DE GRUYTER
OPEN

Research Paper

An Automatic Method to Identify Citations to Journals in News Stories: A Case Study of UK Newspapers Citing Web of Science Journals

Kayvan Kousha†, Mike Thelwall

Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.

Citation: Kayvan Kousha and Mike Thelwall (2019). An Automatic Method to Identify Citations to Journals in News Stories: A Case Study of UK Newspapers Citing Web of Science Journals Journal of Data and Information Science, 4(3), 73–95

DOI: 10.2478/jdis-2019-0016

Received: Feb. 16, 2019
Revised: Apr. 11, 2019
Accepted: Apr. 12, 2019

Abstract

Purpose: Communicating scientific results to the public is essential to inspire future researchers and ensure that discoveries are exploited. News stories about research are a key communication pathway for this and have been manually monitored to assess the extent of press coverage of scholarship.

Design/methodology/Approach: To make larger scale studies practical, this paper introduces an automatic method to extract citations from newspaper stories to large sets of academic journals. Curated ProQuest queries were used to search for citations to 9,639 Science and 3,412 Social Science Web of Science (WoS) journals from eight UK daily newspapers during 2006–2015. False matches were automatically filtered out by a new program, with 94% of the remaining stories meaningfully citing research.

Findings: Most Science (95%) and Social Science (94%) journals were never cited by these newspapers. Half of the cited Science journals covered medical or health-related topics, whereas 43% of the Social Sciences journals were related to psychiatry or psychology. From the citing news stories, 60% described research extensively and 53% used multiple sources, but few commented on research quality.

Research Limitations: The method has only been tested in English and from the ProQuest Newspapers database.

Practical implications: Others can use the new method to systematically harvest press coverage of research.

Originality/value: An automatic method was introduced and tested to extract citations from newspaper stories to large sets of academic journals.

Keywords Citation analysis; News stories; Public engagement; Public impact; UK newspapers; Web of Science journals

† Corresponding author: Kayvan Kousha (E-mail: k.kousha@wlv.ac.uk).

Research Paper

1 Introduction

Disseminating scholarly discoveries outside of academia is important to ensure that they are widely used, to justify public financing of academic research, and to inspire future scholars. Although scientific findings are usually published in articles and books for other scholars, those that are relevant to the public may attract mass media attention (Weigold, 2001; Dudo, 2015). The number of science articles in the news has been increasing (Pellechia, 1997), perhaps reflecting the growing number of people having a degree. For example, a series of 39 surveys in the United States from 1996 through 2002 (showed that 42% of adults claimed to follow health news stories in the mass media (Brodie, 2003). It is therefore reasonable for universities and scholars to seek press coverage to promote their work. From a scientometric perspective, newspaper coverage of research may also be used as evidence of the wider impact of the cited papers (Fogg-Rogers, Grand, & Sardo, 2015) in addition to reflecting or causing (by publicity) its greater value within academia (e.g., Fanelli, 2013). It is important for research institutions to know which types of research attract press attention so that they can (a) focus marketing efforts on media-friendly types of research and (b) interpret press coverage impact evidence in the context of the type of research covered. For example, a newspaper story about a pure mathematics article might be a bigger achievement than one for an article about an astronomical discovery.

Prominent academic journals are a regular source for many newspapers (Entwistle, 1995; Bartlett, Sterne, & Egger, 2002; Lewison, 2002; Weitkamp, 2003). Prestigious health journals are frequently covered (Conrad, 1999; Lewison et al., 2008; Van Trigt et al., 1994), because they may have more relevance for the public. Major UK newspaper stories about biomedical research in 2001 cited New Scientist (7% - a magazine rather than a journal), British Medical Journal (6%), Lancet (4%) and Nature (3%) (Lewison, 2002). In 2008–9, the UK-based weekly science magazine New Scientist drew mainly from Proceedings of the National Academy of Sciences of the USA (7%), Nature (6.5%), Science (4.6%), Lancet (3.9%), and New England Journal of Medicine (2.6%) (Lewison & Turnbull, 2010). News stories about cancer in major US newspapers in 2003 mostly cited New England Journal of Medicine, Journal of the National Cancer Institute, and Journal of the American Medical Association (Moriarty, Jensen, & Stryker, 2010). Most articles in prestigious health journals attract no press attention, however. For example, only 7% of 1,193 Lancet and British Medical Journal articles (1999–2000) were reported in two UK newspapers (the Times and the Sun), mostly for women's health issues and cancer (Bartlett, Sterne, & Egger, 2002). Nevertheless, many news stories reporting research

findings do not directly cite academic articles or journals and instead name authors, researchers, research institutions or other secondary sources. It is therefore difficult to be sure about the extent to which academic articles or journals are reported by the press without reading a large sample of scientific news stories (De Semir, Ribas, & Revuelta, 1998; Lewison, Tootell, Roe, & Sullivan, 2008; Mellor et al., 2011).

Health research is a common theme in science news, whereas the social sciences attract little coverage. The weekly science section of the New York Times 1980–2000 covered medicine and health (39%), physical, earth and life science (31%), technology and engineering (21%), and history and culture (6%) (Clark & Illman, 2006). However, health news may also convey incorrect information or contain omissions or misinterpretations of published research (Chang, 2015). Science news in three US newspapers (New York Times, Washington Post, Chicago Tribune) 1966–1990 was predominantly (70%) about medicine and health (Pellechia, 1997). In the UK, most of the science news in five newspapers 2000–2001 was about health and medicine, compared with biology (18%), social sciences (11%), environment (8%) and physical sciences (8%) (Weitkamp, 2003). Social science has a similarly low share of coverage in the New York Times (11%) and Los Angeles Times (8%) weekly science sections (Evans, 1995). There is also periodic mass media coverage of individual narrow scientific areas with high levels of public interest (Schäfer, 2012).

Despite the above focused investigations into the types of research reported in the press, there is no quick method to estimate the number of citations to academic journals on a large scale from full-text newspaper articles. Previous studies have employed either content analysts or coders to investigate small random or complete samples of news stories from a narrow set of sources (e.g., Clark & Illman, 2006; Lewison & Turnbull, 2010; Mellor, Webster, & Bell, 2011; Weitkamp, 2003), or have read complete sets of news stories to identify mentions of collections of journal articles (e.g., Bartlett, Sterne, & Egger, 2002). Some previous studies have automatically identified relevant articles using keyword queries in news databases (e.g., Lexis–Nexis) to identify mentions of named medicines (Moynihan et al., 2000) but this method only works for topics with a finite set of keywords and a more general automatic method is needed. In response, this article introduces for the first time an automatic method to find press coverage of almost any area of research. This article also uses the method to analyse mentions of a substantial number of science (n=9,639) and social science (n=3,412) academic journals in a set of digitised newspaper stories (eight UK daily newspapers during 2006–2015), to update previous manual studies in a faster and easily replicable way. Previous studies have reported which journals are frequently covered in small sets of

newspaper stories. Moreover, there have been initiatives by Altmetric.com[①] and PlumX[②] to capture mentions of scholarly articles in news sources. However, it seems that article metadata (e.g., title, journal name) is rarely mentioned in news stories (e.g., 12%-13% see: Lewison et al., 2008; De Semir, Ribas & Revuelta, 1998). Hence, this article introduces an automatic method for comprehensive journal-level analysis of press uptake of scientific articles. This method can also be used to aid the evaluation of the wider impacts of academic research.

2 Research questions

This paper introduces and assesses an automatic method to estimate the coverage of published academic research in digitised newspapers on a large scale and uses it to investigate recent UK press coverage of research. Major newspapers were investigated since these are a coherent set and most likely to cover academic research. The UK was chosen as a clear unit of analysis and to ensure familiarity with reporting idioms. Relating to the two broad goals, the following specific questions are addressed, updating previous large-scale studies and broadening knowledge about press treatment of academic research using the new automatic method.

1. Can citations to scientific journals be automatically identified in news stories with an acceptable degree of accuracy?
2. How likely are different academic subjects and journals to be cited by major UK newspapers? This updates previous manual research and, for the first time, includes an analysis of journals.
3. Has newspaper coverage of academic research changed in the last decade?
4. Are recent newspaper stories citing scientific journals in-depth, positive and/or critical?

3 Data and methods

The research design was to apply the new method to identify mentions of academic journals in news stories for a large-scale systematic analysis involving many scientific journals. A follow-up content analysis was conducted to assess the accuracy of method and to give detailed information about the news stories found.

The method identifies mentions of journals in newspaper stories by searching for journal names in the ProQuest Newspapers database with a small set of curated

① https://www.altmetric.com/about-our-data/our-sources/news
② https://plumanalytics.com/tag/news-mentions/

queries, then using a program to automatically extract and filter citations from the news stories identified. For this method WoS was used as a convenient large-scale source of important journals, although any other source of non-abbreviated journal names could be used, and ProQuest was chosen for its press coverage, full text sources, and query syntax.

To apply and evaluate the new method, UK subset of newspapers in ProQuest database (*The ProQuest UK Newsstand*) was selected to have a large enough but manageable number of English language news stories. Eight major UK daily newspapers were selected: *Daily Mail, Times, Daily Telegraph, Daily Mirror, Guardian, Independent, Sun* and *Herald* (Scotland).

On a stylistic note, most newspapers have "The" within their titles (The Sun, The Guardian, The Times, The Daily Telegraph, The Independent, The Herald) but these are omitted here, and academic journals are referred to without definite articles.

3.1 Stage 1: Selection of journals and construction of ProQuest queries

Journal names were extracted from the Journal Citation Reports (JCR) Science (9,639) and Social Science (3,412) editions. Journal names were downloaded separately for each year 2006–2015 and duplicate titles were removed. In a few cases, extra abbreviations or subtitles from the original JCR Full Journal Titles were deleted to increase the power of the subsequent ProQuest citation searching, as shown in square brackets in the following examples. For example, "JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION" gave eight UK newspaper ProQuest matches, whereas over 15,000 were found when "JAMA-" was removed. The abbreviated journal name "JAMA" was not used alone because it could retrieve many false matches by capturing names such as "Mustaf Jama", "Yusuf Jama", "Mohamed Jama" or "Ibrahim Jama". However, we used journal titles containing the term JAMA with other medical specialities such as JAMA Psychiatry, JAMA Surgery, JAMA Oncology, JAMA Internal Medicine, and JAMA Dermatology. Moreover, the main English language journal names in JCR were used to generate the ProQuest citation queries for higher recall (e.g., "Journal of Orofacial Orthopedics" instead of "Journal of Orofacial Orthopedics-Fortschritte der Kieferorthopädie").

- [JAMA-] JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION
- JOURNAL OF OROFACIAL ORTHOPEDICS [-FORTSCHRITTE DER KIEFERORTHOPADIE]

UK newspapers sometimes use British English spellings of journal names in American English. For instance, "AMERICAN JOURNAL OF OBSTETRICS AND GYNECOLOGY" retrieved 73 UK newspaper ProQuest matches but the

British English spelling "GYNAECOLOGY" gave 120 more. Scientific terms that could be in journal names were therefore identified using a list of British English words that have different American English spellings (https://en.wikipedia.org/wiki/Wikipedia:List_of_spelling_variants) such as *Behavior/Behaviour, Pediatric/Paediatric/, Signaling/Signalling, and Labour/Labor*, as the examples below show. For these cases two searches were conducted to cover both British and American spellings and the results were combined.

-JOURNAL OF CELL COMMUNICATION AND **SIGNALING**
-JOURNAL OF **LABOR** ECONOMICS

For journals with "&" in their JCR titles, additional searches were conducted to retrieve newspaper mentions with "and" instead. For instance, the JCR-indexed journal "LANCET DIABETES & ENDOCRINOLOGY" had 141 UK newspaper matches, whilst the same search with "AND" ("LANCET DIABETES AND ENDOCRINOLOGY") retrieved 116 extra results.

For JCR-indexed journals with generic names, such as *Science, Nature, Cancer* or *Cell,* the term "journal" was added before and after their names as a practical method to reduce false matches during ProQuest full text searching. From initial investigations, news stories typically add the term "journal" before, and occasionally after, journal names when reporting research from journals with names that are not clearly academic, such as *"published in the journal **Science**", "reported in the journal **Nature**", or "a finding in the medical journal **Cancer**"*. For instance, there were 66 ProQuest hits for "Nature journal" in the eight UK newspapers during 2006–2015 and 4,259 for "journal Nature". Additional examples are shown below.

-journal BREAST CANCER RESEARCH AND TREATMENT
-journal LANGUAGE SPEECH AND HEARING SERVICES IN SCHOOLS

The term "journal" was not added to publications containing the similar terms *Journal, Quarterly, Annual, Proceedings, Bulletin, Letters* or *Archives*, or the distinctive terms *Acta, ACM, ACS, IEEE, BMC, JAMA, BMJ* or *Revista* or the small set of journals *PLOS Biology, LANCET Infectious Diseases* and *Harvard Business Review* because their names were unique enough for effective searches. For instance, the ProQuest query "journal PLOS Biology" gave 558 newspaper matches, missing 4,600 relevant results from the query "PLOS Biology".

3.2 Stage2: ProQuest searches for news stories mentioning academic journals

To locate mentions of the 13,051 scientific journals in (full text) news stories, the above journal queries were submitted separately in the full text field ("FT") of

ProQuest "Command Line Search" interface as phrase searches in October 2016. Extra search commands were added to the query to limit the results to each of the eight selected UK newspapers and to each year during 2006–2015, as in the example below.

> **FT**("journal BEHAVIORAL AND BRAIN SCIENCES" OR "journal BEHAVIOURAL AND BRAIN SCIENCES" OR "LANCET GLOBAL HEALTH" OR "JAMA PSYCHIATRY" OR "AMERICAN JOURNAL OF PSYCHIATRY" OR "journal ALZHEIMERS & DEMENTIA" OR "journal ALZHEIMERS AND DEMENTIA" OR ...) AND **PUB**("Daily Mail") AND **YR**(2015)

Because the maximum number of Boolean operators in a single ProQuest search was 1,000, fourteen separate searches were conducted for each newspaper/year combination to cover all journals in the data set. Irrelevant publications, such as the Sunday editions *Independent on Sunday* and *Sunday Herald*, were filtered out with the ProQuest "Narrow results" menu. The full text ProQuest records for the search results were downloaded to generate a set of news stories potentially mentioning one or more of the 13,051 academic journals.

3.3 Stage 3: Identifying correct journal matches

The Stage 2 method returns some false and ambiguous results. For instance, the query FT("JOURNAL OF CANCER") matched multiple journals (e.g., EUROPEAN JOURNAL OF CANCER, BRITISH JOURNAL OF CANCER). Adding the term "journal" after journal names ("Science journal" or "Nature journal") also sometimes retrieved many false matches for journals with generic names (e.g., "published in the science journal Advanced Materials Interfaces"). To automatically identify journal names in the full text of each news story, a program was written and added to the free Webometric Analyst software (http://lexiurl.wlv.ac.uk: Services option, "News: Extract journal mentions from ProQuest records") to extract correct mentions of journals. This program scans news stories for mentions of one or more of the terms, "journal", "bulletin", "acta", "annals", "transactions", "proceedings", "zeitschrift", or "lancet" and then uses heuristics to identify if the text before, after or surrounding is a journal name. The heuristics rely upon journal names being in title case (initial capital letters except for small words) and incorporate a list of rules to eliminate common text segments that these rules often mistakenly add (e.g., "Writing in the", "Reporting in the", "Dr", "Prof", "Professor", "University").

The list of journal names extracted was matched against the 13,051 pre-selected journal names using Webometric Analyst (see the Tab-sep option, "Split file"> "Split file 1 based upon…") to exclude non-JCR publications (e.g., Indian Journal

of Scientific and Industrial Research) and other noun phrases (e.g., American Heart Association) that were incorrectly extracted by the software. This gave a list of newspaper stories with at least one citation to a Science Citation Index (SCI) or Social Science Citation Index (SSCI) journal.

3.4 Stage 4: Removing duplicate and near duplicate stories

There are duplicate news stories from different editions of the same newspaper in ProQuest. To avoid counting duplicate citations, stories with duplicate titles in different editions (e.g., Scotland, Eire or Ulster regions) were deleted.

Some news stories had the same contents but slightly different headlines, as shown in the Daily Mirror example below. Hence, news stories with the same abstract as a previous story were removed (in the example below: "Bosses at the Prostate Cancer Charity point out that men who ate the most eggs in the study were also more likely to smoke, be overweight, take less exercise and have a poor diet - all risk factors for the disease which kills 10,000 in the UK annually."). Such title changes are presumably minor copyedits by regional editors.

-3 EGGS A WEEK UPS PROSTATE CHANCE BY 80%
-3 EGGS A WEEK UPS PROSTATE RISK 80%

3.5 Content analysis of news stories citing academic journals

A content analysis was conducted on a random sample of 360 newspaper stories (45 from each of the eight UK newspapers) mentioning a WoS science journal to assess whether newspapers mention scientific journals in their articles to communicate research results and, if so, how newspapers discuss research. The content analysis also estimated the accuracy of the method for capturing journal names in the text of newspapers. Three independent coders with social science PhDs conducted the content analysis after a pilot study on a random sample of 40 newspaper stories (not in the main sample of 360 news stories) to refine the classification procedure. Cohen's kappa values were calculated between the three coders for all classes and compared with standard heuristic guidelines (at least 0.21+ indicates fair agreement, 0.41+ indicates moderate agreement, 0.61+ indicates substantial agreement and 0.81+ indicates almost perfect agreement) (Landis & Koch, 1977). The following classes were used.

1. Research citations, non-research citations and errors:

1.1. Research citation: Story mentions the journal for its published scientific findings (e.g., "*The research team at Penn State University in the US, whose findings were published in the Journal of Clinical Oncology, also found a 57 per cent reduction in deaths among men aged 65 to 75.*").

1.2. Non-research citation: Story mentioning the journal in a non-research/non-scholarly context (e.g., "*More than 100 doctors, including 20 professors, have signed an open letter in the British Medical Journal, criticising the leadership of the British Medical Association, the doctors' union.*").
1.3. Software error capturing journal names: (e.g., "*In her journal, Nature was described in detail on a daily basis.*")

2. The extent of research coverage (just for category 1.1):

2.1. Brief description of the research: Story giving only the main outcome and perhaps a small amount of other information about it.
2.2. Extensive description of the research: Story giving a significant amount of information about the research findings, such as the main result and at least one of the following: (A) the main characteristics of the sample or dataset used, (B) at least two main quantitative results (percentage, proportions and other indicators), or (C) a substantial amount of other descriptive information about the published article, especially for qualitative papers, but not necessarily just for these.

3. Sources reported in the news story (just for category 1.1):

3.1. Only one journal article used in the report.
3.2. Extra sources of information used, such as multiple journal articles or an interview with the authors or other experts.

4. Overall sentiment of the conclusion of the news story about the implications of the research (just for category 1.1):

4.1. Good news: Story reflects promising scientific results such as a treatment, a new surgery technique, cancer drug/diagnosis that is likely to improve life for some people (e.g., "*Vegetable protein helps in fight against strokes and heart disease*").
4.2. Bad news: Story reflects risks, warnings, dangers, and evidence of treatments not working or other bad news (e.g., "*People born in spring more at risk of suicide*").
4.3. Other: Story reflects neither good nor bad news, but just gives general information (e.g., "*Unknown species of man identified from cave DNA*").

5. Research quality judgment (just for category 1.1):

5.1. High quality research: Story states that the research is good, using words like excellent, innovative, good, impressive, impressive, ground-breaking, world-leading, and award-winning. The judgement may be of the study or its researchers.

5.2. Low quality research: Story states that the research is poor, using words like flawed, bad, unimpressive, retracted, fraudulent, trivial, pointless. The judgement may be of the study or its researchers.
5.3. No judgement about the quality of the research or researchers.

4 Results

In answer to the main research question, 100% of the news stories extracted by the method in the manually checked random sample were technically correct (i.e., mentioned an academic journal) and 94% cited the journal to discuss its research. Thus, the method has a high precision (few false matches), although its recall (percentage of matches found) is unknown. More details are given in the content analysis section, after a discussion of the overall results.

4.1 Newspaper citations to WoS Journals

Only 5.2% of Science and 5.7% of Social Science WoS journals were cited at least once by a daily UK newspaper 2006–2015. More WoS Science journals were cited by the Daily Mail (Science: 8.8%; Social Science: 9.1%), Daily Telegraph (7.6%; 8.9%) and Times (7.3%; 8.1%) than the Sun (2.1%; 1.4%), Daily Mirror (3.4%; 1.7%), and Herald (3.9%; 2.9%) (Figure 1), showing large differences between newspapers. No major UK newspaper restricts its coverage to a small number of elite journals, however. Perhaps surprisingly, the tabloid Daily Mail has the widest coverage of academic journals.

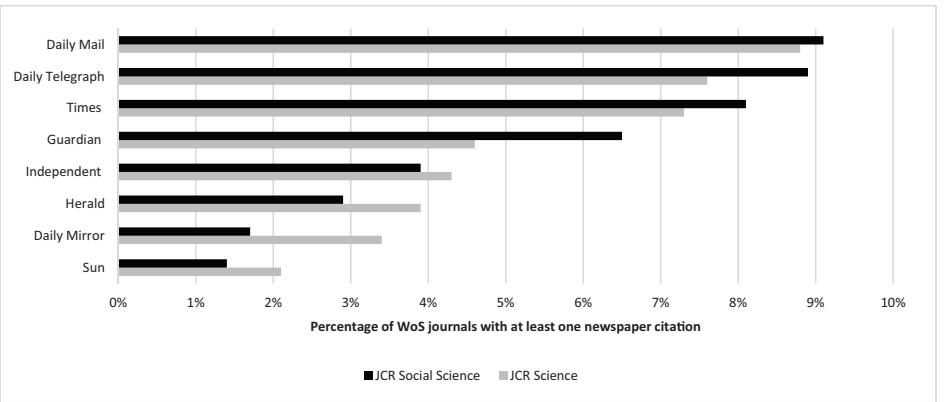

Figure 1. The percentage of WoS Science and Social Science journals (JCR: 2006–2015) with at least one identified citation from (eight) UK newspapers during 2006–2015, based on the ProQuest UK Newsstand database.

Although more Science journals were cited than Social Science journals (Figure 2), in comparison to the number of journals in WoS, a slightly higher proportion of Social Science journals were cited by the UK newspapers (the fourth column). It is therefore possible that the lower total Social Science press coverage is due to the smaller amount of published Social Science articles in WoS or overall.

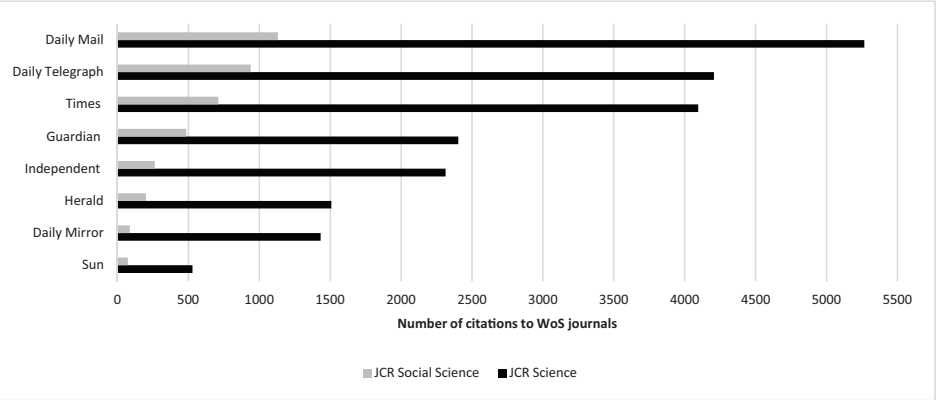

Figure 2. The number of identified citations to Science and Social Science WoS journals in eight UK newspapers during 2006–2015 based on the ProQuest UK Newsstand database.

4.2 Subject areas of WoS journals cited by newspapers

In order to estimate the main academic fields of the cited journals, each was assigned to one of 22 Essential Science Indicators (ESI) research fields[3] (see, Table A in the online appendix, https://doi.org/10.6084/m9.figshare.4796548.v4). To simplify the presentation of the results, related subject areas from 22 ESI fields were combined to represent six broad fields (Figure 3). For instance, the subjects Clinical Medicine, Pharmacology & Toxicology, Psychiatry, Immunology and Neuroscience in the ESI classification scheme were combined to form Medical Sciences and the subjects Plant & Animal Science, Molecular Biology & Genetics, Biology & Biochemistry, Environment/Ecology, Agricultural Sciences, Geosciences and Microbiology were merged as Biological Sciences. Popular science journals, such as Science, Nature and PLOS ONE were classified as Multidisciplinary based on their ESI classification. Overall, most (55%) of the WoS Science journals with at least one citation from a UK newspaper during 2006–2015 were from Medical and Health Sciences and over a quarter (28%) were from Biological Sciences, although there are differences between newspapers. This confirms that medical research is more widely reported in major UK newspapers than other academic

[3] http://ipscience-help.thomsonreuters.com/incitesLiveESI/8289-TRS.html

topics Nearly all WoS Science journals categorised as Social Sciences (8%) in the ESI classification scheme are about public health and epidemiology (e.g., BMC Public Health, American Journal of Epidemiology, Cancer Epidemiology Biomarkers and Prevention), and are indexed in the WoS Science Citation Index as Public, Environmental & Occupational Health rather than the same category (Public, Environmental & Occupational Health) in the WoS Social Science Citation Index. Hence, the Social Sciences category here primarily reflects newspaper citations to public health journals.

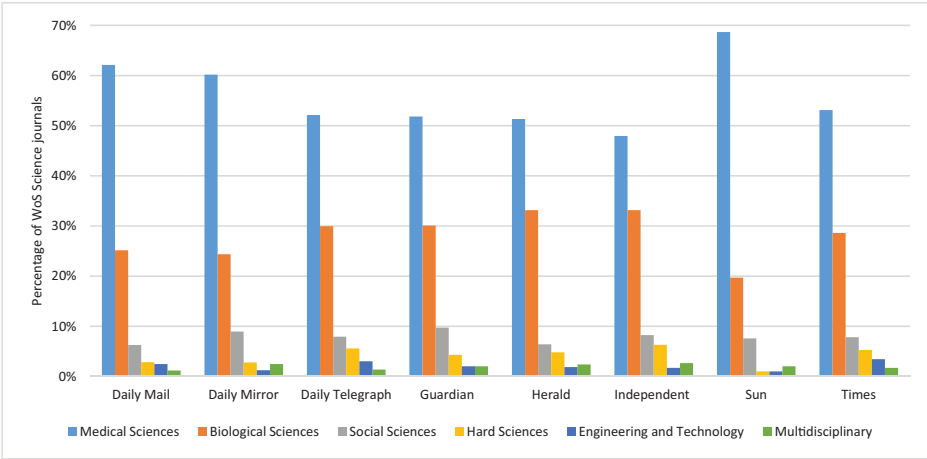

Figure 3. The percentage of WoS Science journals with at least one identified citation from UK newspapers during 2006–2015, by broad subject area.

From the Social Sciences category (Figure 4 and Table B in the online appendix), journals with at least one newspaper citation were mainly from Psychiatry/Psychology (43%), Social Sciences (37%), and Economics & Business (10%) based on their ESI fields. Psychiatry/Psychology is particularly interesting to the tabloids: Sun (54%), Daily Mail (50%), and Daily Mirror (49%). The slightly left-leaning broadsheets Guardian and The Independent cited more Social Science, General journals. These included *Foreign Affairs, British Journal of Sociology, Antiquity* (a journal about world Archaeology) and *British Journal of Political Science*. A few (10%) of the cited journals have a science classification (e.g., Clinical Medicine, Environment/Ecology, and Engineering) due to the ESI classification scheme, such as *Addiction*, *Alcohol and Alcoholism* (classified as Clinical Medicine in ESI) or Energy policy and Applied Ergonomics (classified as Engineering in ESI).

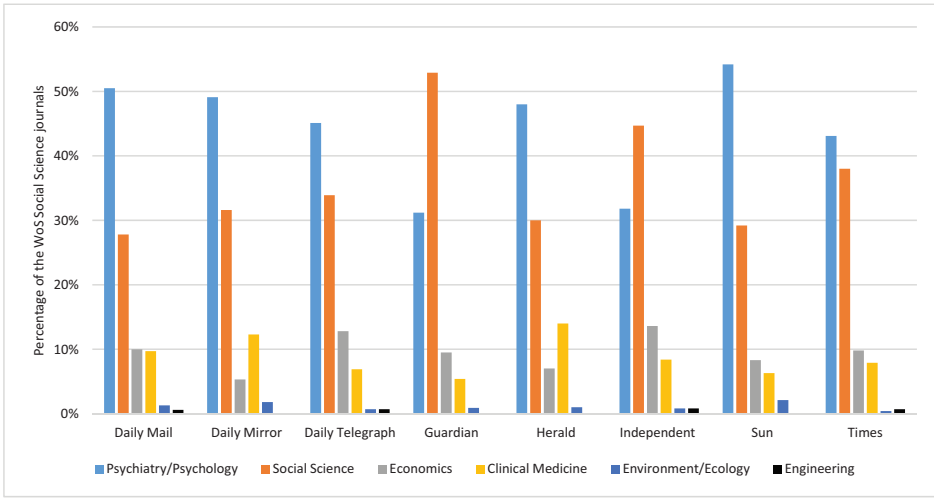

Figure 4. The percentage of WoS Social Science journals with at least one identified citation from UK newspapers during 2006–2015, by broad subject area.

4.3 Journals most cited by major UK newspapers

Within Science, British Medical Journal (now officially called The BMJ) had the most citations from all newspapers, ranging from 12% in the Daily Mail and Daily Telegraph to 22% in the Sun (Figure 5). The multi-disciplinary journals Nature, Science, and PLOS ONE were also frequently cited. Table C in the online appendix (https://doi.org/10.6084/m9.figshare.4796548.v4) shows the additional general medical journals New England Journal of Medicine (NEJM), Lancet and Journal of the American Medical Association (JAMA), and specialist scientific journals, like British Journal of Cancer, British Journal of Sports Medicine, Current Biology, and Neurology. The overall results confirm the importance, but not dominance, of a few prestigious journals for UK newspapers. However, there are large differences. For instance, half of the citations to science journals from the Independent (52%) and Guardian (48%) were to seven world-leading science and medical journals (BMJ, Nature, Science, NEJM, Lancet, JAMA and PLOS ONE), in comparison to less than a third from the Daily Mail (29%), Sun (32%), and Herald (33%).

In Social Sciences, the WoS journals most frequently cited by UK newspapers were about psychology, psychiatry, epidemiology, public health, medical ethics and climate change, such as British Journal of Psychiatry, Psychological Science, Journal of Epidemiology & Community Health (a BMJ publication), Journal of Medical Ethics (a BMJ publication), and Nature Climate Change. Figure 6 and Table D in the online appendix (https://doi.org/10.6084/m9.figshare.4796548.v4) show that there are also some social science journals from other specialisms with many

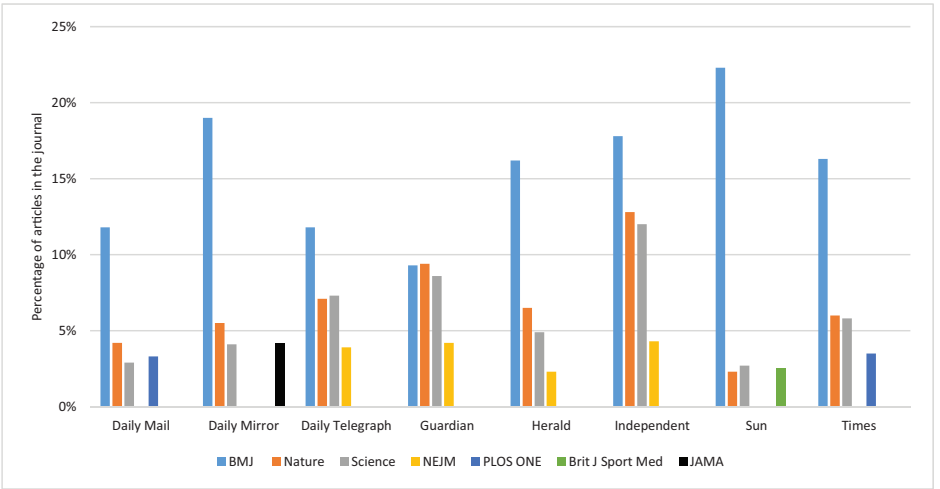

Figure 5. The seven WoS Science journals with the most identified citations (%) from the eight UK newspapers during 2006–2015 in ProQuest UK Newsstand.

citations from UK newspapers, such as Journal of Consumer Research, The Economic Journal, and British Journal of Sociology. This suggests that most UK newspapers were most interested in reporting public health (e.g., psychology and psychiatry) but also had some interest in the environment, economics and business.

There are some differences in reporting social science research between newspapers. For instance, 14% of the citations from the Times were to Journal of Archaeological Science (9%) and Antiquity (3.6%) and more tobacco research in the journal Tobacco Control was cited by the Sun (5.5%) and Daily Mirror (3.6%) than by other newspapers.

4.4 Newspaper citations to journals over time

There has been no universal and systematic overall increase or decrease in terms of the numbers of journals reported in the UK press, although there are patterns for individual newspapers (Figures 7 and 8). The Times seems to have decreased its coverage in both broad areas 2009–2013 and the Daily Telegraph seems to have increased 2010–2014.

The coverage of newspapers by ProQuest was checked to assess whether changes over time could explain the citation variations. There were five times more news stories from the Times indexed by ProQuest during 2006 (about 117,000), 2007 (108,000) and 2008 (105,000) compared with 2009–2015 (19,000 to 20,000), indicating that the fewer indexed news stories during 2009–2012 was the cause of fewer citations to Science and Social Science from the Times during this period but it does not explain the increase 2013–2015 (Figures 7 and 8). This seems to reflect

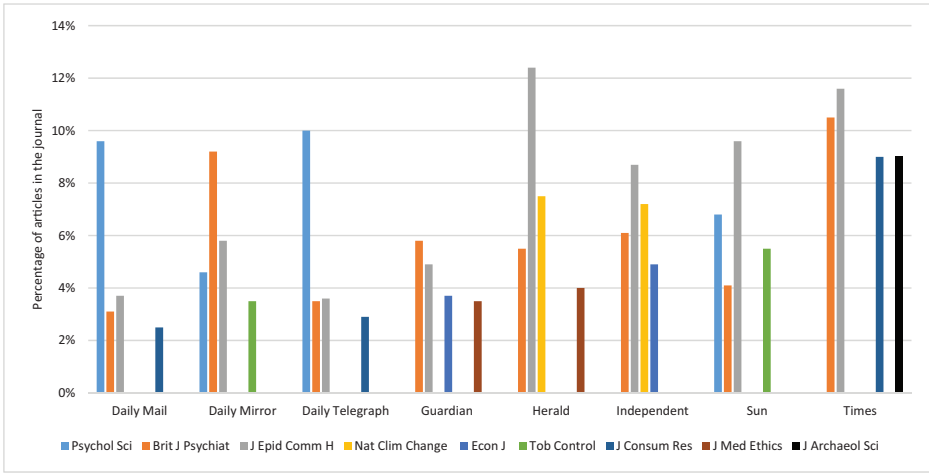

Figure 6. The nine WoS Social Science journals with the most identified citations from eight UK newspapers during 2006–2015 in ProQuest UK Newsstand.

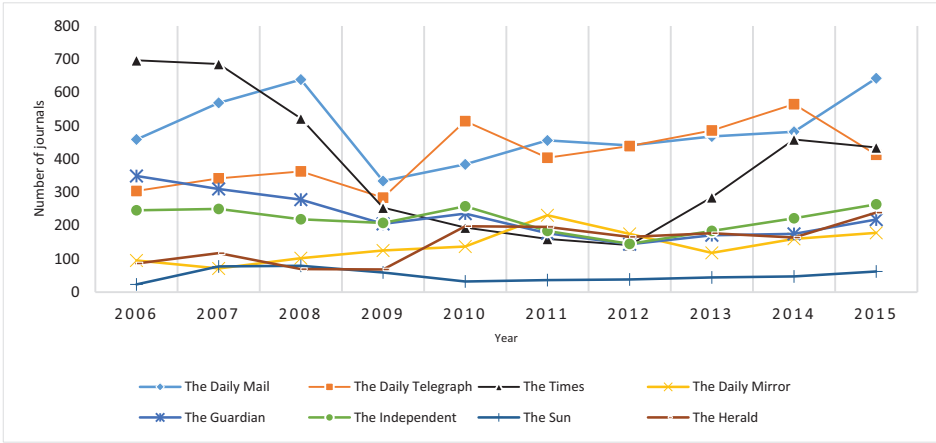

Figure 7. The number of Science journals mentioned by UK newspapers during 2006–2015.

deliberate policy shifts or staff changes within ProQuest rather than an 80% reduction of news content in The Times in 2009.

4.5 Topics of citing news stories

To get a simple overview of the topics of news stories citing at least one journal, all news headline terms were extracted and the Porter algorithm (Porter, 1980) was used to convert plural English terms to singular and to remove verb suffixes (e.g., ing, ed and tion). Term frequencies were calculated for each newspaper and for

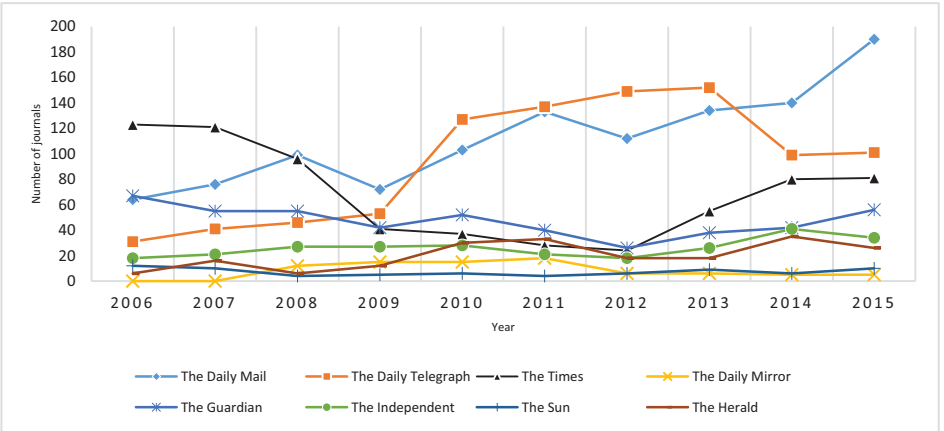

Figure 8. The number of Social Science journals mentioned by UK newspapers during 2006–2015.

Science and Social Science journals separately. Small function words, such as "the", "of", "to", and "in", were then manually removed and common irregular nouns were combined (e.g., woman and women or child and children).

For the news headlines of stories citing WoS Science journals, medical terms are the most frequent, but the Guardian and Independent also had terms relating to climate change, and global warming, as well as dinosaur and fossil (Tables E and F in the online appendix, https://doi.org/10.6084/m9.figshare.4796548.v4). Medical terms described common major diseases, including cancer (e.g., breast and prostate cancers), heart (e.g., heart attack), brain (e.g., stroke), Alzheimer's, diabetes, and dementia. These are probably mentioned in the contexts of warnings rather than cures. For instance, Daily Mail terms risk (frequency: 552), warn (70), death (66), kill (56), harm (53) and danger (53) are more common overall than cure (87), hope (63) and help (221). Medical issues concerning children and women are more apparent than those for men. For instance, child (205 including children), baby (151), woman (176 including women) and mother (65) were more frequent in the headlines of Daily Telegraph news stories than man (111 including men), although these may include some non-medical stories. For Social Science journals, women, children, men and mental health issues such as depression and suicide were common.

4.6 Content analysis of news stories citing journals

Following the Landis and Koch (1977) guidelines, inter-coder agreement between the three coders for the first category (*Research citations, non-research citations and errors*) was either substantial (0.61–0.80) or perfect (0.81–1) and for the second (*The extent of research coverage*) and fourth (*Sentiment of research in news*)

categories the agreement was mainly moderate (0.41–0.60) or substantial (0.61–0.80). The agreement rates between coders for the third *(Sources reported in the news story)* and fifth *(Research quality judgment)* categories include many cases of *fair* agreement (0.21–0.40) and no agreement (less than 0.20). This is probably due to the unbalanced nature of the categories (most in one class) and the subjective nature of the judgements for these classes and hence the results (averaging the three coders) should be interpreted cautiously (see Table G in the online appendix, https://doi.org/10.6084/m9.figshare.4796548.v4).

Nearly all (94%) of the sampled news stories had cited WoS Science journals to report their research findings, with fewer for the Guardian (86%) and the Independent (88%) (Table 1). About 6% of the journals were mentioned in non-research contexts, such as *"Bernard Lagan Sydney The editor of the Medical Journal of Australia has been dismissed after he objected to outsourcing the 101-year-old journal's production to Reed Elsevier"*. No cases of software errors in capturing journal names were identified in the sample.

The descriptions of 59% of the reported journal articles were extensive. The Herald (83%), Independent (74%) and Daily Telegraph (69%) had more extensive reports than the Sun (33%) and Daily Mirror (36%), for example giving more findings or discussing methods. Most newspaper stories had used more than one source (53%) rather than a single journal article (47%). The Guardian (67%), the Independent (62%) and Daily Mail (62%) used multiple sources more often than did the other newspapers.

Similar shares of the newspaper stories reflected good, bad or other aspects of research. However, good news was more common in the Daily Mail (48%) such as "*Cherry aid for the heart; A slice of pie could cut cholesterol and reduce your risk of diabetes*". In contrast, the Sun (47%) reported more bad news, such as "*A lack of international will means the chances of bringing climate change under control may already be slipping out of reach."*.

Most stories (94%) made no comment about the quality of the research covered. However, 5.5% of the reported research was explicitly judged to be important or high quality, such as "*Scientists have created a revolutionary drug that could restore sight to the blind*" and "*In a landmark study,...*". Less than 1% of the stories clearly referred to faults or weaknesses, such as "*The author of an article in the British Medical Journal which incorrectly claimed that statins cause side effects in 20 per cent of patients has hit out at the leading academic who made a complaint to the journal - accusing him of taking "a biased view" on the usefulness of the drugs*."

Table 1. Common reasons for newspaper stories citing journals based on a faceted content analysis of a random sample of 360 newspaper stories mentioning a WoS Science journal.

Broad categories	Narrow classes	Daily Mail	Daily Teleg.	Times	Daily Mirror	Guard.	Indep.	Herald	Sun	%Total
Types of citation	Research	97.8%	98%	89.6%	100.0%	85.9%	88.1%	97.0%	95.6%	94.0%
	Non-res.	2.2%	2%	9.6%	0.0%	14.1%	11.9%	3.0%	3.7%	5.8%
Coverage of research	Brief	37.9%	31.1%	41.7%	63.7%	41.7%	25.6%	17.1%	66.7%	40.7%
	Extensive	62.1%	68.9%	58.3%	36.3%	58.3%	74.4%	82.9%	33.3%	59.3%
Sources used	Single	37.9%	58.3%	53.3%	54.8%	33.3%	37.6%	54.3%	50.0%	47.4%
	Multiple	62.1%	41.7%	46.7%	45.2%	66.7%	62.4%	45.7%	50.0%	52.6%
Type of news (sentiment towards res.)	Good	48.5%	31.1%	19.2%	38.5%	32.4%	35.0%	26.4%	23.8%	31.9%
	Bad	24.2%	32.6%	37.5%	30.4%	41.7%	29.9%	42.6%	46.8%	35.7%
	Other	27.3%	36.4%	43.3%	31.1%	25.9%	35.0%	31.0%	29.4%	32.4%
Research quality judgment	High	8.3%	0.0%	2.5%	3.7%	8.3%	14.5%	2.3%	4.0%	5.5%
	Low	0.8%	0.0%	2.5%	0.7%	1.9%	0.9%	0.0%	0.0%	0.8%
	Other	90.9%	100.0%	95.0%	95.6%	89.8%	84.6%	97.7%	96.0%	93.7%

5 Limitations

The newspaper citation search method in this study has a several limitations. Heuristics were used to identify mentions of journals in news stories. Although the manual checks of a sample of the results found no false matches (i.e., an estimated 100% technical precision), there might be a few errors in capturing journal names in the whole data set. For instance, the software may identify journal names incorrectly in some complex cases, such as in "published in Scientific Reports, a Nature journal", where Scientific Reports is a correct match and Nature is not. For practical reasons, it was not possible to search generic and short journal names (e.g., Science, Nature or Cell) without adding the term journal to the ProQuest query to avoid downloading a huge number of news stories using these above terms in other contexts ("Science fiction stories are", "Nature conservation issues are"). Nevertheless, there could also be relevant results not mentioning the term *journal*, such as "a study published in Science" or "The findings, which have been published in Science magazine", where these are indirectly flagged as academic journals.

Many news stories do not explicitly name their sources. For instance, internationally, newspapers (The New York Times, Le Figaro and Le Monde, El País and La Vanguardia, La Repubblica, and the International Herald Tribune) seem to cite journal articles in about 13% of their scientific news stories (De Semir, Ribas, & Revuelta, 1998) and only 12% of US newspaper stories in 2003 about cancer (n=3,600) identified journal articles (Lewison, Tootell, Roe, & Sullivan, 2008). A content analysis of 640 BBC News stories with science themes in 2010 also showed that a third named a publication. Online news stories were more likely to mention journal names, however. Two thirds of online BBC news stories about research explicitly mentioned journal names. Only a fifth of the stories naming journal

articles had hyperlinks to the main articles online (Mellor, Webster, & Bell, 2011). The recall of the method (i.e., the percentage of news stories about research that the method found) is therefore probably low and almost certainly lower than studies in which complete sets of newspaper articles were individually read to identify mentions of academic research (e.g., Bartlett, Sterne, & Egger, 2002). The level of recall also probably varies between journals and may be lower for those with long names that newspapers might abbreviate or with short names that heuristics were used to identify (e.g., adding "journal" before). Moreover, it was not practical to estimate the overall recall of searches for individual journals with generic names, such as Science, Nature, Cancer or Cell. This is because searching above journal names in the ProQuest Newspapers database gives many results. For instance, searching the term "Science" in the full-text of the eight UK newspaper during 2006–2015 retrieves over 162,000 news stories which requires extensive manual work to download and to identify relevant results.

Editorial policies of newspapers about mentioning academic sources might have influenced the results and may change over time. For instance, newspapers may interview authors or researchers about their published research without mentioning journal names or article titles. Hence, newspapers with more emphasis on mentioning research sources would presumably have a larger presence in the overall results. The recall of the method may therefore differ between newspapers, making comparisons between them potentially misleading.

There are some discrepancies between the ESI classification for journals and their subject categories within SCI or SSCI. The subject category Public, Environmental & Occupational Health is in both WoS SCI and WoS SSCI. Some journals have been indexed in Public, Environmental & Occupational Health in WoS SCI rather than WoS SSCI. However, ESI has classified them as Social Sciences, General, presumably because there is no public health category in the 22 ESI research fields. Such issues are a limitation with using ESI for subject analyses of journals.

Finally, the ProQuest UK Newsstand database seems to be incomplete for some of the newspapers and years, so comparisons over time are not reliable.

6 Discussion and Conclusions

In answer to the first research question, the content analysis showed that the method used here captures mentions of academic journals in news stories with high levels of precision (94% technically correct, meaningful matches) but probably low recall (a majority of relevant news stories were probably not found). Since the method is general, it could also be used to extract citations to journals from other news sources in text format such as US newspapers indexed by ProQuest or other online news sources (e.g., BBC, CNN or Reuters).

In answer to the second research question, few WoS Science (5.2%) and Social Science (5.7%) journals were cited by any of the eight UK daily newspapers 2006–2015. Whilst it was already known that some journals are more likely to be covered by the press, this is the first large scale evidence that most are completely ignored. The results also show differences between newspapers. For instance, the Daily Mail had the most and the Sun had the fewest citations to different academic journals. About half of the cited Science (42%) and Social Science (43%) journals in the eight UK newspapers were medical (including Clinical Medicine, Immunology and Pharmacology) or about Psychiatry/Psychology respectively, indicating that UK newspapers tend to report medical and mental health care research more commonly than other subjects. The British Medical Journal (BMJ) was the most frequently cited journal (from 12% in both the Daily Mail and Daily Telegraph to 22% in the Sun). New England Journal of Medicine and other specialist British journals, such as Lancet, British Journal of Cancer, British Journal of Sports Medicine, and The British Journal of Psychiatry were also frequently reported. Within Social Sciences, a range of Psychiatry/Psychology journals were frequently reported in news stories, such as British Journal of Psychiatry and Journal of Epidemiology & Community Health (a BMJ journal). This suggests that UK newspapers tend to report a significant proportion of research from British journals (see also Entwistle, 1995) in addition to prestigious multi-disciplinary journals like Nature and Science. One reason might be that UK newspapers prefer to report research findings that have more general interest or benefit for public in the UK such as public or mental health issues and British journals may publish more research of this kind. Another reason could be that science journalists or editors have more access to British-based information from journals, such as publishers' press releases, authors or experts in the field for writing up their news stories. For instance, there is evidence that UK science writers more frequently look at popular science journals such as Nature and Science in addition to the UK's leading medical journals, such as BMJ and Lancet for ideas for their reports (Weitkamp, 2003). Discoveries by UK academics may also be intrinsically more newsworthy for UK newspapers and UK university press offices may have more contacts with UK-based science journalists.

In answer to the third research question, overall the Daily Mail and Daily Telegraph had the most coverage of scientific research over time, eclipsing the broadsheets that may be thought of as the more logical source of research news, whereas the Sun had the lowest. During the period analysed there was no general trend for an increase or decrease in the coverage of science, although there was an increase for one newspaper and a decrease for another. In the current era of increasing competitiveness of newspapers and cost cutting, it is encouraging that coverage of research has not suffered, at least in major UK daily newspapers.

In answer to the fourth research question, almost all academic journals mentioned in the sampled news stories were cited to report research and only 6% were used in a non-research context. Most news stories reported research extensively (about 60%) and beyond one source (53%), typically other published research or interviews with experts. There were differences in the reporting good or bad implications from the research between newspapers, with the Daily Mail for reporting more good news (48%) than bad news (24%) compared with the Sun reporting more bad news (47%) than good news (24%). Nevertheless, the keyword-based topic analysis of news stories citing Science journals suggests that risk factors for cancer, heart diseases and medical issues about children and women were particularly common topics in all newspapers (Tables E and F in the appendix). This is the first reported comprehensive analyses of these issues.

In terms of practical implications for academics, the results emphasise the difficulty in getting major press coverage of published research, since most journals were ignored. Health continues to be an exception, as are prestigious journals, and so work fitting these criteria should continue to be promoted to the media. The most encouraging news is for the social sciences. Although the study confirms the ongoing low level of coverage of social research, as a proportion of SSCI/SCI journals the difference between the sciences and social sciences is small. Hence, social scientists that publish health related research or in prestigious journals should not be discouraged from promoting their work to the press. Articles are likely to be discussed in reasonable depth using additional sources but without subjective evaluations. Thus, press releases that point to full text versions of papers and make available other sources of information might be particularly helpful for journalists.

Finally, the automatic method introduced here could help researchers, universities, news organisations, governments and scientific publishers to assess the press uptake of published research. With the new method, this can be achieved more quickly and on a larger scale than before.

Author Contributions

Kayvan Kousha (k.kousha@wlv.ac.uk) developed the research questions and methods, collected the data, conducted the analysis and wrote the main body of the paper. Mike Thelwall (m.thelwall@wlv.ac.uk) deigned a tool for citation extraction from digitised newspapers and helped to write the paper.

References

Bartlett, C., Sterne, J., & Egger, M. (2002). What is newsworthy? Longitudinal study of the reporting of medical research in two British newspapers. British Medical Journal, 325(7355), 81–84.

Brodie, M., Hamel, E. C., Altman, D. E., Blendon, R. J., & Benson, J. M. (2003). Health news and the American public, 1996–2002. Journal of Health Politics, Policy and Law, 28(5), 927–950.

Chang, C. (2015). Inaccuracy in health research news: A typology and predictions of scientists' perceptions of the accuracy of research news. Journal of Health Communication, 20(2), 177–186.

Clark, F., & Illman, D. L. (2006). A longitudinal study of the New York Times science times section. Science Communication, 27(4), 496–513.

Conrad, P. (1999). Uses of expertise: Sources, quotes, and voice in the reporting of genetics in the news. Public Understanding of Science, 8(4), 285–302.

De Semir, V., Ribas, C., & Revuelta, G. (1998). Press releases of science journal articles and subsequent newspaper stories on the same topic. Journal of the American Medical Association, 280(3), 294–295. doi:10.1001/jama.280.3.294

Dudo, A. (2015). Scientists, the media, and the public communication of science. Sociology Compass, 9(9), 761–775. doi:10.1111/soc4.12298

Entwistle, V. (1995). Reporting research in medical journals and newspapers. BMJ, 310(6984), 920.

Evans, W. (1995). The mundane and the arcane: Prestige media coverage of social and natural science. Journalism & Mass Communication Quarterly, 72(1), 168–177.

Fanelli, D. (2013). Any publicity is better than none: Newspaper coverage increases citations, in the UK more than in Italy. Scientometrics, 95(3), 1167–1177. doi:10.1007/s11192-012-0925-0

Fogg-Rogers, L., Grand, A., & Sardo, M. (2015). Beyond dissemination—science communication as impact. Journal of Science Communication, 14(3), C01–CO7.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Lewison, G. (2002). From biomedical research to health improvement. Scientometrics, 54(2), 179–192. doi:10.1023/A:1016005710371

Lewison, G., & Turnbull, T. (2010). News in brief and features in new scientist magazine and the biomedical research papers that they cite, August 2008 to July 2009. Scientometrics, 85(1), 345–359.

Lewison, G., Tootell, S., Roe, P., & Sullivan, R. (2008). How do the media report cancer research? A study of the UK's BBC website. British Journal of Cancer, 99(4), 569–576.

Mellor, F., Webster, S., & Bell, A. R. (2011) Content analysis of the BBC's science coverage. Science Communication Group, Imperial College London. http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/our_work/science_impartiality/appendix_a.pdf

Moriarty, C. M., Jensen, J. D., & Stryker, J. E. (2010). Frequently cited sources in cancer news coverage: A content analysis examining the relationship between cancer news content and source citation. Cancer Causes and Control, 21(1), 41–49.

Moynihan, R., Bero, L., Ross-Degnan, D., Henry, D., Lee, K., Watkins, J., & Soumerai, S. B. (2000). Coverage by the news media of the benefits and risks of medications. New England Journal of Medicine, 342(22), 1645–1650.

Pellechia, M. G. (1997). Trends in science coverage: A content analysis of three US newspapers. Public Understanding of Science, 6(1), 49–68.

Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

Schäfer, M. S. (2012). Taking stock: A meta-analysis of studies on the media's coverage of science. Public Understanding of Science, 21(6), 650–663.

Van Trigt, A. M., De Jong-Van Den Berg, L. T. W., Haaijer-Ruskamp, F. M., Willems, J., & Tromp, T. F. J. (1994). Journalists and their sources of ideas and information on medicines. Social Science and Medicine, 38(4), 637–643.

Weigold, M. F. (2001). Communicating science: A review of the literature. Science Communication, 23(2), 164–193.

Weitkamp, E. (2003). British newspapers privilege health and medicine topics over other science news. Public Relations Review 29(3), 321–33.