SEMALYTIX

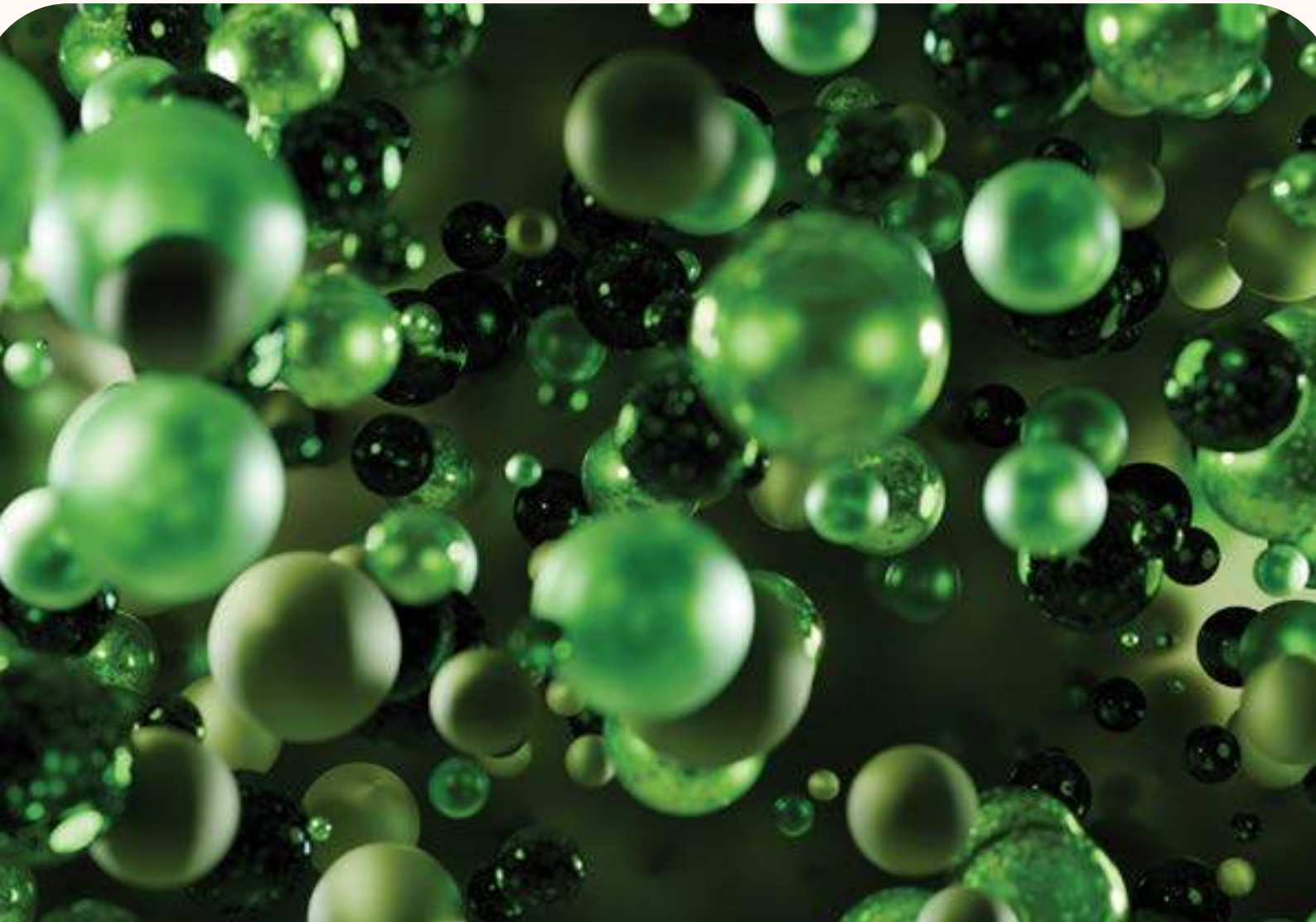
White Paper

The Semalytix Patient Listening Methodology

Prof. Dr. Philipp Cimiano, Dr. Thomas Andreu, Janik Jaskolski MSc

Semalytix GmbH

05 January 2025



About Semalytix

Semalytix's mission is to improve patient-centricity, inform patient engagement, and provide insights for developing more patient-focused drugs and therapies.

Semalytix is a Germany-based AI startup that has developed PatientGPT™ and Pharos™.

Both are AI-powered patient-centricity solutions. They amplify the global patient voice, making it easier and more scalable to identify unmet patient needs and to gain a deeper understanding of how people genuinely experience diseases.

Semalytix's expertise in artificial intelligence, natural language processing, and Large-Language Models (LLM) enables them to comprehensively anonymise and accurately analyze vast amounts of real-world patient experience data in over 25 languages. This unique approach positions Sematytix at the forefront of aiding pharmaceutical companies in designing patient-centric strategies, improving patient engagement and advancing patient-focused drug development.

The company is focused on establishing new, scalable, and reliable solutions that improve patient understanding for a simple reason, patient-centricity will only be achieved and maintained over time if complexity, cost, and the time needed to create insights, identify unmet needs, and even craft entire, patient experience data-based patient journeys, are dramatically reduced.

This white paper describes the methodology applied by Semalytix to extract patient insights from different data sources in which patients discuss their condition with peers. These data sources include general and disease-specific discussion fora, social networking sites, patient communities and support groups.

Semalytix has developed a proprietary analytical stack that comprises supervised machine learning components that are able to detect key concepts discussed by patients in online fora.

Example of key concepts detectable by these models are:

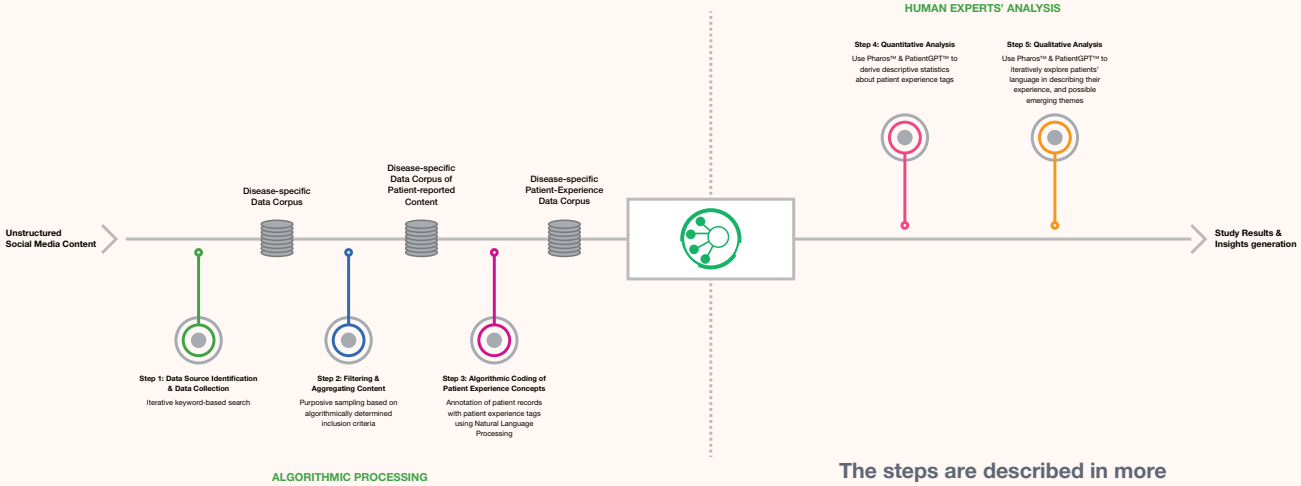
- Demographic variables, including gender and age of patients
- Symptoms mentioned by patients including their perceived severity
- Patient affliction: the symptoms and condition that patients experience
- Quality of Life aspects discussed by patients, captured with respect to the WHOQOL taxonomy
- Treatments discussed by patients and whether they had a positive or negative experience and sentiment
- Aspects related to treatment management
- Mentions of adherence problems and reasons for treatment discontinuation

For each of these aspects, Semalytix has developed specific models that have been trained with data annotated in-house by qualified and trained annotators. Based on these models, the experience of patients can be algorithmically extracted from unstructured data, that is from the posts shared by patients on online fora. For this purpose, transformer architectures and other types of models are trained to minimize the errors on a training set. Hyperparameters are optimized with respect to validation sets as is standard in machine learning based methods.

The methodology comprises in particular the following steps:

1. Data Source Identification and Data Collection
2. Filtering and Aggregating Content
3. Algorithmic Coding of Patient Experience Concepts
4. Quantitative Analysis
5. Qualitative Analysis

The following flow diagram depicts the different steps and their intermediate products.



The steps are described in more detail in the following sections.

I Data Source Identification and Data Collection

For each project and research question, Semalytix identifies the most relevant publicly available fora and social networking sites in which patients living with the targeted condition discuss their experiences. The research and studies are therefore solely based on unsolicited peer-to-peer conversations.

Semalytix has contracts with different providers and access to more than 150 Mio. different sites across the world and across geographies. As an important part of Quality Assurance, it is of key importance that the fora selected are validated by the in-house analysts based on reviewing a set of sample posts from each forum candidate.

Once a set of data sources / fora have been manually identified, the crawling functionality at Semalytix downloads the complete web fora to ensure that not only single posts are captured, but the context of posts and patients (thread) and the sequence of their messages is included as much as possible. To ensure compliance with General Data Protection Regulation (GDPR) and to ensure that no (real) patient can be identified, each post is automatically anonymized by removing any user ID as well as personal attributes such as names, email addresses and telephone numbers.

If the posts are not written in English, they are automatically translated into English using state-of-the art translation services. At the moment, Semalytix supports over 25 languages including all major European languages.

II Filtering and Aggregating Content

Based on the research question targeted, inclusion and exclusion criteria are defined and applied to the collected dataset to determine a set of patients to analyze further. Here, the Natural Language Processing (NLP) stack of Semalytix is used to extract the above-mentioned key concepts (e.g., symptoms, severity, Quality of Life (QoL) aspects, treatments, etc.). Based on the extracted variables, it is determined if patients identified in the dataset satisfy the inclusion criteria. Typically, in this step only patients are selected that mention to suffer from a condition or symptom of interest (called “patient affliction”). Additionally, certain segments of patients can be selected based e.g. on gender and age.

III Algorithmic Coding of Patient Experience Concepts

The algorithmic coding of patient experience concepts as used in this Semantic Social Media Listening (SSML) approach, goes beyond simple keyword tracking commonly associated with social media listening; it leverages a scientifically grounded ontology (Pharma Knowledge, PK), where patient terms and language are systematically coded in a bottom-up manner, linking individual expressions to comprehensive, research-backed scientific terms.

The Semalytix pharma knowledge is structured information on diseases, symptoms, treatments, comorbidities, and many other topics. The PK data base is designed, filled, and maintained by medical as well as linguistics experts. It contains disease-specific medical information, as well as all terms and synonyms of relevance for each entry, and relations between these entries.

Examples of sources building the pharma knowledge framework are:

- https://www.fda.gov/
- https://www.ema.europa.eu/
- https://www.nhs.uk
- https://www.ncbi.nlm.nih.gov/
- https://www.who.int/news-room/fact-sheets/
- https://www.cdc.gov/nchs/icd/index.htm
- https://en.wikipedia.org
- https://www.mayoclinic.org
- https://clinicaltrials.gov/

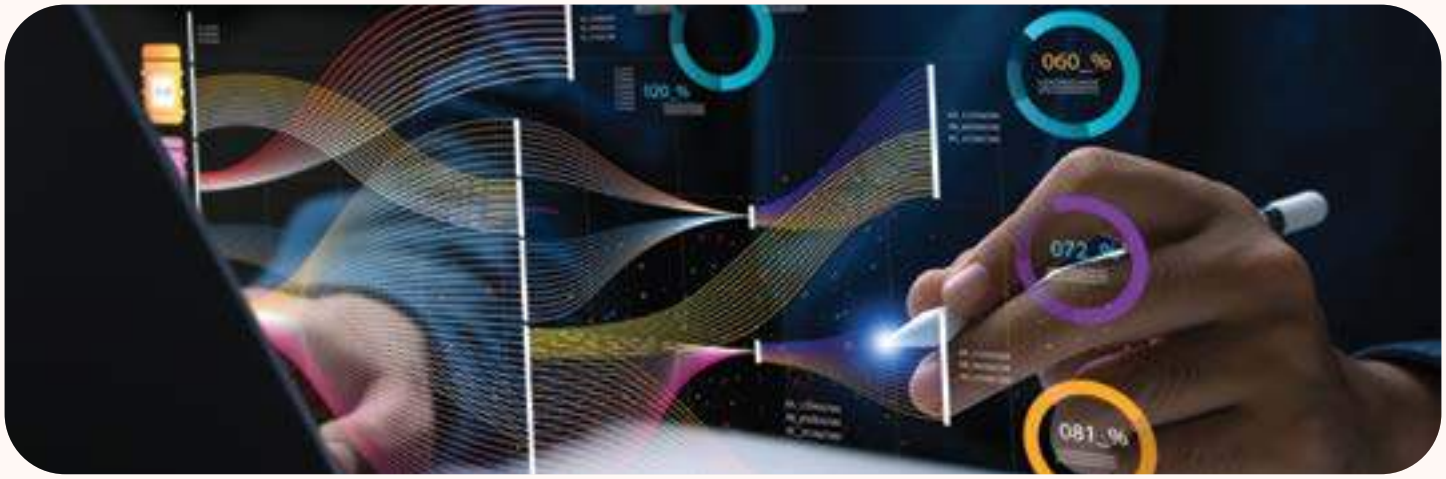
The pharma knowledge-based information is finally implemented as a concept and describes a topic of interest which can be detected in the text data. Concepts are defined by medical experts and then implemented using distinct kinds of AI (NLP analyzers).

The dataset selected as part of step II is automatically analyzed using the full set of NLP analyzers available at Semalytix.

In particular, the following analyzers are available:

Age Analyzer: extracting age information mentioned explicitly by patients.

Gender Analyzer: extracting gender information from patients’ posts, if explicitly mentioned.



Country / Language Analyzer: identifying the country or language of the domain or of the author.

Author Type Analyzer: identifying whether the author is a patient, family member or caregiver.

Patient Affliction Analyzer: identifying the conditions / symptoms that a patient reports.

Severity of symptoms: The "severity" concept is aimed at detecting the degree of severity with which a person is affected by a disease based on their experiences and categorised in high, medium, and low: High severity, if the author gives linguistic cues for high significance. Medium severity, if the disease is merely mentioned without emphasizing the effects or giving cues for either overall high or low severity and Low severity, if the author downplays the effects of the disease or gives cues for insignificance.

Treatment Usage Analyzer: identify treatments mentioned as being used by the patient in question.

Targeted Sentiment Analyzer: identifying sentiment (positive, negative, neutral) towards different entities (Drug agents, Drug Products, non-pharmacological treatments, etc.). The concept describes the affective state in which a person describes their opinions or experiences. The idea of sentiment detection is to answer questions like "How does this patient experience a treatment with drug X?"

Quality of Life Analyzer: extracting quality of life concepts discussed by patients.

QoL Importance Analyzer: The "importance" (of QoL facets) categorizes the subjective influence a Quality-of-Life concept has on the patient's life into high, basic, and low and can thus be interpreted as the "intensity" or "severeness" of a Quality-of-Life facet. It answers questions like "How high is the perceived subjective importance of this concept for the patient?"

Treatment Management Analyzer: identifying statements related to treatment start, treatment discontinuation, treatment switch, treatment access, etc.

Impairment Analyzer: identifying statements in which a patient mentions explicitly that their Quality of Life is impaired / reduced by a certain symptom.

IV Quantitative Analysis

Based on the algorithmic analysis of the posts / data from all selected patients, a quantitative analysis can be carried out that is usually conducted by human experts based on the identified factors. Descriptive statistics are computed to understand key elements of the patient's experience. Typical analysis include:

- Frequency of symptoms experienced or mentioned by patients
- Frequency with which certain Quality of Life aspects are mentioned by patients
- Frequency of treatments experienced or mentioned by patients
- Statistical Association between concepts, e.g. based on Fisher's test, chi-square test or regression models
- Distribution of sentiment for specific Drug Products/ brands, Drug Agents or other non-drug treatments

In this step, statistical association analysis such as Fisher's test, t-test or other hypothesis testing methods could be applied. Regression analysis is also applied to understand correlations between variables.

V Qualitative Analysis

Beyond the descriptive statistics mentioned above, arbitrary research questions can be automatically answered relying on generative AI technology that is implemented as part of PatientGPT™, developed by Semalytix. PatientGPT™ relies on Large Language Models and a Retrieval Augmented Generation (RAG) architecture to determine the most relevant posts for a question to be answered. This allows a scalable explorative approach to form hypotheses in an agile fashion. PatientGPT™ accomplishes this by sampling patient statements from the overall patient experience database and leveraging constrained LLMs to answer questions, summarize insights, and generate outcomes strictly based on what patients have written in unsolicited posts from online sources. It is also worth emphasising the importance of maintaining a complete and transparent audit trail in every analysis, guaranteeing accountability, reproducibility, and scientific rigor at each step of the research process.

References

The above methodology has been used in a number of social media listening (SML) studies with demonstrated results, as conveyed by the following publications:

- **Exploring the Perspectives of Patients Living With Lupus: Retrospective Social Listening Study,** Spies E, Andreu T, Hartung M, Park J, Kamudoni P (2024), JMIR Form Res 8:e52768
- **Using AI- based technology to gain insights from Osteoarthritis patients in UK via Social Listening Authors:** Neil Betteridge, Gudula Petersen, Thomas Andreu, Matthias Hartung (2023) In Annals of the Rheumatic Diseases: Volume 82, Supp. 1, page 244
- **Exploration of Melanoma Patient-Generated Real-World Data Using an AI Based Social Listening Approach:** Tadmouri A, Alivon M, Andreu T, Hartung M, Ryll B, Rauch G, Kiecker F, Cimiano P (2022) In Value in Health 25(12): S476-S477
- **Retrospective Social Listening Study of Patients Living with Systemic Lupus Erythematosus (SLE): Understanding the Patient Experience,** Spies E, Andreu T, Koelling J, Hartung M, Kamudoni P, Park J (2022), In Value in Health 25(12): S438
- **An Exploratory Retrospective Social Listening Study to Identify Patient Experiences Associated with Cutaneous Lupus:** Erythematosus (CLE), Spies E, Andreu T, Koelling J, Hartung M, Kamudoni P, Park J. In Value in Health 25(12): S393
- **Continuous Post-Market Real World Evidence Generation from Online Drug Reviews using Natural Language Processing:** Matthias Hartung, Arne Kramer Sunderbrink, Soufian Jebbara, Yannick Loonus, Bassam Mokbel, Philipp Cimiano, In Proceedings of IQWiG Information Retrieval Meeting (IRM), 2022.
- **Automatically Analyzing Online Patient Experience Data with Natural Language Processing. An Instrument to Investigate Health Status and Help-Seeking Factors in Patients with Obesity,** Matthias Hartung, Nathalie Schwering, Yannick Loonus, Philipp Cimiano, Anna Jäger, Ben Collins, In Qual Life Res 30(Suppl 1): S28, 2021
- **My Daughter Loves the New Pens: Quantifying the Patient Experience with Machine Reading and Applied Semantic Computing**, Bichteler, A.; Collins, B. G.; Walter, S.; Wendler, K.; Koelling, J.; Loonus, Y.; Hoewelkroeger, J.; Matheus, C.; Jebbara, S.; Hommel, F.; Badmaeva, E.; Verissimo, S.; Mokbel, B.; Cimiano, P.; Hartung, M., In ISPOR 2019

Authors

Dr. Thomas Andreu is a Deployment Strategist at Semalytix, with extensive expertise in the pharmaceutical industry. His scientific knowledge and business skills include data and market analysis of health care markets, combining scientific rigor with strategic insights through customized software solutions.

Prof. Dr. Philipp Cimiano is the Chief Technology Officer and Co-Founder of Semalytix. He is an internationally renowned leading academic in the fields of semantic technologies, knowledge representation and natural language processing.

Janik Jaskolski MSc is the Chief Product Officer and Co-Founder of Semalytix. He is an innovative leader in AI life-science technology and has been instrumental in advancing patient-focused drug development through the development of AI/NLP technology.

Important Information

For information or permission to reprint (in whole or in part) this document, please contact Semalytix at **info@semalytix.com**

If you would like to discuss this report, please contact the authors.

To find the latest Semalytix content, please visit **semalytix.com**