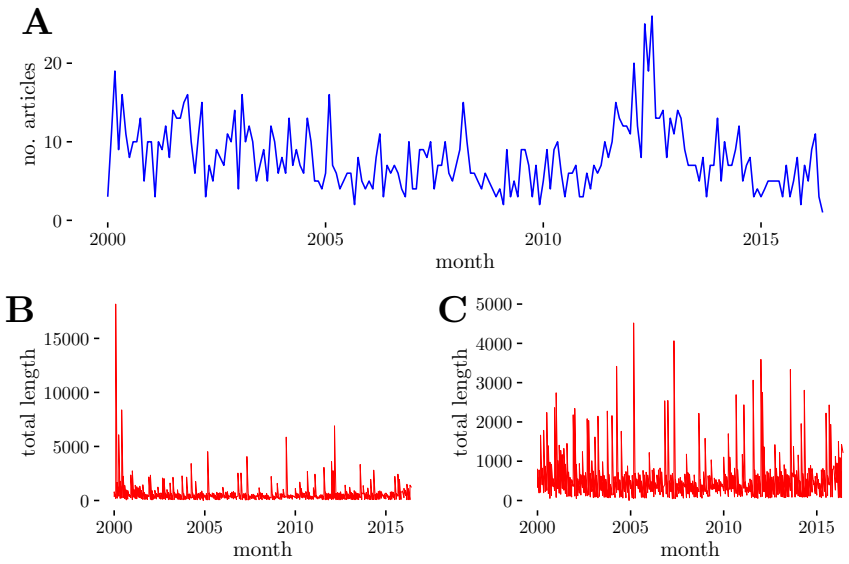

Figure 9: Overview of the BNA dataset. **A**: Number of articles published each month. **B** and **C**: Total length of articles per month (total token acount after normalization). **C** omits months with more than 5,000 total tokens.

Table 8: ToxCast 100 terms, normalized

actor	aggregated	aims	analyze	animal-based
assay	assays	assistant	attending	automated
bahadori	battery	biological	cell	cells
cellular	century	clinical	collaboration	computational
cost-effective	cost-effectively	dominate	ecosystems	efficiently
embryos	emerging	expensive	expocast	expose
fast	faster	forecaster	high-speed	high-throughput
host	hts	humane	initiative	integrate
investment	jim	join	jones	judson
kavlock	launched	maintain	mostly	ncct
non-agency	offices	outside	paradigm	pat
pharmaceutical	phase	predict	predicted	predicting
prediction	predictive	priorities	prizzuto	proof–concept
proteins	quick	quickly	rapid	rizzuto
robert	robot	robotic	screen	screening
screens	scrutiny	signaling	software	start
summit	sustainability	technologies	tens	thousands
throughput	tina	tools	tox	toxcast
toxicities	transformation	translational	trials	unveiled
virtual	vision	vital	vitro	vivo

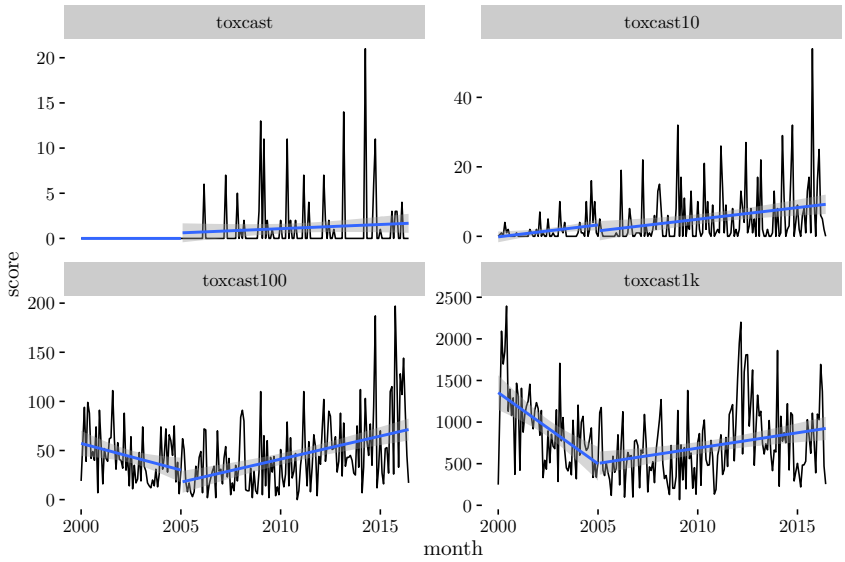

Figure 10: ToxCast scores. Scores are calculated as monthly total occurrences of "toxcast" and its 10, 100, or 1000 most closely-associated terms. Blue lines indicate linear regressions before and after January 2005. Articles from February 2000 are excluded from this plot.

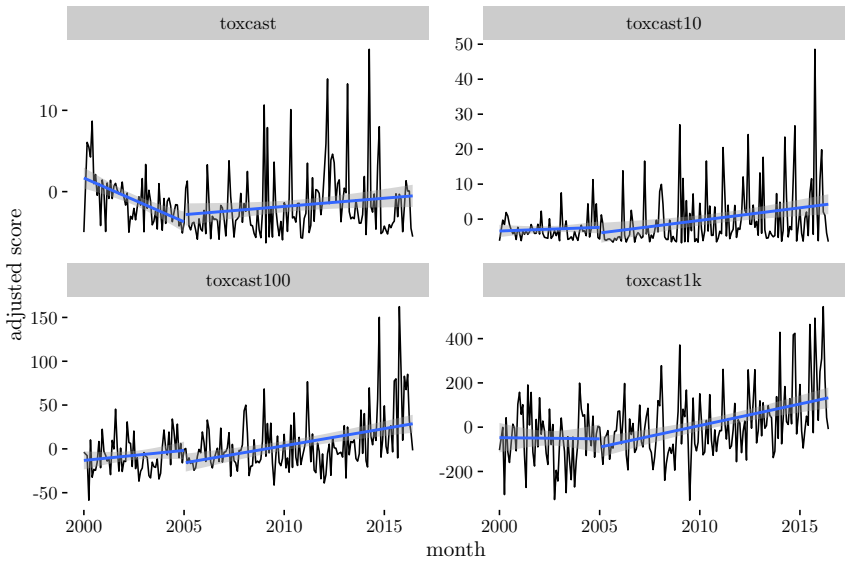

Figure 11: ToxCast scores after adjusting for monthly article and word totals. Articles from February 2000 are excluded from this plot.

Since scores are based on word frequencies, they are likely to be affected by the number of articles per month and the total length of those articles. Figure 11 plots scores after adjusting for these two factors. After adjustment, "toxcast" shows a striking downward trend before 2005; the length of the articles were gradually changing over this period, and so the model expects to see an increase, but "toxcast" does not occur at all, so the trend line runs down. The other three sets of terms exhibit roughly the same pattern, with flat trends before 2005 and modestly increasing trends afterwards. This suggests, first, that the decreasing pre-2005 trends in figure 10 are due to changes in the number of articles and total length; and second, that the increasing post-2005 trends are not only due to changes in the number of articles and total length.

Aside of the general trends, the plots show a number of "peaks," months with especially high total scores. Table 10 lists the articles with toxcast100 scores greater than 0 for those months with total toxcast100 scores at least 120 (excluding February 2000). There are five such peaks, in October 2014, October 2015, November 2015, January 2016, and March 2016. The October 2015 peak is essentially the result of a single article with an extremely high toxcast100 score; the other four peaks are due to combinations of multiple articles with modest to high scores.

Table 9: Comparison of linear trends before and after January 2005. Trend values are annual changes in the number of articles per month. z: Z statistics for the difference in trends. p: p-values of z statistics against null hypotheses of no differences in trends. Articles from February 2000 are excluded from this analysis

	trend (before)	(se)	trend (after)	(se)	z	p
toxcast	0.0	0.0	0.09	0.08	1.14	1.26e-01
toxcast10	0.7	0.3	0.67	0.22	-0.05	4.82e-01
toxcast100	-5.4	2.2	4.69	0.86	4.27	9.81e-06
toxcast1000	-170.8	35.6	36.74	10.76	5.58	1.20e-08

Table 10: Articles with non-zero toxcast100 scores, from months with total toxcast100 scores at least 120. February 2000 is not included in this table.

date	title	score
October 1, 2014	Chemical Testing: Non-EPA Scientists Will Peer Review Uses of Emerging Toxicity Tests, Jones Says	50
October 2, 2014	Chemical Testing: Scientists Describe Diverse Applications Of Data From EPA's High Throughput Assays	72
October 2, 2014	Risk Assessment: EPA to Issue Solvent Risk Assessment Soon; Road Map for Flame Retardants Also Coming	22
October 3, 2014	Chemicals: EPA Issues Final Report on Benefits, Limits Of New Data Sources, Analysis Techniques	30
October 3, 2014	Chemicals: Rubber Chemical Is Human Carcinogen, Three Other Compounds May Be, HHS Says	6
October 23, 2014	Chemicals: Studies of Bisphenol A, Thermal Paper Reach Opposite Conclusions on Significance	3
October 24, 2014	Pesticides: EPA Reorganizes Registration Division, Announces Lewis as New Division Director	1
October 31, 2014	Chemicals: In Light of Overlapping Toxicology Efforts, Agencies Need Help Coordinating, GAO Says	3
October 8, 2015	Risk Assessment: NTP Invites Data on Chemicals, Mountaintop Removal	5
October 19, 2015	Utility of Emerging Technologies in Chemical Safety Assessment, Sustainability	182
October 19, 2015	Chemicals: Volunteers Sought to Wear Wristband Measuring Chemicals	9
October 22, 2015	Chemicals: Sunscreen Chemical Toxic to Coral Reefs, Study Says	1
November 19, 2015	Drinking Water: EPA Releases Algal Bloom Strategy to Protect Water	2
November 20, 2015	Chemical Testing: EPA Expects to Authorize More Rapid Toxicity Tests	40
November 24, 2015	Chemical Testing: Agencies Lay Scientific Foundation for Rapid Toxicity	79
January 6, 2016	Chemicals: FDA Revokes Use of Three Perfluorinated Chemicals	4
January 7, 2016	Chemicals: Cancer Hazard of Metal Fluid, Flame Retardant Analyzed: NTP	3
January 8, 2016	Chemical Testing: Agencies Offer Up to $1 Million in Toxicity Testing Challenge	53
January 12, 2016	Mining: Clash Over Mining Rules Ahead	5
January 13, 2016	Risk Assessment: Data, Computational Tools for Risk Analysts Key in 2016	51
January 19, 2016	Risk Assessment: Academies Hosts Workshop on Low Doses of Chemicals	11
January 25, 2016	Nanotechnology: OECD to Share System With Nanomaterials Chemical Hazard Data	1
March 1, 2016	Chemicals: Making In Vitro Chemical Data Useful for Decisions	65

March 11, 2016	Risk Assessment: EPA Science Adviser Discusses Public Health Role	29
March 11, 2016	Risk Assessment: EPA: Lower Doses of Explosive May Harm Health	6
March 11, 2016	Chemical Testing: New Software Expected to Help Chemical Safety Reviews	28
March 18, 2016	Chemical Testing: Webinars Set for Acute Inhalation Toxicity Tests	2
March 21, 2016	Pesticides: Fewer Animals Used Under Draft EPA Pesticide Guide	10
March 24, 2016	Pesticides: As Zika Spreads, No New Funding for Insecticides	4

The titles of the articles from these peaks indicate, first, that the toxcast100 score is a precise detector of CSS-relevant stories in BNA, with a low false positive rate (i.e., few articles with a high toxcast100 score but that are not relevant to CSS). Second, the stories in these peaks are generally not about internal scientific developments, but instead about near- and medium-term regulatory uses of CSS tools. However, third, stories that focus on specific hazards seem to have lower scores than stories that give broad overviews of high-throughput toxicology, at least within these peaks. (Notably, the EDSP Pivot in June 2016 does not appear on this list; as far as I have been able to tell, BNA's only coverage of the EDSP Pivot was a brief note in one of the *Federal Register* overview articles.)

4.3.2 Sentiment Analysis

1,177 stories had non-zero toxcast100 scores. Figure 12 shows the distribution of sentiment analysis scores for each emotional valence across this set of stories, normalized by article length. Except for a substantial number of 0 scores, the distribution of scores for each valence is roughly log-Gaussian. Anger, disgust, joy, sadness, and surprise have consistently low scores, so the analysis focuses on anticipation, fear, trust, negative, and positive.

Figure 13 shows the distribution of sentiment scores over time, along with local regressions. All five emotional valences are basically stable over time. Notably, positive scores appear to be higher on average than negative scores. Figure 14 shows the distribution of the difference between positive and negative scores. 97% of articles have a greater positive than negative score, and this positive difference is stable over time.

Figure 15 shows that there is no relationship between toxcast100 score and length-normalized positive scores. There is a statistically significant ($p = 6.13 \times 10^{-11}$) negative association ($b = -0.074$) between toxcast100 score and negative valence; however, as indicated by the plot, this relationship is small and non-explanatory ($R^2 = 0.036$). That is, coverage that is more focused on CSS-relevant research does not tend to be more or less positive, but does have a slight tendency to be less negative.

Figure 13 also suggests that "trust" scores are generally greater than "fear" scores. Figure 16 confirms this; trust scores are greater than fear scores for 96% of articles.

4.4 Discussion

This study examined trade media coverage of CSS research using two text-mining techniques. The first technique identified a set of standardized terms that appeared to be specific to CSS-relevant stories; this study showed increasing attention to CSS research since 2005, thereby addressing research question 1. This change over time is due in part, though not entirely, to changes in the number of stories per month and the length of stories. Coverage tends to be more relevant to CSS research when it deals with general or broad regulatory uses of this research in the short- and medium-term. The second technique assessed the emotional valence of terms used in CSS-relevant stories, thereby addressing research question 2. These stories were consistently more positive than negative, and expressed trust more than fear. These patterns were consistent over time, and did not appear to be associated with the degree to which the article was CSS-relevant.

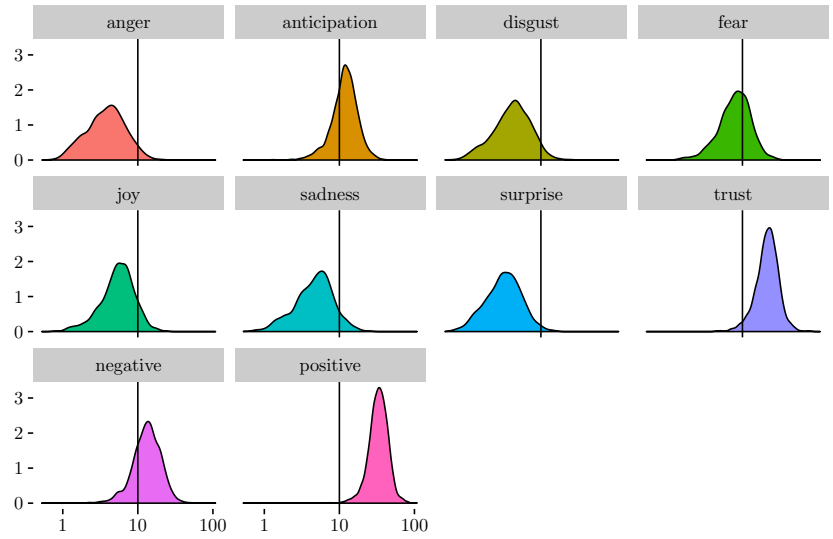

Figure 12: Kernel density estimates of length-normalized sentiment analysis scores. Scores are on x-axis (note log scale), with densities on the y-axis. Vertical line indicates scores = 10 (1% of all article words). February 2000 is excluded from this plot.

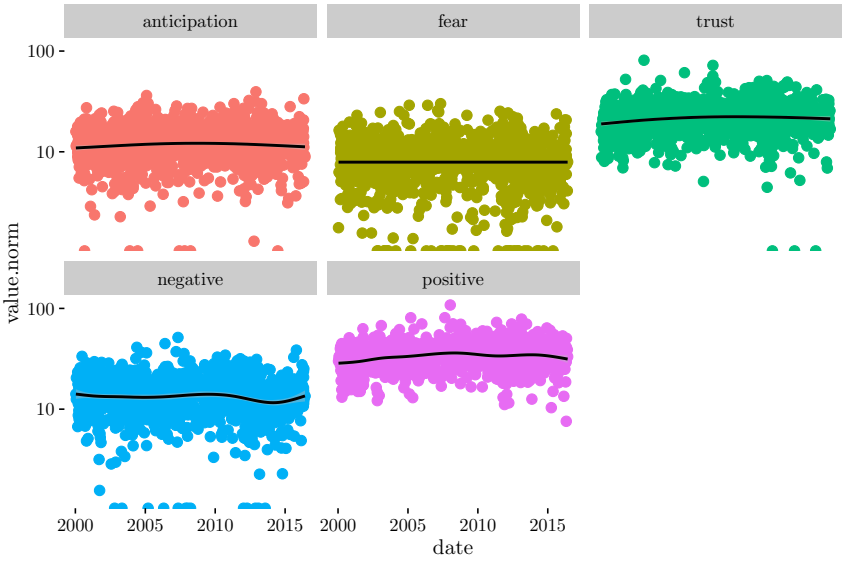

Figure 13: Distribution of sentiment analysis scores over time. Black lines indicate loess regressions. February 2000 is excluded from this plot.

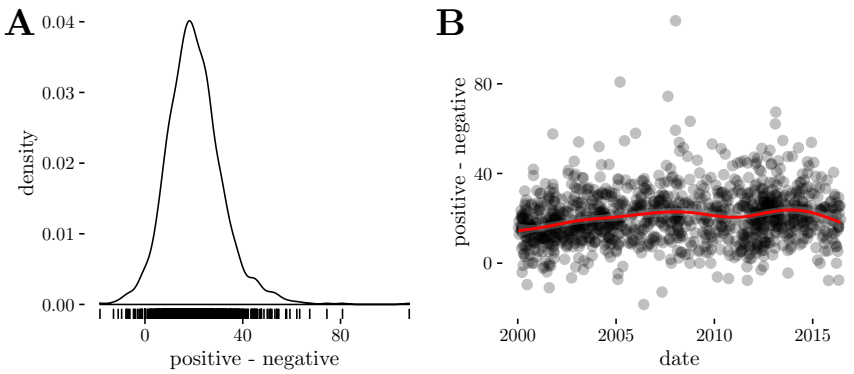

Figure 14: Distribution of differences between positive and negative emotional valence scores. **A**: Density of differences. **B**: Differences over time; red line is a loess regression. February 2000 is excluded from these plots.

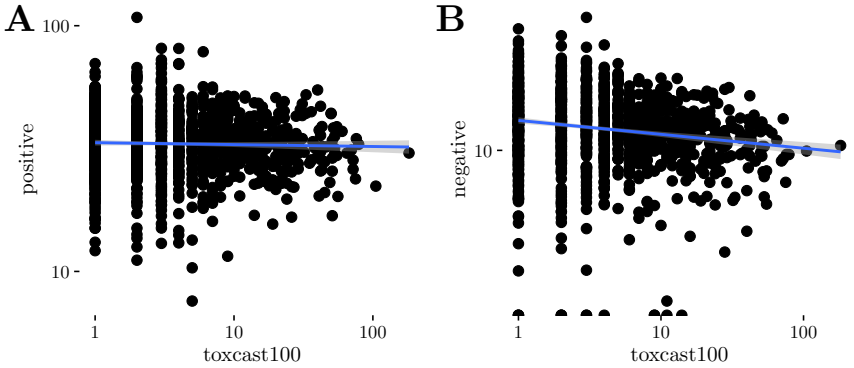

Figure 15: Correlations between toxcast100 scores and (**A**) positive and (**B**) negative emotional valences. Blue lines are linear regressions. Note log scales on all axes.

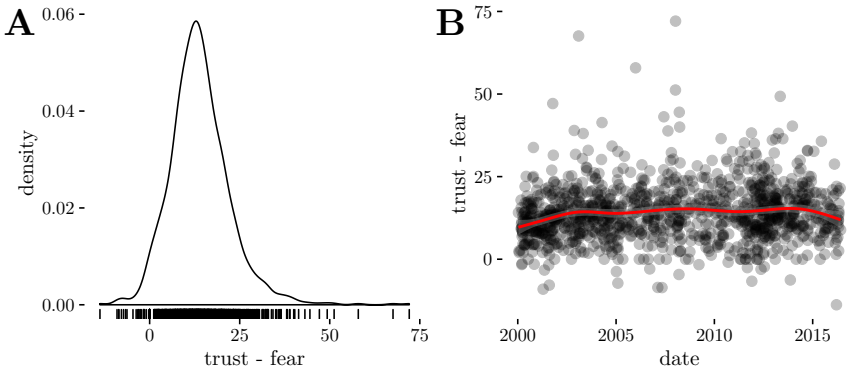

Figure 16: Distribution of differences between trust and fear emotional valence scores. **A**: Density of differences. **B**: Differences over time; red line is a loess regression. February 2000 is excluded from these plots.

The analysis in this study depends heavily on a substantial number of analytical decisions and assumptions. Regarding the underlying data, key decisions include the terms used for the initial BNA search, and the inclusion/exclusion criteria used to refine the set of results. Both of the subsequent analyses are based on simple word counts, which completely ignore structure and context; aggregate at the article or monthly level (rather than sentence, paragraph, quarter, or year); involve various decisions to use raw or normalized/adjusted scores or statistics; and use statistical techniques that assume Gaussian distributions and linearity. For the "ToxCast" scores analysis, key decisions include decisions about how vocabulary would be standardized (which stopwords were removed, personal names were not identified and handled separately, common stems such as -s and -ing were not removed), the decision to use "toxcast" as the starting point for analysis, the use of Jaccard distance to identify associated terms, the size of the sets of associated terms (and the use of size, and rather than a distance threshold, to define these sets), and a choice of a date threshold. The sentiment analysis depended heavily on a particular sentiment analysis tool, which in turn had its own substantive assumptions (Mohammad and Turney 2013).

While these decisions and assumptions do not invalidate the study's findings, it is highly plausible that they could have made a difference in the findings — that the findings could have been different if decisions had been made differently — which does complicate the interpretation of the findings. This point is illustrated by the robustness analysis of the construction of the toxcast100 sets given in the appendix. However, because of the large number of decisions and assumptions, it is impractical to check the robustness of these findings by surveying the entire space of possible analyses.

4.5 Appendix: Robustness Analysis of Four Term-Distance Metrics

This subsection examines concordance and discordance between four different ways of calculating distances from "toxcast" in the BNA vocabulary, and thus four different ways of determining which terms are included in the toxcast100 set. In paticular, two different metrics are considered — Jaccard distance and ℓ_1 distance — across every article in the dataset and excluding February 2000 articles.

Jaccard distance is typically used to compare the distance between two sets A and B. In set-theoretic notation, Jaccard distance is defined as

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|},$$

or 1 minus the fraction of elements that are in both sets. Applied to the BNA vocabulary, a term's associated set is the set of all documents that use the term at least once. Then, for two terms A and B, their union $A \cup B$ is the set of all documents that use at least one of the two terms, and their intersection $A \cap B$ is the set of all documents that use both of the terms. Note that this distance does not take into account how frequently a term occurs *within* a given document. Jaccard distances range from 0 to 1, where 0 means the two terms occur in exactly the same documents and 1 means the terms never occur together.

The ℓ_1 distance takes within-document frequencies into account. For a term A, let a_i be the

number of times it occurs in the ith document. Then the ℓ_1 distance is

$$d_1(A, B) = \sum_i |a_i - b_i|,$$

or the sum of the absolute difference in frequencies across every document. The minimum value of ℓ_1 distance is 0, but unlike the Jaccard distance there is no maximum ℓ_1 distance.

Two terms can have a small Jaccard distance and a large ℓ_1 distance if they occur in exactly the same documents (so that their intersections and unions are the same), but one term occurs much much more often than the other (so that the absolute differences in frequencies are large). Two terms can have a small ℓ_1 distance and a large Jaccard distance if they are both very rare (so that the absolute differences in frequencies are always 0 or small) but never occur together (so that their intersection is 0).

The analysis in the main body of this study used Jaccard distance across the full set of articles (i.e., including February 2000) to define the toxcast10, toxcast100, and toxcast1000 sets. To analyze the implications of this analytical choice, we calculate distances from "toxcast" across both the full set of articles and across a "restricted" set that excludes February 2000, using both Jaccard and ℓ_1 metrics.

Figure 17 shows scatterplots and correlation coefficients for each pair of these distance calculations, for 293 terms that are included in each at least one toxcast100 set. The plot indicates strong aggreement between full and restricted calculations, but strong disagreement between Jaccard and ℓ_1 distances. Similarly, figure 18 shows scatterplots and correlation coefficients for inclusion in the toxcast100 sets. There is strong agreement between full and restricted calculations using the same metric, but almost complete disagreement between the two metrics, i.e., the two metrics produced almost completely different toxcast100 sets.

Decisions about which metric to use had a substantial downstream effect on the findings of this study. An early version of this study used the ℓ_1 distance, rather than Jaccard distance. The former norm generates a list of toxcast100 terms that includes several obvious typos, such as "dataespecially," suggesting that the construction was substantially picking other rare terms, rather than terms that were actually semantically related to "toxcast." As indicated by table 8, Jaccard distance produces a much more meaningful set of terms. Using the ℓ_1 distance, the especially high-scoring months were January 2009, March 2011, March 2012, April 2014, and October 2014. Except for October 2014, none of these months were especially high-scoring using the Jaccard metric. Among other things, a story from March 2012 about a partnership between EPA and L'Oréal was prominent when the ℓ_1 distance was used, but no longer prominent once Jaccard distance was adopted instead. Thus, a subtle analytical decision — which mathematical function to use to calculate distances between pairs of terms — can have substantial downstream effects.

All together, the toxcast100 term lists cannot be considered robust. Their construction, use, and interpretation should be tailored carefully to the particular aims of any given analysis.

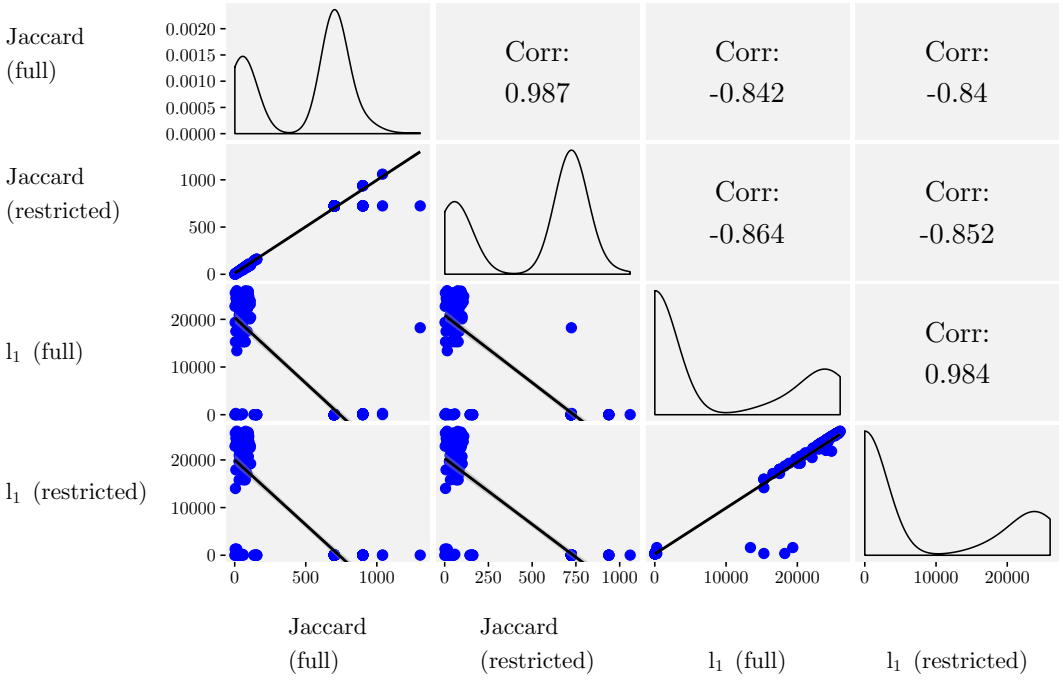

Figure 17: Pairs plot for four ways of calculating rankings/distances from "toxcast" in the BNA vocabulary, for terms included in at least one of the four toxcast100 sets. **Main diagonal**: X-axis values are rankings, with 1 = "toxcast". Curves are kernel density estimates of distribution of ranking values. **Lower triangle**: Scatterplots of ranking values across pairs of distance calculations; both axis are rankings, with 1 = "toxcast". Black lines are linear regressions. **Upper triangle**: Pearson correlation coefficients on pairs of rankings, equivalent to Spearman rank correlation coefficients on distance calculations.

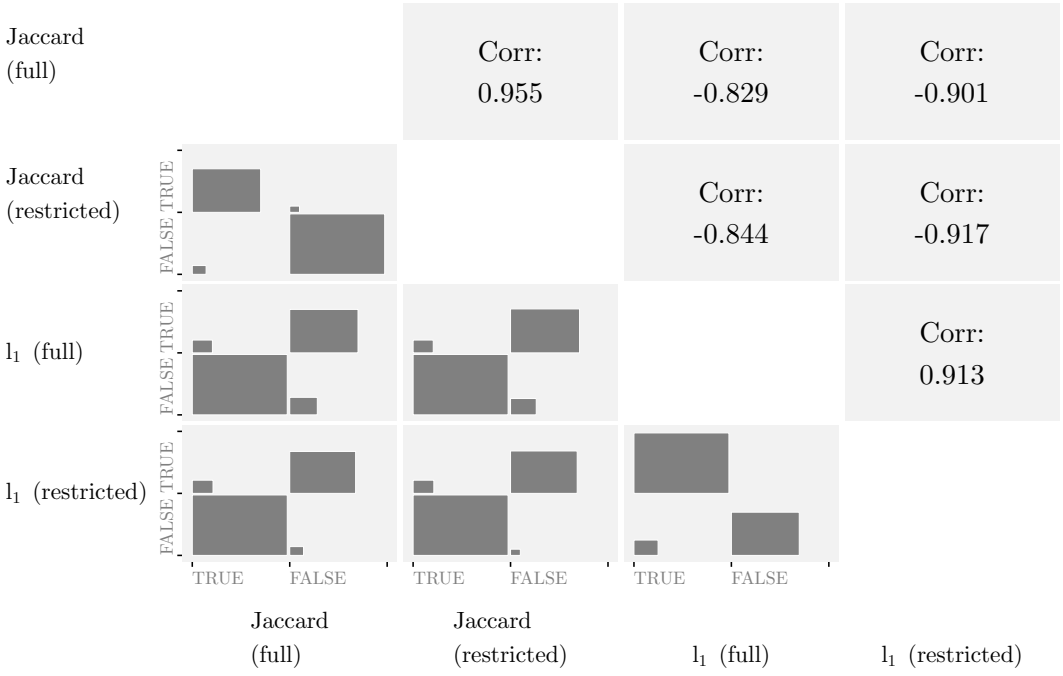

Figure 18: Pairs plot for four ways of determining membership in toxcast100 sets. **Lower triangle**: Inclusion/exclusion size plots. Size of rectangles corresponds to number of terms in each cell of the subplot. Upper-left cell corresponds to terms in both toxcast100 sets; upper-right cell corresponds to terms in the y-axis set but not the x-axis set; and so on. **Upper triangle**: Pearson correlation coefficient for set membership, taking inclusion = 1, exclusion = 2.

References

Bush, Vannevar. 1945. "Science, the Endless Frontier: A Report to the President on a Program for Postwar Scientific Research." Washington, DC: National Science Foundation.

Holbrook, J. Britt. 2005. "Assessing the Science–Society Relation: The Case of the US National Science Foundation's Second Merit Review Criterion." *Technology in Society* 27 (4): 437–51. doi:10.1016/j.techsoc.2005.08.001.

Holbrook, James Britt, and Steven Hrotic. 2013. "Blue Skies, Impacts, and Peer Review." *RT. A Journal on Research Policy and Evaluation* 1 (1). doi:10.13130/2282-5398/2914.

Kraker, Peter, Asura Enkhbayar, and Elisabeth Lex. 2015. "Exploring Coverage and Distribution of Identifiers on the Scholarly Web." *ArXiv*, March. http://arxiv.org/abs/1503.05096.

Mingers, John, and Loet Leydesdorff. 2015. "A Review of Theory and Practice in Scientometrics." *European Journal of Operational Research* 246 (1): 1–19. doi:10.1016/j.ejor.2015.04.002.

Mohammad, Saif M., and Peter D. Turney. 2013. "Crowdsourcing a Word–emotion Association Lexicon." *Computational Intelligence* 29 (3): 436–65. doi:10.1111/j.1467-8640.2012.00460.x.

Wolf, Birge, Thomas Lindenthal, Manfred Szerencsits, J. Britt Holbrook, and Jürgen Heß. 2013. "Evaluating Research Beyond Scientific Impact: How to Include Criteria for Productive Interactions and Impact on Practice and Society." *GAIA* 22 (2): 104–14.