Using Machine Learning to explore Scientific and Social media Engagement with Medical Publications

MSR203

Adamos Spanashis[1], Wenli Sun[2], Nicola Lazzarini[1], Avgoustinos Filippoupolitis[1], Simon Francis[1*] and Helen Stewart[1] [1]IQVIA, UK [2] IQVIA Inc., US

*simon.francis@iqvia.com

KEY TAKEAWAY

Machine learning and explainability techniques can be leveraged to predict engagement of studies in social media and scientific communities, providing insights into what drives engagement. We confirm that journals with higher impact factors tend to drive higher engagement.

INTRODUCTION

- Generating scientific evidence that can inform health care decisions and improve health outcomes is a top priority for medical affairs teams.
- Traditionally, impact of scientific evidence has primarily been measured by looking at citations in peer review journals.
- Social media like X (Twitter) offers a new dynamic environment for the dissemination and discussion of research findings, becoming crucial for the scientific communication of Medical Affairs teams. [1]
- To address these challenges, two complementary impact scores were defined to capture scientific publication engagement in both traditional scientific and social media environments.
- Two machine learning (ML) models were built to predict engagement and identify key factors driving interactions.

METHODOLOGY

STEP 1: We defined an impact score to quantify and aggregate the engagement of a scientific study on X:

$$social\ media\ score = a * \sqrt{tweets} + b * \sqrt{likes} + c * \log(followers)$$

The score considers the number of tweets and likes the study receives in the 12 months following its publication, along with the followers count of the user tweeting about it.

Impact with the scientific community is defined by the number of citations it receives within 24 months from publication, weighted by the Scimago Journal Impact Factor (SJR):

$$scientific\ score = \sum_i (no.of\ citations * SJR\ of\ journal\ i)$$

STEP 2: We selected 7,000 oncology papers published from 2017 to 2021 and used multiple data sources to characterise each study. For an unbiased performance evaluation, the data was split into training (80%) and test (20%).

STEP 3: We used discussions with Subject Matter Expert (SMEs) and data profiling to group the scores into categories:

Social media impact:
72.8% of studies: No Impact
24.5% of studies: Low Impact
2.7% of studies: High Impact

Traditional scientific media impact:
11.7% of studies: No Impact
61.8% of studies: Low Impact
17.7% of studies: Medium Impact
8.8% of studies: High impact

STEP 4: We developed two machine learning models [2], to predict the impact of an oncology publication in each domain.

STEP 5: We used SHAP (SHapley Additive exPlanations) [3] to interpret the model predictions and identify key factors driving publication impact.

DATA SOURCES

- **Scientific evidence:** clinical study attributes extracted from PubMed abstracts and meta data with IQVIA Applied AI Science NLP engine.
- **Author information:** from IQVIA Onekey, including metrics on individual health care professionals which characterise activity and influence in traditional scientific and social media domains.
- **Publication journal information:** extracted from Scimago's journal ranking.
- **X engagement:** collected using X and Crossref events APIs.

Study characteristics	Scientific evidence	Communication strategy
• Is the study randomised? • Type of study (clinical trial, observational) • Type of arm (placebo, active comparator) • Study design (blind/control) • Support of Real-World Evidence • Clinical outcome question (e.g., safety)	• Hazard ratio value for endpoint (e.g., OS) • Hazard ratio p-value for endpoint • Survival rate at x-years	• Journal impact factor (SJR) • Number of journal X followers • Author Digital Insights • Author Scientific Insights
Sources: IQVIA Applied AI Science and PubMed	Sources: IQVIA Applied AI Science and PubMed	Sources: OneKey (by IQVIA), X, and Scimago

Figure 1. List of features to build the models used

RESULTS

- The SJR is associated with both impact scores. As expected, the evidence published in journals with higher impact factor tends to achieve higher impact overall. However, our data show that high SJR values do not always guarantee higher engagement.

SJR vs. Impact scores

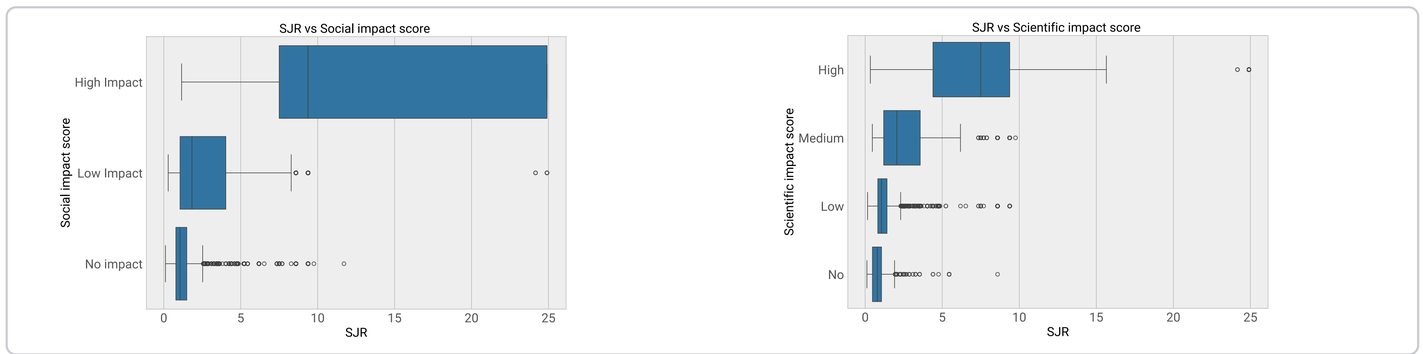

Figure 2. How SJR impact factor is associated with both impact scores

- Both ML models show best test performance when predicting High Impact papers in both social media and traditional scientific domains (17x and 7x better than baseline predictions based on prevalence).
- Differentiating between Low and No impact classes represents a more complex challenge, with predictive performance declining (ranging from 3x to 1.2x higher than prevalence baseline).

	High Impact	Low Impact	No Impact
Precision	48%	47%	85%
Recall	43%	54%	80%
F- score	46%	50%	83%

	High Impact	Medium Impact	Low Impact	No Impact
Precision	67%	47%	73%	76%
Recall	56%	27%	93%	26%
F- score	61%	34%	81%	38%

Figure 3. Predictive performance for social media (left) and traditional scientific media (right) impact models, per class. Recall measures the % of impactful papers correctly predicted. Precision indicates % of predicted papers which are impactful. F-score is the harmonic mean of the two. A perfect model obtains scores of 100%

- In both models, SJR emerges as the most potent predictor. For social media engagement, the follower count on X is the secondary key driver. Conversely, traditional scientific engagement is influenced by the average citation count within the respective publishing journal. End point results are equally important for both engagement predictions.
- The authors' importance metrics, indicating their relevance, does not contribute to the prediction of social media engagement. It holds higher relevance however in the traditional scientific context, as expected.

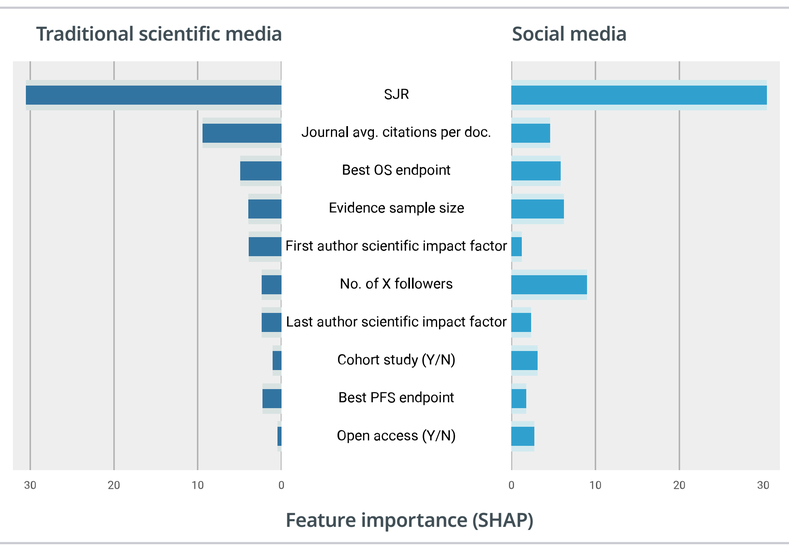

Figure 4. Top predictive features (using SHAP values). The length of the bar indicates the importance of each feature.

CONCLUSIONS

Machine learning and explainability techniques can be leveraged to predict engagement with study publications in social media and traditional scientific domains, providing insights into what drives engagement and guide teams in optimising their decisions.

The SJR emerged as the primary factor driving engagement on X and within the traditional scientific domain. The followers count of the journal on X also influences the social media engagement, but has less impact in traditional settings, where the average citation count per paper holds more value.

The suggested methodology can be extended to include other social media platforms (e.g., Research Gate) to understand different types of impact.

BIBLIOGRAPHY:

1. Fang Z, Costas R, Tian W, Wang X, Wouters P. An extensive analysis of the presence of altmetric data for Web of Science publications across subject fields and research topics. *Scientometrics.* 2020;124(3):2519-2549.
2. Ke, G, et al. LightGBM: a highly efficient gradient boosting decision tree. presented at: *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017; Long Beach, California, USA
3. Lundberg, et al. A unified approach to interpreting model predictions. presented at: *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017; Long Beach, California, USA

ABBREVIATIONS: AI, artificial intelligence; **API**, application programming interface; **avg**, average; **DOC**, document; **ML**, machine learning; **N**, no; **NLP**, natural language processing; **OS**, overall survival; **PFS**, progression-free survival; **SHAP**, SHapley Additive exPlanations; **SJR**, Scimago Journal rating; **Y**, yes.