Measuring Correlation-to-Causation Exaggeration in Press Releases

Bei Yu[1], Jun Wang[2], Lu Guo[3], and Yingya Li[1]

[1]School of Information Studies, Syracuse University
[2]Independent Researcher, Syracuse, NY
[3]School of Public Administration, Sichuan University
{byu,yli48}@syr.edu, {junwang4,luguo2019}@gmail.com

Abstract

Press releases have an increasingly strong influence on media coverage of health research; however, they have been found to contain seriously exaggerated claims that can misinform the public and undermine public trust in science. In this study we propose an NLP approach to identify exaggerated causal claims made in health press releases that report on observational studies, which are designed to establish correlational findings, but are often exaggerated as causal. We developed a new corpus and trained models that can identify causal claims in the main statements in a press release. By comparing the claims made in a press release with the corresponding claims in the original research paper, we found that 22% of press releases made exaggerated causal claims from correlational findings in observational studies. Furthermore, universities exaggerated more often than journal publishers by a ratio of 1.5 to 1. Encouragingly, the exaggeration rate has slightly decreased over the past 10 years, despite the increase of the total number of press releases. More research is needed to understand the cause of the decreasing pattern.

1 Introduction

Recent studies have found that press releases issued by research institutions are a major source of exaggeration in science communication, which is later spread to mainstream media (Woloshin and Schwartz, 2002; Brechman et al., 2009; Sumner et al., 2014; Sumner et al., 2016). Exaggeration in press releases threatens to undermine public trust in science, which is a foundation for the scientific research enterprise.

The influence of press releases on science communication has increased since the 1980s. With the growing competition for reputation and funding, research institutions are increasingly using press releases as a public relations tool for research promotion (Brechman et al., 2009; Sumner et al., 2014). At the same time, independent journalism faces financial challenges and staff shortage (Galewitz, 2006; Schwitzer, 2008). Newspapers, especially small newspapers, rely heavily on press release material to write science news stories (De Semir et al., 1998; Schwitzer, 2008; Woloshin et al., 2009; Taylor et al., 2015). Prior research also shows that high quality press releases seem to improve the quality of associated news stories (Schwartz et al., 2012). Since press releases are now the dominant link between academia and news media, the information quality of press releases plays a critical role in communicating science research to the public.

The problem of exaggeration in press releases is particularly severe in health research. For example, a manual analysis of a sample of UK press releases in health research found that over a third of them contained exaggerated advice, causal claims from correlational findings, or inference of animal studies to humans (Sumner et al., 2014). The errors and inaccuracies in reporting health research findings may also misinform the public about their conditions, diagnosis, and treatments. On the other hand, medicine and health are the dominant topic in science news articles. research outside the health domain was much less covered in science news (Suleski and Ibaraki, 2010). Therefore, studying exaggeration in the health domain is particularly important for both science communication and public health.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4860–4872
Barcelona, Spain (Online), December 8-13, 2020

Researchers and media watchdogs such as HealthNewsReview have been conducting manual content analysis to estimate and monitor press release quality (Smith et al., 2005; Cassels and Lexchin, 2008; Schwitzer, 2008; Sumner et al., 2014). Due to the large number of press releases and scarce funding, this labor-intensive approach is difficult to maintain. In fact, HealthNewsReview stopped operation at the end of 2018 after a nearly 13-year run. Manual content analysis is also inadequate for answering important research questions that require large-scale, real-time analyses. For example, has the exaggeration problem worsened or improved over the years? Do some institutions exaggerate more than others? Seeking answers to these research questions requires developing a computational approach to automatically analyze exaggeration in press releases.

In this study we propose an NLP approach for automatically detecting exaggerated causal claims, one of the most common types of exaggeration, by comparing claims in press releases to the corresponding claims in the original research papers. An exaggerated causal claim is defined as a *causal* statement in a press release with a *correlational* counterpart in the corresponding research paper.

Our study consists of several subtasks. First, we developed a corpus with paper-press pairs. Press releases were downloaded from EurekAlert![1], a nonprofit news-release distribution platform operated by the American Association for the Advancement of Science as a resource for journalists and the public. Doi links were obtained from EurekAlert! and ScienceDaily to associate the press releases with the original research papers in PubMed. The corpus includes 64,177 press releases with references to 62,317 PubMed-indexed journal papers. We then trained a classification model to identify research papers on observational studies; they are designed to establish correlational findings, but are often exaggerated as causal (Woloshin et al., 2009; Sumner et al., 2014; Zweig and DeVoto, 2018). Focusing on research papers with structured abstracts, in which the sentences in the conclusion subsection were used as main statements in research papers, we identified 15,962 observational studies. The headline and first two sentences of a press release are considered as its main statements, according to the "inverted pyramid" structure commonly used by news stories. We then paired the conclusion subsections in the research papers with the main statements in the corresponding press releases for exaggeration analysis. Some research papers were associated with multiple press releases, resulting in 16,431 paper-press pairs dated from 2008 to October 2020 in the final dataset.

Second, we developed BERT-based sentence classifiers to categorize the main statements in press releases and research papers respectively. The sentences were classified by their strength as either "direct causal", "conditional causal", "correlational", or "not claim".

Third, by applying the sentence classifiers to the 16,431 paper-press pairs, we were then able to identify the press releases that used direct causal claims to describe correlational findings in observational studies. Then we used the identified exaggeration cases to answer the following questions: (1) What is the trend of correlation-to-causation exaggeration from 2008 to 2020? (2) Did university press releases make more exaggerated claims than journal press releases?

2 Related Work

Institutional press releases are an important presentation of science research to the public (Autzen, 2014). A press release serves the dual purpose of science communication and public relations (Carver, 2014). Now that research and innovation have become essential drivers of economic growth and international competitiveness, research institutions are facing growing competition for research funding and talents, and are increasingly emphasizing the commercialization of university research. Such pressure results in tension between the goal of responsible science reporting and the goal of marketing when issuing press releases (Caulfield and Ogbogu, 2015; Samuel et al., 2017). A survey study on UK science press officers found that they were aware of such tension, and were trying to balance and achieve both sensation and accuracy in a delicate way (Samuel et al., 2017).

In addition, prior studies have found an increasing influence of press release on the content of science news since the 1980s (Bauer and Bucchi, 2007; Göpfert, 2007). Actually, press releases have become the dominant link between academia and news media (Sumner et al., 2014; Brechman et al., 2009). A

[1] https://www.eurekalert.org/

number of studies have found that journalists rely heavily on press release materials to write science news stories, instead of acting as watchdogs by conducting independent investigations (De Semir et al., 1998; Schwitzer, 2008; Lewis et al., 2008; Woloshin et al., 2009; Taylor et al., 2015). For example, over one third of U.S health news stories were largely or solely based on press releases (Woloshin et al., 2009). Under-staffed smaller newspapers even used press releases verbatim (Ashwell, 2016). This trend is likely to continue as science news reporting positions are cut and journalists are rushed for time (Schwitzer, 2008; Galewitz, 2006).

Journalists have long been criticized for pursuing sensational stories that overstate research results (Fahnestock, 1998). However, recent studies have found that press releases may contain more exaggerations than news stories. According to (Sumner et al., 2014), 58%-86% of news stories derived from the exaggerated press releases contained similar exaggerations, compared with the exaggeration rate of 10%-18% in news articles when the press releases were not exaggerated. In the context of the high institutional expectation for generating "hype" and the shrinking of journalism as the watchdog, it has been argued that press releases have become a major source of exaggeration in science communication, e.g. (Woloshin and Schwartz, 2002; Brechman et al., 2009; Sumner et al., 2014; Sumner et al., 2016). Furthermore, university press releases were found to contain more exaggerations than journal press releases, probably because the university press officers face more pressure to generate expectations and hype (Sumner et al., 2016).

Prior studies on exaggeration in press releases have been limited to manual content analysis, which have provided valuable insights for understanding the types of exaggeration and estimating the problem severity at certain times. For example, (Sumner et al., 2014) is the first study on estimating the exaggeration rate in press releases by manually checking a sample of press releases from 20 UK univerities in 2011. However, the manual content analysis method falls short for answering important research questions that require large-scale, real-time analyses, such as longitudinal trend of exaggeration rate over time. This creates a need to develop a computational approach that can consistently measure exaggeration in a large number of press releases, and further study its trend and difference by factors such as the source of press releases (e.g. academic institutions vs. journal publishers).

Although computational modeling of exaggeration in press releases is a new research problem, it is closely related to the field of NLP applications for information credibility assessment, especially in the health domain, given the strong need for curbing misinformation in recent years. The task of credibility assessment has often been modeled as a text categorization problem. A common way to categorize credibility is true/fake binary prediction, e.g. (Abbasi et al., 2012; Dhoju et al., 2019). However, many online documents are not completely true or fake. Instead, multiple claims can be made in one document, and claims can be inaccurate in different ways. Therefore, a more precise way to assess credibility is faceted evaluation guided by established credibility frameworks with multiple quality criteria.

Some frameworks were designed for rating online health information, such as the HON principles and DISCERN. For example, the DISCERN instrument includes 16 questions, such as "are the aims clear?" and "is it balanced and unbiased?" (Charnock et al., 1999). In comparison, HealthNewsReview.org designed 10 criteria specifically for news stories. These criteria include exaggeration, missing information, etc. Expert ratings based on these criteria have been used to train NLP models. For example, an automated system was developed to assess health websites' conformity to the HON principles (Boyer and Dolamic, 2015). AutoDISCERN is an NLP tool designed to automatically assign DISCERN quality ratings to webpages using hierarchical encoder attention-based neural network (Kinkead et al., 2020). NLP models were also developed to automate ratings based on the ten HealthNewsReview criteria (Dai et al., 2020). These methods rely on human expert ratings to train credibility prediction models. In comparison, our method links press releases to the original research papers for fact checking, and we focus on exaggeration, a specific criterion used by HealthNewsReview, due to its significant impact on the quality of press releases (Sumner et al., 2014).

Our work is also built upon prior studies on identifying causal relations, e.g. (Li and Mao, 2019). A causal relation can be broadly defined as any type of cause-effect relation between events, for example Event A is the consequence of Event B (Bethard and Martin, 2008; Mirza and Tonelli, 2014). It is

commonly realized by causal indicators such as "because", "as a result of", "as a consequence of" (Prasad et al., 2008). This broad definition is also adopted in the NLP tasks SemEval-2007 (Girju et al., 2007) and SemEval-2010 (Hendrickx et al., 2009). However, a causal relation presented in a research finding sentence is in a more specific linguistic format, i.e. a relation between an independent variable and a dependent variable (Cofield et al., 2010), such as "smoking causes cancer". To identify causal claims in research findings, Yu et al. (2019) developed a corpus of claims with different strength levels and trained a BERT-based model to distinguish them. Compared to research papers, press releases are a different genre. Li et al. (2017) found that genre differences may result in more challenge for identifying causal claims in news articles. Based on these prior studies, we decided to construct a new corpus for identifying causal claims in press releases, which are more similar to news stories, and to develop prediction models based on BERT.

3 Corpus Construction

3.1 Collecting press releases

The press releases used in this study were downloaded from EurekAlert!. Since 1996 EurekAlert! has been the major platform for more than 3,000 scientific journals, research institutes, and government agencies to distribute press releases to journalists and the public. To date, EurekAlert! includes about 137,000 health and medicine research-related press releases[2]. From 2015 EurekAlert! started to add a section named *related journal article* onto their webpage that has a doi link to the original research paper. Currently, about 34,000 health and medicine-related press releases contain links to the original papers. However, the press releases with doi links account for less than 25% of research-related ones in EurekAlert!.

To link more press releases to the associated journal articles, especially those published before 2015, We used ScienceDaily as an additional data source. ScienceDaily is an independent website that reposts press releases. The editors manually added doi links dated back to 2008. However, we can not adopt the press releases directly from ScienceDaily because its editors sometimes modify the headline and the beginning part of a press release. Given that we need to use the original press releases posted by press officers rather than the edited ones, we borrowed the doi links from ScienceDaily to help fill the missing links between the EurekAlert! press releases and the associated papers. The basic idea is: for each EurekAlert! press release, find a best match from ScienceDaily; if the best match has a doi link, pair the EurekAlert! item with the doi link.

To implement this idea, we created a simple search engine using Elasticsearch[3], a popular search engine that is based on the open-source Apache Lucene library. Elasticsearch supports a type of decay function that can decay relevance score depending on how far a value is from a given origin. We used this function to match publication dates between EurekAlert! and ScienceDaily press releases, which should be as close as possible. For each press release from EurekAlert! and ScienceDaily, we extracted its metadata such as publication date and headline, and computed 150 TF-IDF weighted keywords from its full text. A query was formed based on the metadata and keywords for each EurekAlert! item. Given the query, the Elasticsearch returned a ScienceDaily item with the highest relevance score. If the score is higher than a threshold, a match between EurekAlert! and ScienceDaily is found. We set the threshold to a value such that Elasticsearch achieved an accuracy of 99.5% and a recall of 90% when evaluated on the 34,000 EurekAlert! items that already have associated doi links. Using this method we found doi links for an additional 30,000 EurekAlert! press releases.

In the end, our final corpus consists of 64,177 press releases, spanning from 2008 to October 2020. They are associated with 62,317 journal publications —some collaborative work were reported by two or more press releases from different research institutions— for which we downloaded their titles and abstracts from PubMed.

[2]Currently, EurekAlert! has accumulated 444,000 press releases, in which 181,000 are labeled with Medicine/Health. Through analyzing the "type" metadata in each EurekAlert! press release page, we found that about 24% of the 181,000 press releases are not labeled as research; rather, they are labeled as grant, business, award, meeting, book, and etc.

[3]https://www.elastic.co/

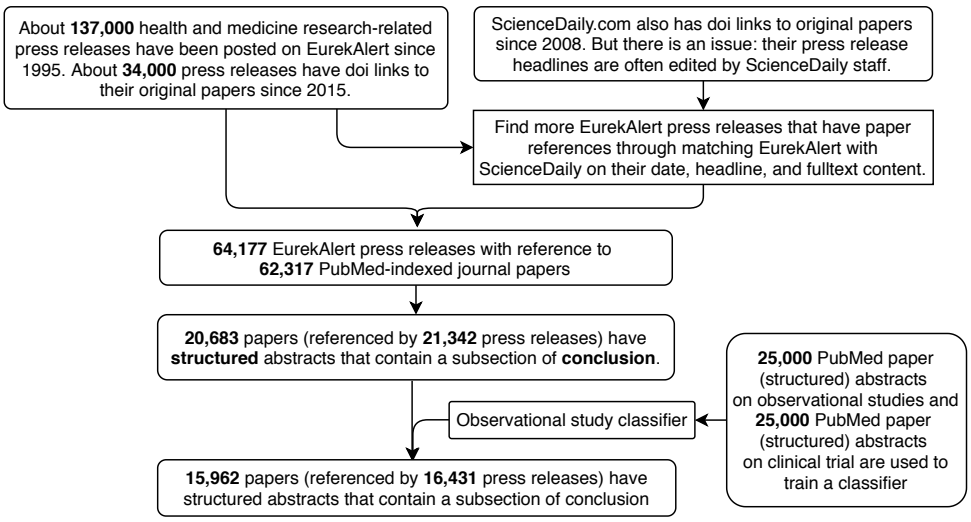


Figure 1: Collecting press releases and associated journal papers that report on observational studies.

3.2 Locating claim statements

For a press release, we assume that its main claim statements are presented in the headline or the first two sentences, here referred to as the "three-sentence opening section". According to Sumner et al. (2014), these sentences are usually the main statements, since nearly all press releases and news stories follow the "inverted pyramid" structure of stating their main claims first (Pöttker, 2003).

To verify the above assumption, we randomly sampled 100 press releases, and checked how many press releases violated the assumption. The result shows only two press releases put the main statements outside the three-sentence opening section. Hence, the assumption holds well for our data.

For an academic paper, its main claim statements or findings are usually presented in its *conclusion* sections. However, the conclusion sections often include not only the main findings but also statements on limitations and future work. In comparison, the abstracts of biomedical and health research papers often have a structure that usually includes subsections of *introduction, method, result*, and *conclusion*. For simplicity, we focused on those papers with a structured abstract, and extracted the conclusion subsections as the main statements. This reduces our data to 20,683 research papers and 21,342 associated press releases.

3.3 Observational study classification

Our next subtask is to identify the paper-press pairs that report on observational studies. Since PubMed has made a significant effort to identify the study designs in biomedical and health literature, we developed our identification method by expanding on PubMed's existing data.

In 2014, the PubMed staff introduced the tag *observational study* to their list of *publication types*, a metadata item for annotating study designs, such as observational studies and clinical trials [4]. To date they have manually annotated more than 60,000 research papers as observational studies. PubMed also established a category of *clinical trial* in 2008, and has since annotated over 861,000 papers as clinical trials.

We sampled 50,000 PubMed-annotated articles as training data, including 25,000 observational studies and 25,000 clinical trials. Since the fulltext of a article is often not available, our data are in the format of abstracts; specifically, our training data consist only of those articles with structured abstract because in this study, as we explained above, the claim statements of a research paper are identified from the conclusion subsection in the structured abstract.

LightGBM[5], a decision tree-based gradient boosting algorithm, is chosen to train the observational study classifier based on paper titles and abstracts. Using 80% of them for training, and 20% for testing,

[4]https://www.ncbi.nlm.nih.gov/mesh/68064888
[5]https://github.com/microsoft/LightGBM

Type	Description	Language Cue	Example Sentence
Correlational	The statement describes the association between variables, but causation cannot be explicitly stated.	association, associated with, predictor, at high risk of	Suicide risk greater for people living at higher elevations.
Conditional causal	The statement shows that one variable directly changes the other. However, the relation carries an element of doubt in it, which is normally via hedges or modalities.	increase, decrease, lead to, effect on, contribute to, result in (*Cues indicating doubt:* may, might, appear to, probably)	Being overweight or obese during middle age may increase the risk of certain dementias.
Direct causal	The statement says that the independent variable directly alters the dependent variable.	increase, decrease, lead to, effective in, contribute to, reduce, can	Traffic noise increases the risk of having a stroke.
Not claim	The statement is not for current study findings or no correlation/causation relationship is mentioned in the statement.		Many rheumatoid arthritis patients not getting recommended drugs.

Table 1: A taxonomy of claim types and examples of commonly used language cues.

Label	Count	Ratio
Correlational	737	35.5%
Conditional causal	284	13.7%
Direct causal	568	27.4%
Not claim	487	23.4%
Total	2076	100%

Table 2: Sentence type distribution.

we trained a LightGBM model with an F1 score of 0.95. Applying this classifier to the 20,683 research papers, we identified a total of 15,962 observational studies, associated with 16,431 press releases. The corpus construction procedure described above is also illustrated in Figure 1.

3.4 Annotating a training corpus

Following the coding schema in (Yu et al., 2019), which itself is a simplified version of the taxonomies defined in previous studies (Kilicoglu et al., 2015; Sumner et al., 2014), we annotated the following four claim types: *correlational*, *conditional causal*, *direct causal*, and *not claim*. Table 1 lists the category definitions and some common language cues used to identify the relation type for each category. Example sentences of different claim types are also shown in the table.

We randomly sampled 700 press releases from our corpus for manual annotation. They consist of 2,100 sentences in the three-sentence opening sections. Occasionally, a statement could have more than one claim type. These mixed-type statements were excluded, resulting in 2,076 sentences in the final training data set. Table 2 shows the claim type distribution of single-type sentences in the training corpus.

A sample of 200 sentences were randomly selected for conducting the inter-coder reliability test. Two annotators labeled the claim type for each sentence. The overall Cohen's Kappa agreement (Cohen, 1960) was 0.95, indicating a near-perfect inter-coder agreement (McHugh, 2012). The two annotators then annotated the rest of the data. All disagreements during the annotation were later resolved by the team through discussion.

4 Sentence-level Claim Type Classification

Due to the significant difference in writing styles between academic papers and press releases, we chose to train separate prediction models, one for each genre. For academic papers, we adopted the classification model developed by Yu et al. (2019) that has achieved a performance of 0.88 macro-F1 and 0.90 accuracy. For press releases, we trained a new prediction model using the press release corpus that we constructed as described above.

4.1 Prediction model

Three models were trained for main statement sentence classification: bag-of-words based Linear SVM, BERT (Devlin et al., 2018), and BioBERT (Lee et al., 2019). We hypothesized that BioBERT (Lee et al., 2019) would perform the best, since it is a domain specific language representation model that is based on BERT and further trained on large-scale biomedical corpora. BioBERT has shown to outperform BERT on three representative biomedical text mining tasks, including text classification (Lee et al., 2019).

All models were evaluated on our annotated corpus via 5-fold stratified cross-validation. As shown in Table 3(a), BioBERT achieved a macro-averaged F1 score of 0.868, outperforming BERT and LinearSVM, whose scores are 0.844 and 0.732 respectively.

To further improve performance, we augmented our press release data with the academic paper data (Yu et al., 2019)[6]. In our implementation, we set the weight of the data in the academic genre to half of the weight in the press release genre. As shown in Table 3(b), the data augmentation approach can boost performance from 0.868 to 0.892.

	Precision	**Recall**	**F1**
Correlational	0.908	0.900	0.904
Conditional causal	0.827	0.958	0.887
Direct causal	0.885	0.868	0.876
Not claim	0.826	0.780	0.802

(a) macro-F1: 0.868

	Precision	**Recall**	**F1**
Correlational	0.924	0.883	0.903
Conditional causal	0.919	0.954	0.936
Direct causal	0.885	0.921	0.903
Not claim	0.829	0.825	0.827

(b) macro-F1: 0.892 (with data augmentation)

Table 3: Performance of BioBERT on each type of press release sentences.

4.2 Error Analysis

Although the prediction model achieved strong performance, there were challenges in distinguishing the "not claim" sentences from others. An examination of the misclassified sentences shows that many such sentences actually talk about research background instead of research findings. We defined these sentences as belonging to the "not claim" category. However, sometimes these sentences used causal language, which may confuse the prediction model. For example, *"choking is a leading cause of injury among children"*.

There was also some confusion between the "direct causal" and the "correlational" categories. This type of error was mainly due to the causal markers in the subordinate clauses instead of the main clauses. For example, the sentence *"many children are at increased risk for sleep breathing disorders that can impair their mental and physical development"* describes a correlational claim but was misclassified as "direct causal" due to the cue "can impair". In our taxonomy, "can + causal cue" belongs to the category of "direct causal".

The prediction model also had some difficulty in recognizing less commonly used relation markers, especially when nouns were used to describe correlational findings in headlines, such as the phrase "coupled with" in the headline *"Kidney disease coupled with heart disease common problem in elderly"*.

5 Exaggeration in Press Releases

5.1 Strategy for identifying exaggeration

Our prediction model operates at sentence level, so each press release will receive three predictions for its three-sentence opening section, and each paper will receive one or more predictions depending on the number of sentences in the conclusion subsection. However, exaggeration is determined at article level, i.e. whether a press release made exaggeration or not. Hence, a strategy is needed to determine article-level exaggeration by aggregating sentence level predictions. Since the study goal is to detect

[6] https://github.com/junwang4/causal-language-use-in-science

the correlation-to-causation type of exaggeration, we focused on the research papers that made correlational claims only. A research paper's finding is regarded as "correlational" if at least one sentence in the conclusion subsection uses correlational language and no sentence uses direct or conditional causal language. In addition, we only considered the direct causal claims from correlational findings as exaggerations, since sometimes researchers would also speculate a conditional causal claim from a correlational finding, which often serves as the goal of future investigation.

We then designed a strategy for identifying exaggeration at article level by considering the characteristics of press release as a news genre. First, headlines serve the role of attracting readers to the story; therefore exaggeration is more likely to occur in headlines, and can be more impactful than those in content. Hence, predictions of causal claims in headlines should carry more weight than those from first and second sentences. Second, most, but not all headlines contain main claims. In our annotated corpus, 15.6% headlines were annotated as "not claim". An examination of these cases showed that the main claims were significantly shortened or paraphrased. An extremely short headline is "Brain Power". In these cases, the first and second sentences are more likely to contain the main claims. Third, a small number of first and second sentences used causal language to describe research background instead of findings. Error analysis in the previous section has shown that our prediction model has difficulty in distinguishing them from real causal findings. Therefore, when one sentence is predicted as "causal" while the other as "correlational", it may or may not be a case of exaggeration.

Considering the above concerns, we designed a conservative strategy for identifying exaggeration in all paper-press pair with the paper's finding predicted as "correlational", as illustrated in Figure 2.

Input: Three sentence predictions on headline, 1st sentence, and 2nd sentence of a press release
Output: Did the press release exaggerate correlational findings as causal? (Yes or No)
Process: If headline prediction is direct causal, exaggeration=Yes
Else if headline prediction is "not claim", check predictions for 1st and 2nd sentences:
If both predictions are "direct causal", exaggeration=Yes
Else if one prediction is "direct causal" and the other "not claim", exaggeration=Yes
Else exaggeration=No
Else exaggeration=No

Figure 2: Claim aggregation algorithm for press releases that correspond to papers identified as correlational.

5.2 Overall exaggeration in press releases

After applying the exaggeration identification strategy to our corpus of 16,431 paper-press pairs, out of 6,244 cases in which the original research papers drew correlational conclusions, we found 1,391 (22%) cases of exaggeration. This exaggeration rate is in a similar range to previous studies; however, our exaggeration identification strategies differ.

Sumner et al. (2014) reported an exaggeration rate of 33% by manual analysis of 462 press releases issued by 20 leading UK universities in 2011. They used seven scales to measure the claim strength: no statement, explicit statement of no relation, correlational, ambiguous (e.g. 'linked to"), conditional causal (e.g. "may" and "might"), "can" causal, and unconditionally causal. An exaggeration was reported when the press release used stronger claims than those present in the associated research paper.

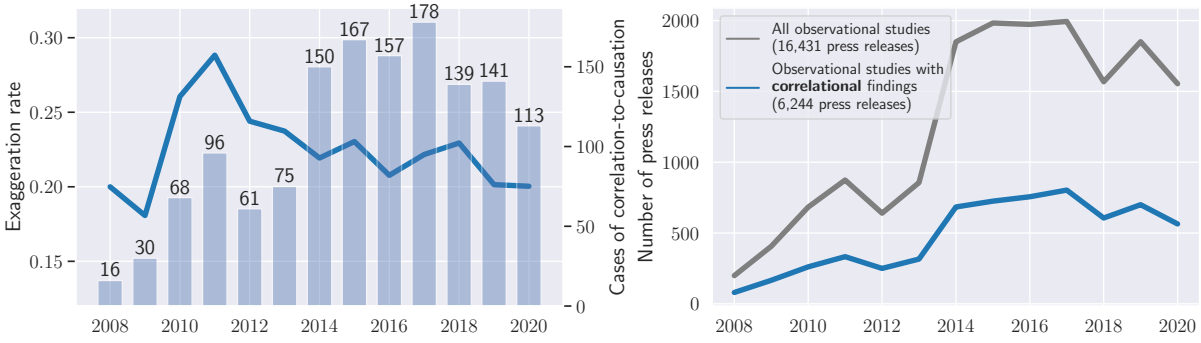
Later, the same research team re-analyzed the data using a five-scale measure by merging the correlational, ambiguous, and conditional causal categories, in order to focus on the types of exaggeration that have a more significant impact on readers' understanding. The new analysis resulted in a lower rate of 19%, since the nuanced exaggeration cases were no longer counted (Adams et al., 2017).

Compared to (Sumner et al., 2014) and (Adams et al., 2017), our schema includes four categories: the "direct causal" category equates to the combined categories of "can" causal and unconditional causal defined in (Sumner et al., 2014); the "conditional causal" category is equivalent to the category with the same name in (Sumner et al., 2014); the "correlational" category corresponds to the combined categories of "correlational" and "ambiguous" in (Sumner et al., 2014); the "not claim" is similar to the "no

statement" category in (Sumner et al., 2014). We did not use the "statement of no relation" category, which refers to the statements of negative findings, because our annotations are based on correlational vs. causal language use. A negative finding on correlation would still use correlational language, and thus be annotated as correlation. Similarly, a negative finding on causal relation are annotated as causal.

In comparison, our exaggeration identification strategy is even more conservative than (Adams et al., 2017), in that we only considered the most significant cases of exaggeration, in which the research papers' conclusions used correlational language only, whereas the associated press releases used direct causal claims. We did not count minor exaggeration cases, such as those with conditional causal claims in research papers and direct causal in press releases.

5.3 Trend of exaggeration in press releases



(a) (b)

Figure 3: (a) Trend of exaggeration: there is an overall decreasing trend of exaggeration since 2010. (b) Number of press releases that are used as the basis for calculating the exaggeration rate in (a).

Figure 3(a) plots the trend of exaggeration rate over the past decade. This result shows a slight but significant decreasing trend since 2010, with a Spearman rank correlation coefficient of -0.88 (p-value$<$0.001). Note the years 2008 and 2009 were excluded from this analysis due to insufficient data, as shown in Figure 3(b).

Figure 3(b) shows the total number of observational studies each year, and among them the total number that used correlational language only. These numbers provide more data context for better understanding the trend. The sharp increase of observational studies in 2014 is caused by a big jump in the total number of health-related press releases in EurekAlert!. The dip in 2018 is a result of ScienceDaily starting to block robot access that year. However, ScienceDaily is less likely to be needed in the future since EurekAlert! has been adding more doi links since 2015. For 2020, we collected only 10 months of data to date.

More research is needed to understand the cause of the decrease in exaggeration rate. Bratton et al. (2020) hypothesized that warnings within the science community might be impactful, after observing a drastic reduction of exaggeration rate (by half) in the UK's press releases six months after the publication of (Sumner et al., 2014). However, confounding factors cannot be ruled out without further evidence. Our result did not replicate the pattern of sharp decrease in 2014 or 2015. Instead, it shows that the exaggeration rate was mildly decreased globally.

5.4 Comparing academic institutions and journal publishers

To investigate whether universities exaggerated more than journal publishers, Sumner et al. (2016) compared press releases from these two sources, and found a higher exaggeration rate in university press releases, exceeding the journals' by 33% to 21%, or a ratio of 1.6 to 1. Adams et al. (2017) re-analyzed the data in (Sumner et al., 2016) based on a simplified, 5-scale claim strength schema, and reported a similar ratio (1.5 to 1, or 19% to 13%).

Here we examine whether their findings still hold for our corpus. To identify the source type of a press release, we used two EurekAlert! metadata items: the media contact person's email address and the name of the source institution. Specifically, if the domain name contains `.edu` or `.ac`, or the institution name contains the word "University", the source is categorized as a university. If the email address matches with a list of well-known publisher domain names such as `bmj.com` and `plos.org`, or the institution name contains the word "Publisher" or "Press", the source is categorized as a journal publisher. With this method, out of 6,244 research papers with correlational findings, we found 3,021 press releases from universities and 1,534 from journal publishers.

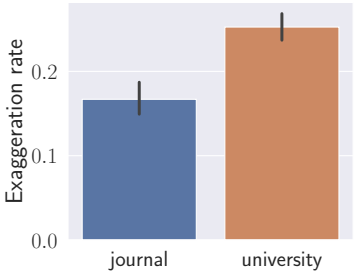


Figure 4: University press releases have higher exaggeration rate than journal's.

Figure 4 shows that the exaggeration rate of university press releases is 25.3%, higher than the 16.7% of journal publishers, by a ratio of 1.5 to 1. This result is consistent with the ratios reported in (Sumner et al., 2016) and (Adams et al., 2017).

6 Conclusion and Future Work

In this study we examined patterns of correlation-to-causation exaggeration in press releases. We developed a new corpus with 16,431 pairs of observational study papers and associated press releases, and a prediction model for identifying cases of correlation-to-causation exaggeration in press releases. Applying the prediction model to the new corpus, we found that among all observational study papers that made correlational claims only, 22% of the associated press releases made exaggerated causal claims. Furthermore, universities made more exaggerated claims than journal publishers by a 1.5 to 1 ratio. The good news is the exaggeration rate has slightly decreased in the past ten years, despite the increase in the total number of press releases. More research is needed to understand the cause of the decreasing pattern.

This study has a few limitations which we plan to improve in future work: First, the current exaggeration detection method relies on the structured abstract of a research paper to identify its main claims from the conclusion subsection. This restriction limits our data set size since only about one third of papers have a structured abstract. In future work we will design an algorithm that can automatically identify the conclusion sentences from unstructured abstracts. Second, the current approach uses the doi links provided by EurekAlert! and ScienceDaily to link research papers and press releases; however, these manually-added links cover less than 50% of EurekAlert! press releases. In future work we would like to develop a content-based matching system to link press releases to scientific literature. Third, the current approach matches the main claims in research papers and press releases at article level, by utilizing the inverted-pyramid structure in press release. This approach works well for press releases, but might not be suitable for identifying exaggeration in other genres. For example, inverted-pyramid structure is not common in online discussion forums (e.g. Reddit (Zhou and Yu, 2020)) and social media posts. Some news articles may also cite findings from multiple scientific studies. In future work we would like to develop a sentence-level approach for pairing claims in research papers and their paraphrased versions in news and social media.

Code and Data

Our code and data are available on github:
`https://github.com/junwang4/correlation-to-causation-exaggeration`.

Acknowledgements

This research is supported by the US National Science Foundation under grant 1952353 and the Microsoft Investigator Fellowship program. We thank the reviewers for helpful comments, and Albert Wang for help with editing the manuscript.

References

Ahmed Abbasi, Fatemeh Mariam Zahedi, and Siddharth Kaza. 2012. Detecting fake medical web sites using recursive trust labeling. *ACM Transactions on Information Systems (TOIS)*, 30(4):1–36.

Rachel C Adams, Petroc Sumner, Solveiga Vivian-Griffiths, Amy Barrington, Andrew Williams, Jacky Boivin, Christopher D Chambers, and Lewis Bott. 2017. How readers understand causal and correlational expressions used in news headlines. *Journal of Experimental Psychology. Applied*, 23(1):1–14.

Douglas James Ashwell. 2016. The challenges of science journalism: The perspectives of scientists, science communication advisors and journalists from new zealand. *Public Understanding of Science*, 25(3):379–393.

Charlotte Autzen. 2014. Press releasesthe new trend in science communication. *Journal of Science Communication*, 13(3):C02.

Martin W Bauer and Massimiano Bucchi. 2007. *Journalism, science and society: Science communication between news and public relations*. Routledge.

Steven Bethard and James H Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 177–180. Association for Computational Linguistics.

Célia Boyer and Ljiljana Dolamic. 2015. Automated detection of honcode website conformity compared to manual detection: an evaluation. *Journal of Medical Internet Research*, 17(6):e135.

Luke Bratton, Rachel C Adams, Aimée Challenger, Jacky Boivin, Lewis Bott, Christopher D Chambers, and Petroc Sumner. 2020. Causal overstatements reduced in press releases following academic study of health news. *Wellcome Open Research*, 5.

Jean Brechman, Chul-joo Lee, and Joseph N Cappella. 2009. Lost in translation? a comparison of cancer-genetics reporting in the press release and its subsequent coverage in the press. *Science Communication*, 30(4):453–474.

Rebecca B Carver. 2014. Public communication from research institutes: Is it science communication or public relations? *Journal of Science Communication*, 13(3):C01.

Alan Cassels and Joel Lexchin. 2008. How well do canadian media outlets convey medical treatment information?: Initial findings from a year and a half of media monitoring by media doctor canada. *Open Medicine*, 2(2):e45.

Timothy Caulfield and Ubaka Ogbogu. 2015. The commercialization of university-based research: Balancing risks and benefits. *BMC Medical Ethics*, 16(1):70.

Deborah Charnock, Sasha Shepperd, Gill Needham, and Robert Gann. 1999. Discern: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology & Community Health*, 53(2):105–111.

Stacey S Cofield, Rachel V Corona, and David B Allison. 2010. Use of causal language in observational studies of obesity and nutrition. *Obesity facts*, 3(6):353–356.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 853–862.

Vladimir De Semir, Cristina Ribas, and Gemma Revuelta. 1998. Press releases of science journal articles and subsequent newspaper stories on the same topic. *JAMA*, 280(3):294–295.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sameer Dhoju, Md Main Uddin Rony, Muhammad Ashad Kabir, and Naeemul Hassan. 2019. Differences in health news from reliable and unreliable media. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 981–987.

Jeanne Fahnestock. 1998. Accommodating science: The rhetorical life of scientific facts. *Written communication*, 15(3):330–350.

Phil Galewitz. 2006. Ongoing newsroom cutbacks hit health reporting ranks. *Association of Health Care Journalists. HealthBeat*, 9:1–12.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18. Association for Computational Linguistics.

Winfried Göpfert. 2007. The strength of PR and the weakness of science journalism. In MW Bauer and M Bucchi, editors, *Journalism, Science and Society: Science Communication Between News and Public Relations*, chapter 20, pages 215–226. Routledge, New York.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics.

Halil Kilicoglu, Graciela Rosemblat, Michael Cairelli, and Thomas Rindflesch. 2015. A compositional interpretation of biomedical event factuality. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 22–31.

Laura Kinkead, Ahmed Allam, and Michael Krauthammer. 2020. Autodiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks. *BMC Medical Informatics and Decision Making*, 20.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Justin Lewis, Andrew Williams, and Bob Franklin. 2008. A compromised fourth estate? UK news journalism, public relations and news sources. *Journalism Studies*, 9(1):1–20.

Pengfei Li and Kezhi Mao. 2019. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, 115:512–523.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP Analysis of Exaggerated Claims in Science News. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism*, pages 106–111.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106.

Horst Pöttker. 2003. News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Gabrielle Samuel, Clare Williams, and John Gardner. 2017. UK science press officers, professional vision and the generation of expectations. *Public Understanding of Science*, 26(1):55–69.

Lisa M Schwartz, Steven Woloshin, Alice Andrews, and Therese A Stukel. 2012. Influence of medical journal press releases on the quality of associated newspaper coverage: retrospective cohort study. *BMJ*, 344:d8164.

Gary Schwitzer. 2008. How do US journalists cover treatments, tests, products, and procedures? An evaluation of 500 stories. *PLOS Medicine*, 5(5):e95.

David E Smith, Amanda J Wilson, and David A Henry. 2005. Monitoring the quality of medical news reporting: early experience with media doctor. *Medical Journal of Australia*, 183(4):190–193.

Julie Suleski and Motomu Ibaraki. 2010. Scientists are talking, but mostly to each other: a quantitative analysis of research represented in mass media. *Public Understanding of Science*, 19(1):115–125.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy Boy, and Christopher D Chambers. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, 349:g7015.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andrew Williams, Lewis Bott, Rachel Adams, Christos A Venetis, Leanne Whelan, Bethan Hughes, and Christopher D Chambers. 2016. Exaggerations and caveats in press releases and health-related science news. *PLOS ONE*, 11(12):e0168217.

Joseph W Taylor, Marie Long, Elizabeth Ashley, Alex Denning, Beatrice Gout, Kayleigh Hansen, Thomas Huws, Leifa Jennings, Sinead Quinn, Patrick Sarkies, et al. 2015. When medical news comes from press releasesa case study of pancreatic cancer and processed meat. *PLOS ONE*, 10(6):e0127848.

Steven Woloshin and Lisa M Schwartz. 2002. Press releases: Translating research into news. *JAMA*, 287(21):2856–2858.

Steven Woloshin, Lisa M Schwartz, Samuel L Casella, Abigail T Kennedy, and Robin J Larson. 2009. Press releases by academic medical centers: not so academic? *Annals of Internal Medicine*, 150(9):613–618.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4656–4666.

Haichen Zhou and Bei Yu. 2020. Information quality of reddit link posts on health news. In *Proceedings of the 2020 iConference*, pages 186–197. Springer.

Mark Zweig and Emily DeVoto. 2018. Observational studies: Does the language fit the evidence? Association vs. causation. http://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/.