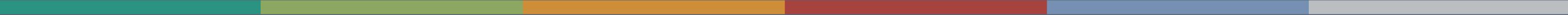




TDM: Unlocking the hidden potential from scholarly content



Making sense of unstructured content

Until recently, text mining has mostly been restricted to post-publication PDFs and has proved slow and difficult. The focus for scholarly content has often been limited to metadata and abstracts.

TDM is evolving to extract a wealth of information that can support the entire scholarly community – from authors to publishers.



scholarcy

Landscape

Too many manuscripts. Not enough time.

Submission to publication time expanding.



48 Hours
Screening



13 Weeks
First review
round



400 Days
Submission to
publication

The format challenge

XML often made available for Open Access articles, but not all publishers make XML available to TDM services (API).

Rise of preprint servers and number of journals inviting article submission via these servers increases need to mine non-XML content.

Most authors still submit manuscripts to publishers & preprint servers in Word or PDF.

Some servers convert content into XML, but majority of platforms only allow for the preprint to be downloaded in the same format it was uploaded in.

Software used by authors

Word still the preferred format

1. Microsoft Word

2. Google docs

3. LaTeX

4. Overleaf

5. GitHub

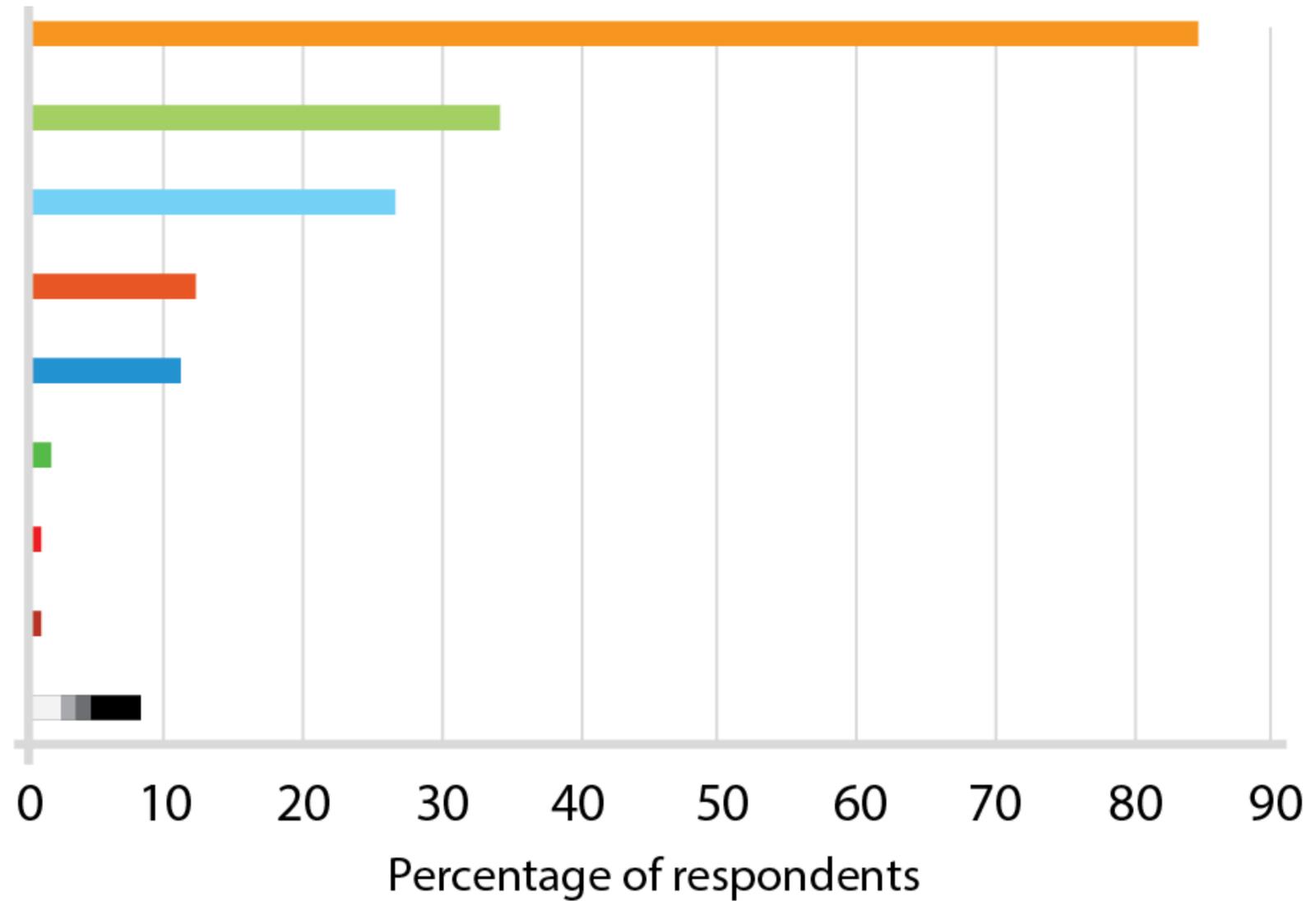
6. Adobe InDesign

7. Authorea

8. Scrivener

9. Other

- Other - LibreOffice
- Other - R / R Studio
- Other - Apple Pages
- Other



n=4079

Writing software used by authors submitting to bioRxiv.

Source: Sever et al (2019) [bioRxiv: the preprint server for biology](https://doi.org/10.1101/833400). <https://dx.doi.org/10.1101/833400>



Format shouldn't matter

Extracting structured content from any document

ARTICLE OPEN How the weather affects the pain of citizen scientists using a smartphone app

William G. Dixon^{1,2,3,4}, Anna L. Beukenhorst⁵, Belay B. Yimer¹, Louise Cook¹, Antonio Gasparini^{6,7}, Tal El-Hay⁸, Bruce Hellman⁹, Ben James⁷, Ana M. Vicedo-Cabrera¹⁰, Malcolm MacLure¹¹, Ricardo Silva^{12,13}, John Alneworth¹⁴, Hual Leng Pisanelli¹⁵, Thomas House^{16,17}, Mark Lunt¹⁸, Carolyn Gamble^{19,20}, Caroline Sanders^{14,15}, David M. Schultz²¹, Jamie C. Sergeant^{22,23,24} and John McBeth^{25,26}

Patients with chronic pain commonly believe their pain is related to the weather. Scientific evidence to support their beliefs is inconclusive, in part due to difficulties in getting a large dataset of patients frequently recording their pain symptoms during a variety of weather conditions. Smartphones allow the opportunity to collect data to overcome these difficulties. Our study Cloudy with a Chance of Pain analysed daily data from 2658 patients collected over a 15-month period. The analysis demonstrated significant yet modest relationships between pain and relative humidity, pressure and wind speed, with correlations remaining even when accounting for mood and physical activity. This research highlights how citizen-science experiments can collect large datasets on real-world populations to address long-standing health questions. These results will act as a starting point for a future system for patients to better manage their health through pain forecasts.

npj Digital Medicine (2019)3:105 | <https://doi.org/10.1038/s41746-019-0180-3>

INTRODUCTION

Weather has been thought to affect symptoms in patients with chronic disease since the time of Hippocrates over 2000 years ago.¹ Around three-quarters of people living with arthritis believe their pain is affected by the weather.^{2,3} Many report their pain is made worse by the cold, rain, and low atmospheric pressure. Others report that their pain is made worse by warmth and high humidity. Despite much research examining the existence and nature of the weather-pain relationship,⁴ there remains no scientific consensus. Studies have failed to reach consensus in part due to their small sample sizes or short durations (commonly fewer than 100 participants or one month or less); by considering a limited range of weather conditions; and heterogeneity in study design (e.g. the populations studied, methods for assessing pain, assumptions to determine the weather exposure, and statistical analysis techniques).^{5,11} Resolving this question requires collection of high-quality symptom and weather data on large numbers of individuals. Such data also need to include other factors potentially linked to daily pain variation and weather, such as mood and amount of physical activity. Collecting this kind of multi-faceted data in large populations over long periods of time, however, has been difficult.

The increasing uptake of smartphones offers new and significant opportunities for health research.¹² Smartphones allow the integration of data collection into daily life using applications (apps). Furthermore, embedded technologies within the smartphones, such as the Global Positioning System (GPS), can be used

to link the data collection to specific locations. We created Cloudy with a Chance of Pain,^{13,14} a national United Kingdom smartphone study, to collect a large dataset to examine the relationship between local weather and daily pain in people living with long-term pain conditions.

RESULTS

Recruitment and retention

The study app was downloaded by 13,267 users over the 12-month recruitment period (Figs 1 and 2a) with recruitment from all 124 UK postcode areas. A total of 10,584 participants had complete baseline information and at least one pain entry, with 6850 (65%) participants remaining in the study beyond their first week and 4692 (44%) beyond their first month (Fig. 2b). Further description of engagement clusters is provided in Supplementary Table 2 and Supplementary Figs 1–3. A total of 2658 participants had at least one hazard period matched to a control period in the same month (Fig. 3) and were included in the final analysis. There were 9695 hazard periods included in the analysis for the final 2658 participants, matched to 81,727 control periods in 6401 participant-months. A total of 1235 participants contributed one month, and the remaining 1423 participants contributed 2–15 months.

The final cohort was active for a median of 165 days (interquartile range, IQR 84–245) with symptoms submitted on an average of 73% of all days. Cohort members were

¹Centre for Systemic Myology, Imperial College, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. ²Health Research Centre, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. ³MBR Greater Manchester Biomedical Research Centre, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. ⁴Department of Public Health, Environment and Society, London School of Hygiene & Tropical Medicine, London, UK. ⁵Centre for Statistical Methodology, London School of Hygiene & Tropical Medicine, London, UK. ⁶IBM Research, Haifa, Israel. ⁷AMRI Limited, London, UK. ⁸Department of Anesthesiology, Pharmacology and Therapeutics, University of British Columbia, Vancouver, BC, Canada. ⁹Department of Statistical Science, UCL, London, UK. ¹⁰Ran Tsing Institute, London, UK. ¹¹Discipline of Medicine, The University of Adelaide, Adelaide, Australia. ¹²School of Mathematics, The University of Manchester, Manchester, UK. ¹³IBM Research, Haifa, Israel. ¹⁴Sci-Tech, Daresbury, UK. ¹⁵Greater Manchester Patient Safety Translational Research Centre, The University of Manchester, Manchester, UK. ¹⁶MBR School of Primary Care, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. ¹⁷Centre for Atmospheric Science, School of Earth and Environmental Science, The University of Manchester, Manchester, UK. ¹⁸Centre for Biostatistics, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. ¹⁹These authors contributed equally: Jamie C. Sergeant, John McBeth. ²⁰email: william.dixon@manchester.ac.uk



How the weather affects the pain of citizen scientists using a smartphone app

William G. Dixon, Anna L. Beukenhorst, Belay B. Yimer, et al.

2019

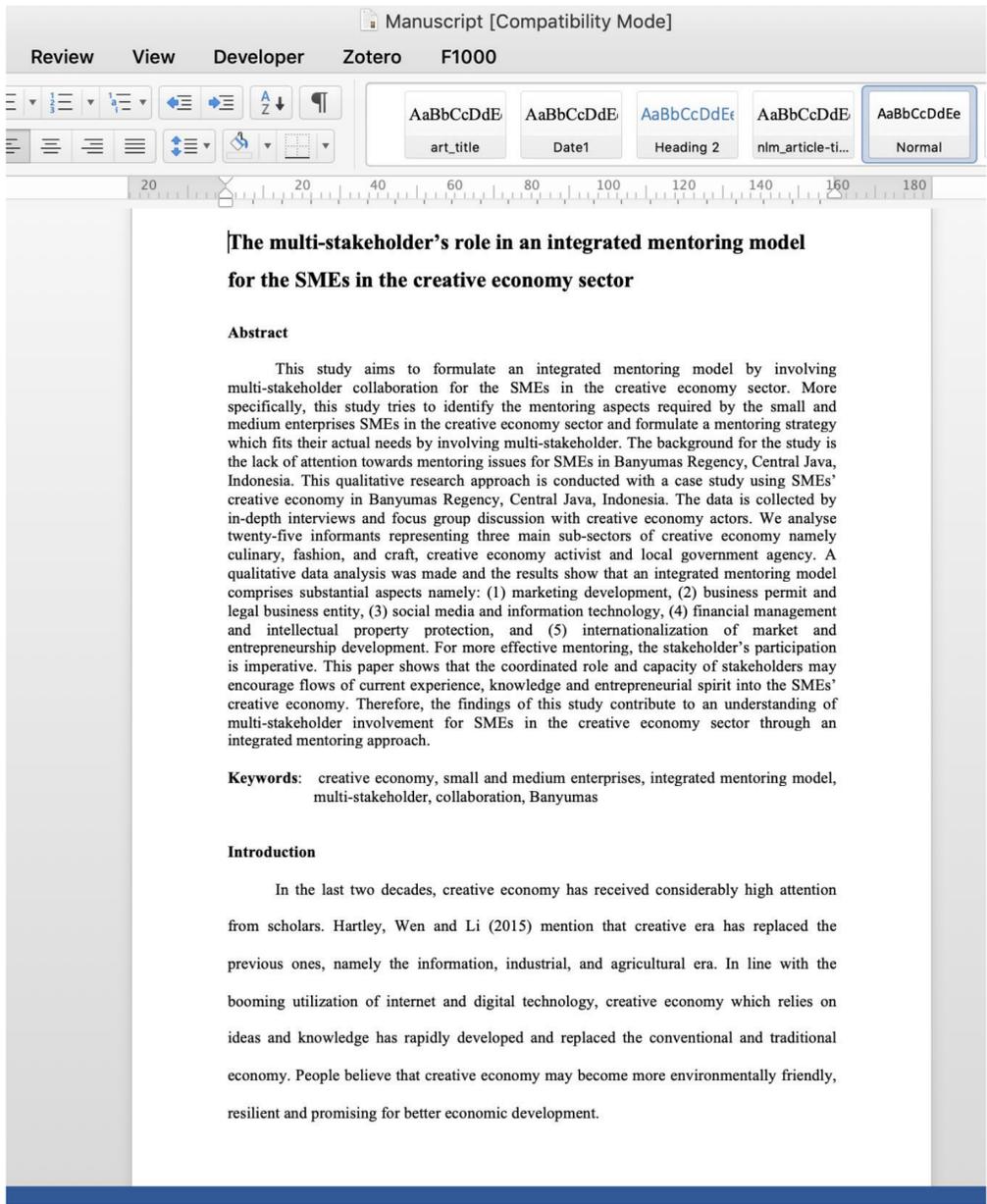
[10.1038/s41746-019-0180-3](https://doi.org/10.1038/s41746-019-0180-3)

“ This study has demonstrated that higher relative humidity and wind speed, and lower atmospheric pressure, were associated with increased pain severity in people with long-term pain conditions ”



- Key concepts
- Abstract
- Scholarly highlights
- Scholarly summary
- Tables
- Figures
- Introduction
- Methods
- Results
- Discussion
- Conclusion
- Funding
- Participants and statistics
- Limitations
- Ethics
- References (30)

Distilling research into headlines and key information



The multi-stakeholders role in an integrated mentoring model for the SMEs in the creative economy sector

2015

“ Our study results show that SMEs in [creative economy](#) sector need mentoring activities to empower them in the field of [financial management](#), marketing, protection of [intellectual property](#), legal business entity, utilization of [information technology](#), and [business permit](#) ”



Key concepts

Abstract

Scholarcy highlights

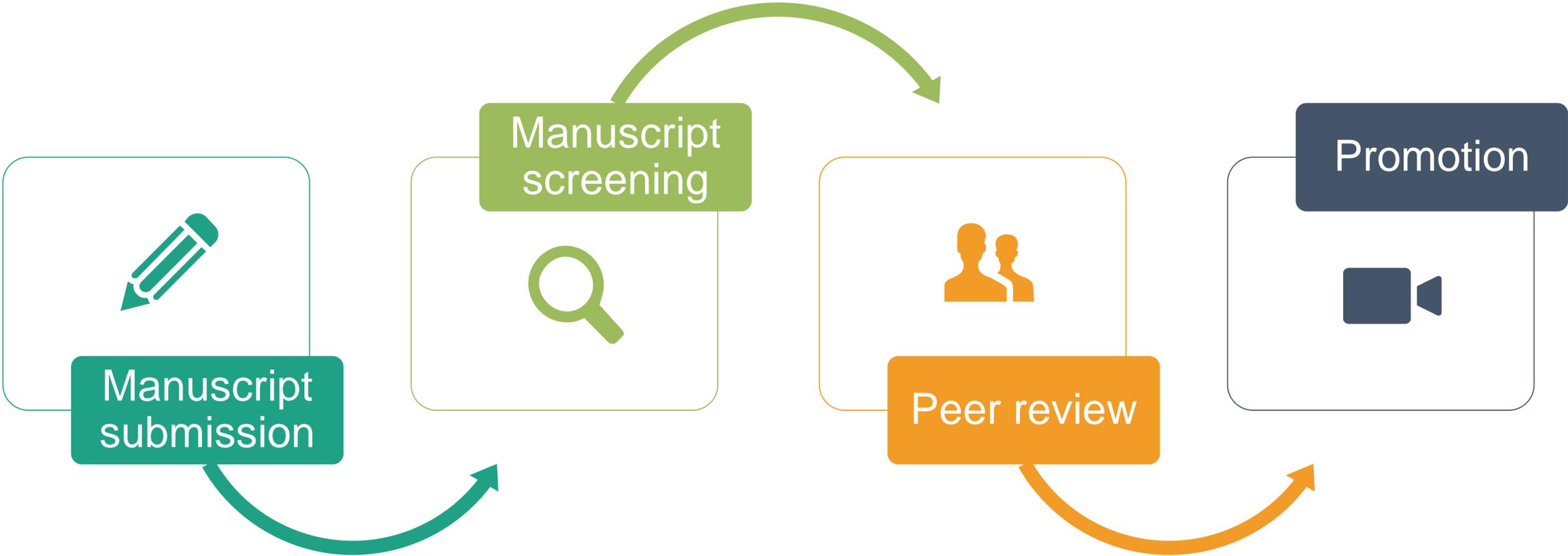
- In the last two decades, [creative economy](#) has received considerably high attention from scholars. [Hartley, Wen and Li \(2015\)](#) mention that creative era has replaced the previous ones, namely the information, industrial, and agricultural era
- Based on the results of focus group discussion with the informants, there are some important aspects to include in the mentoring model: (1) [financial management](#), (2) marketing, (3) protection of [intellectual property](#), (3) legal business entity, (4) utilization of [information technology](#), and (5) [business permit](#)
- Our study results show that SMEs in [creative economy](#) sector need mentoring activities to empower them in the field of [financial management](#), marketing, protection of [intellectual property](#), legal business entity, utilization of [information technology](#), and [business permit](#)
- Such mentoring activities should be integrated into a mentoring program for SMEs to improve their capacity to compete in [creative economy](#) business
- The results of this study confirm that the development of SMEs at local level through a mentoring model requires the [collaboration](#) of Penta Helix actors consisting of [local government](#), universities, business actors, creative community, and [social media](#) community
- The [social media](#) community is able to help the SMEs to promote the utilization of [social media](#) technology for wider market development of [creative economy](#) products

scholarcy

Opportunities

TDM: What are the opportunities?

TDM can work at any stage of the publishing process, opening up a huge number of opportunities from manuscript drafting and screening to promoting the published article.



Automating submissions process



- Metadata extraction to automate population of submissions system (Title, author, affiliations, abstract, keywords).
- Reduces author friction / duplication of effort.
- Previous work in this area has focused on the biomedical domain, but this opportunity can apply to any domain.

Speeding up peer review



- Data extraction for manuscript screening (key methods, results, sample size, participants, ethical compliance etc.)
- Clear article context/overview for reviewers.
- One-click access of cited sources & main findings.
- Table extraction for analysis of statistical calculations.

Surfacing cited sources & their main findings

medRxiv preprint first posted online Nov. 5, 2019 ; doi: <http://dx.doi.org/10.1101/19010991> . The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/) .

Krohn 14

Discussion

Our results confirm the association between *GBA* variants and increased risk for iRBD, and suggest that severe and mild *GBA* variants have differential effects on risk, similar to previous reports in PD.⁵ Furthermore, our results suggest that iRBD patients with severe *GBA* variants may have earlier AAO, and may convert faster to overt neurodegenerative disease. However, the results on AAO and conversion should be considered as preliminary only and with caution, due to several limitations discussed below.

Three previous small sample size studies have examined the association between *GBA* variants and iRBD.¹⁵⁻¹⁷ Two of these studies included full sequencing of the gene,^{15, 17} and the third only examined two specific variants (p.N370S and p.L444P).¹⁶ Due to their size, analyses of specific variants or types of variants such as severe or mild were not possible. The current study includes two of the previously published cohorts,^{16, 17} and additional cohorts of European ancestry. With the larger sample size accrued, we were able to demonstrate a much larger risk in carriers of severe *GBA* variants. However, given the small numbers of these variants and the wide range of the confidence interval, the risk estimates may be different in future, larger studies. Nevertheless, the current results are in line with previous results from PD, which clearly demonstrated similar relationships between severe and mild *GBA* variants and risk for PD.⁵ Furthermore, previous studies have also suggested that the type of *GBA* variants may affect PD progression,^{13, 14} which is further supported by our preliminary findings on AAO and conversion of iRBD.

In recent years, it has been demonstrated that the two coding variants, p.E326K and p.T369M, which do not cause GD, are risk factors for PD.²¹⁻²³ In DLB, the association between



Discussion

- "Our results confirm the association between *GBA* variants and increased risk for iRBD, and suggest that severe and mild *GBA* variants have differential effects on risk, similar to previous reports in PD.⁵ Furthermore, our results suggest that iRBD patients with severe *GBA* variants may have earlier AAO, and may convert faster to overt neurodegenerative disease. However, the results on AAO and conversion should be considered as preliminary only and with caution, due to several limitations discussed below.

Three previous small sample size studies have examined the association between *GBA* variants and iRBD.¹⁵⁻¹⁷ Two of these studies included full sequencing of the gene,^{15, 17} and the third only examined two specific variants (p.N370S and p.L444P).¹⁶ Due to their size, analyses of specific variants or types of variants such as severe or mild were not possible. The current study includes two of the previously published cohorts,^{16, 17} and additional cohorts of European ancestry. With the larger sample size accrued, we were able to demonstrate a much larger risk in carriers of severe *GBA* variants. However, given the small numbers of these variants and the wide range of the confidence interval, the risk estimates may be different in future, larger studies. Nevertheless, the current results are in line with previous results from PD, which clearly demonstrated similar relationships between severe and mild *GBA* variants and risk for PD.⁵ Furthermore, previous studies have also suggested that the type of *GBA* variants may affect PD progression,^{13, 14} which is further supported by our preliminary findings on AAO and conversion of iRBD.

In recent years, it has been demonstrated that the two coding variants, p.E326K and p.T369M, which do not cause GD, are risk factors for PD.²¹⁻²³ In DLB, the association between

Conclusions

- The risk for developing a clinically-defined Lewy body condition was similar in patients with idiopathic REM sleep behavior disorder with and without glucocerebrosidase gene variants
- Investigated whether *GBA* variants are overrepresented in IRBD and if their presence shortens the time to conversion to clinically-defined LBD

Funding

- All *GBA* coding exons from 69 polysomnography-confirmed IRBD patients and 84 matched controls were sequenced by the Sanger method

Participation

- The risk of developing a LBD was similar in IRBD subjects with *GBA* variants than in those without variants

Data and

- In IRBD, *GBA* variants are 1) more frequent when compared to controls, 2) associated with impending PD and DLB but 3) not indicative of a short-term risk for LBD after IRBD diagnosis

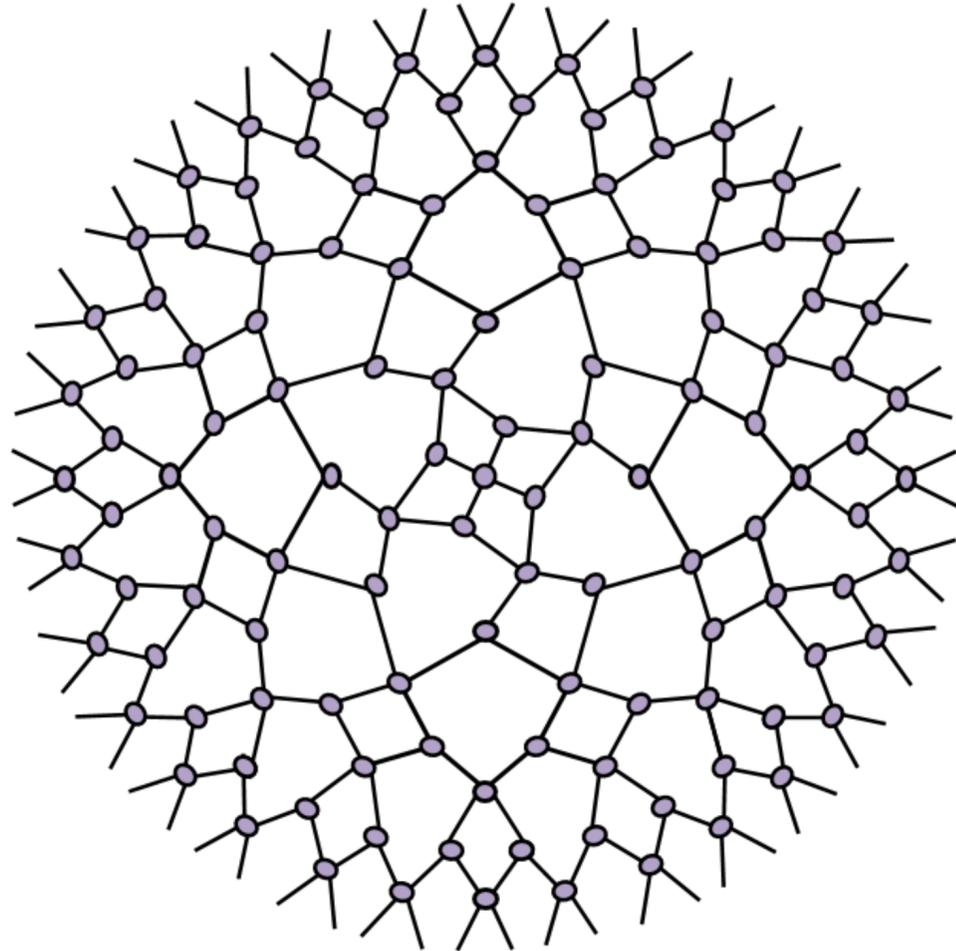
Ethics

References

Krohn L, Ruskey JA, Rudakou U et al. 2019. doi:10.1101/19010991

Cited sources and their main findings surfaced

Exposing more content through citation networks



- Extract, parse and link citations from archives dating back hundreds of years.
- Large scale reference population of open citation networks (BMJ Case study)
- Improve exposure/discovery of older research.

scholarcy

What's needed?

How publishers can help.



Make XML available for all Open Access articles rather than just the final PDF for text mining.



Make abstracts available so that citation networks can be enriched with key findings.



Help authors to write articles natively in a machine-readable format.

And finally...

...equal rights for friendly bots!

