POMELO: Protocol for Social Media Listening Online

Philipp Cimiano[1,2], Nicole Brazda[1], Matthias Hartung[3], Cornelius Starke-Knäusel[1], Ana Lucia Schmidt[4], Maria Carmela De Vuono[5], Aditya Tyagi[6], Jürgen Gottowick[4], Raul Rodriguez-Esteban[4], Ben Collins[7], Krzysztof Wieckowski[7], Thierry Escudier[7]

[1]Semalytix GmbH, Bielefeld, Germany, [2]Faculty of Technology, Bielefeld University, Bielefeld, Germany, [3]Boehringer Ingelheim GmbH, Ingelheim, Germany, [4]Roche Innovation Center Basel, Basel, Switzerland, [5]Chiesi Farmaceutici S.p.A., Parma, Italy, [6]Pistoia Alliance, Wakefield, MA, United States, [7]Boehringer Ingelheim International GmbH, Ingelheim, Germany

ABOUT POMELO

Social media listening emerges as a passive, observational method that uses algorithmic analysis to interpret insights voluntarily and publicly shared by patients across online platforms, forums, and communities. Social media data has been recognized by regulators such as the FDA and EMA as a valuable source of patient experience data. However, clear methodological guidelines and best practices for generating rigorous evidence from this data have been lacking. To address this gap, the Pistoia Alliance's Social Media for Real-World Evidence Expert Community Group is working to advance the responsible adoption of social media listening - ensuring that patient voices shared openly online are systematically captured, understood, and considered in drug development.

The Expert Community Group has developed POMELO, the Protocol for sOcial Media listEning Online, to define key aspects that need to be considered when setting up a social media listening study. POMELO provides guidance on important methodological choices helping organizations apply social media listening in a consistent, transparent, and scalable way.

The protocol is illustrated by an example social media listening study aimed at understanding how Type 2 Diabetes patients are affected by different co-morbidities.

CASE STUDY

01 START

DEFINE RESEARCH QUESTION

- What conditions (symptoms, comorbidities, complications, etc.) do patients experience and how severe are they?
- How do these conditions affect patients' daily life / quality-of-Life (QoL)?

02

DETERMINE REQUIRED SKILL SET

- Data scientists developing AI models and responsible for ETL process
- Medical experts and language experts defining content
- RWE/Epidemiologists for data analysis and interpretation

03

BASE INCLUSION CRITERIA ON PATIENT REPORTS

Patients were required to

- explicitly mention suffering from or being diagnosed with T2DM
- or report an HbA1c of 6.5% or higher in at least one post

04

OPERATIONALIZE RESEARCH QUESTION

- Research question was operationalized using a conceptual patient experience model (CPEM)
- Level of symptom or comorbidity severity as described in patient posts
- QoL was categorized using existing schemes as identified in single posts

05

CHECK IF APPFOVAL FROM ETHICS COMMITTEE IS REQUIRED

- Purely passive observation of publicly shared, pseudonymized data without targeting vulnerable groups was viewed as acceptable without explicit consent or ethics approval
- Beyond the General Data Protection Regulation (GDPR, see below), no other regulations or national laws to comply with were identified

06

IDENTIFY SUITABLE PUBLIC FORA

- The T2DM study relied on available search interfaces to query for online communities containing diabetes related keywords
- A manual analysis of a sample of posts for each forum was performed to ensure relevance of the data
- 32 online forums were included in the case study

07

DECIDE ON THE APPROACH TO COLLECT DATA

The content of identified forums was crawled via data provider APIs.

08

COMPLY WITH TERMS AND CONDITIONS OF FORA

- Sites whose terms do not exclude crawling for non-commercial purposes were selected; licensing agreements with data providers were established where necessary.
- Terms and conditions were recorded.

09

PROTECT DATA AND RESTRICT ACCESS VIA TOMS

The technical and organizational measures (TOMs) comprised

- A database was set up to securely store the non-anonymized data. Only selected system administrators had access to this database.
- Once data was anonymized and transferred to an analysis database (see below), the non-anonymized data was deleted after a period of 30 days.

10

ANONYMIZE DATA BEFORE PROCESSING

- State-of-the-art AI-based licensed 3rd party software was used to automatically remove any personal identifiers from the T2DM data
- Data scientists in the project were regularly trained on data protection matters.

11

IDENTIFY SUITABLE ALGORITHMS

- Patient affliction: rule-based algorithm to identify patient statements conveying diagnosis of T2DM.
- HbA1c Extraction: rule-based algorithm to extract HbA1c values
- Severity Classifier: supervised algorithm to classify symptoms and comorbidities into discrete severity levels.

12

IDENTIFY AUTHORS ROLE USING ALGORITHMS

A supervised algorithm was applied to classify whether a post originated from a patient, caregiver, or healthcare professional (HCP).

13

ASSESS THE PERFORMANCE OF ALGORITHMS

The accuracy of the severity and patient affliction classifiers was evaluated on a hold-out data sample. The accuracy of the severity classifier was 0.77. The accuracy of the patient affliction classifier was 0.88.

14

APPLY STRATEGIES FOR HANDLING FAKE INFORMATION

Advertisements and obvious fake content was filtered out using a heuristic approach, removing any author with more than 5,000 posts within the two-year observation period.

15

CHOOSE APPROPRIATE ANALYSIS METHOD

A mixed methods approach consisting of quantitative and qualitative analysis was applied:

- Quantitative: Frequency distributions of and severity levels.
- Qualitative: LLM-driven summaries using Retrieval-Augmented Generation (RAG)

16

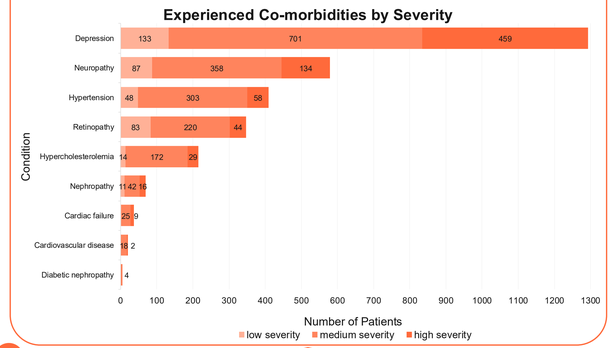
TRIANGULATE RESULTS

The study aimed at supporting a conceptual model of the disease. The results are ready for triangulation with other methods or patient organizations.

17

REPORT RESULTS

Results of the study were reported through descriptive statistics, including pie charts for gender distribution, stacked bar charts for comorbidity severity, and tabular qualitative summaries with paraphrased quotes.



FULL PUBLICATION DOWNLOAD

Cimiano P, Brazda N, Hartung M, et al. Best practices for the collection and analysis of patient experience data from social media for patient-focused drug development. Frontiers in Medicine. 2025;12

Contact Information: cimiano@semalytix.com

FUNDED BY PISTOIA ALLIANCE – JOIN US

We are looking for more companies to act as funders and contributors, if you are interested, please contact: aditya.tyagi@PistoiaAlliance.org

@pistoiaalliance www.pistoiaalliance.org