How Health-Related Social Media Can Complement Traditional Real-World Evidence Approaches to Offer Unique Patient Insights
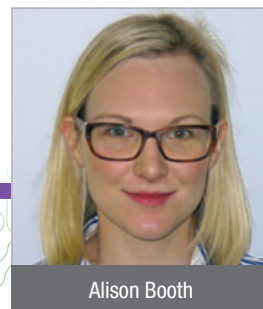
Alison Booth, MSc
Research Associate, Real-World Evidence
Evidera, a PPD business

Traditional real-world evidence (RWE) is based on data collected in normal clinical practice outside of randomized controlled trials. It is used to complement clinical data in regulatory and health technology assessment submissions. This data is most often generated through retrospective or prospective observational studies using electronic health records, medical claims, disease registries, etc. As more emphasis is placed on the importance of RWE, other sources of useful real-world data have been identified, specifically social media. Online technology platforms have allowed patients to interact with each other and provide unique insight into specific diseases, conditions, and treatments. Mining this health-related social media data provides invaluable real-world data.

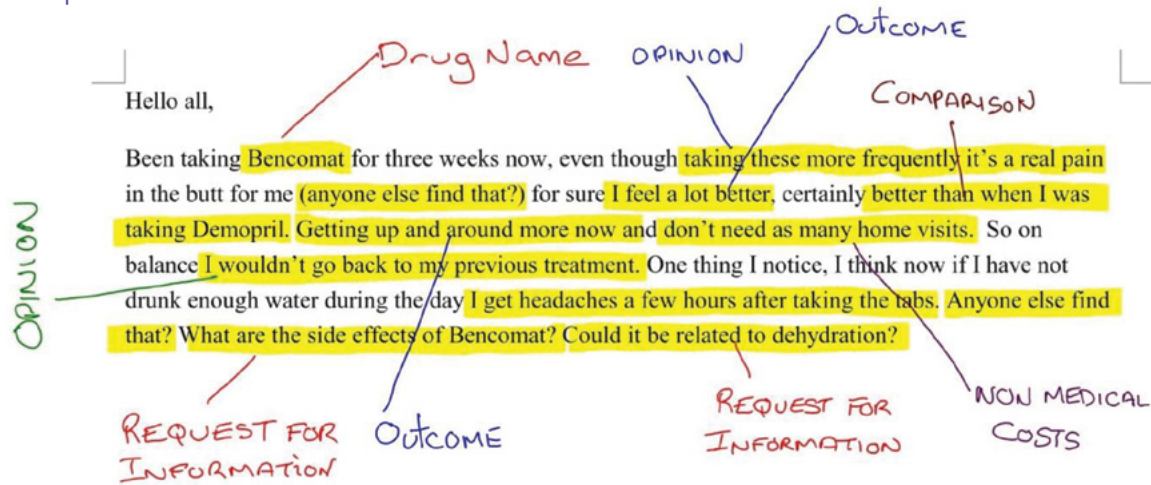## What is Health-Related Social Media Data?

When we think of social media, we typically think of generic sites like Facebook, Twitter, and Instagram. However, when discussing health-related social media we are referring to forums, blogs, and online communities that are often condition-specific and focus on discussions related to patient experiences such as the American Cancer Society's Cancer Survivors Network, HealthUnlocked, and the Psoriasis Association. There are thousands of similar communities used by both patients and caregivers spanning a broad range of indications, making health-related social media a rich source of real-world evidence.

Posts on these sites contain a vast amount of information ranging from a treatment or drug a patient has received and how long they have received it, information on symptoms

Alison Booth

Figure 1. Example of a Health-Related Social Media Post

and side effects, and subsequent impacts and questions to other community users. Figure 1 provides an annotated hypothetical post.

Social Media Listening as a Source of Real-World Data

Health-related social media is constantly growing and gaining momentum as a complementary source of real-world data for both payers and regulators. For example, in June 2018, the United States Food and Drug Administration (FDA) encouraged the use of social media to shed light on the patient's perspective, illustrating the increasing importance being placed on incorporation of the patient experience.[1] In addition, the FDA has its own programs for and perspectives on collecting information about adverse events (AEs) from social media.[2]

Another example illustrating the increasing importance placed on the patient experience is the proposed outcome-based payment model to link the price that the National Health Service (NHS) pays for a cancer drug to patient outcomes[3] (See Figure 2). Certain outcomes evaluated

Figure 2. Value of a Drug for Patient vs. Only Efficacy/Safety[3]
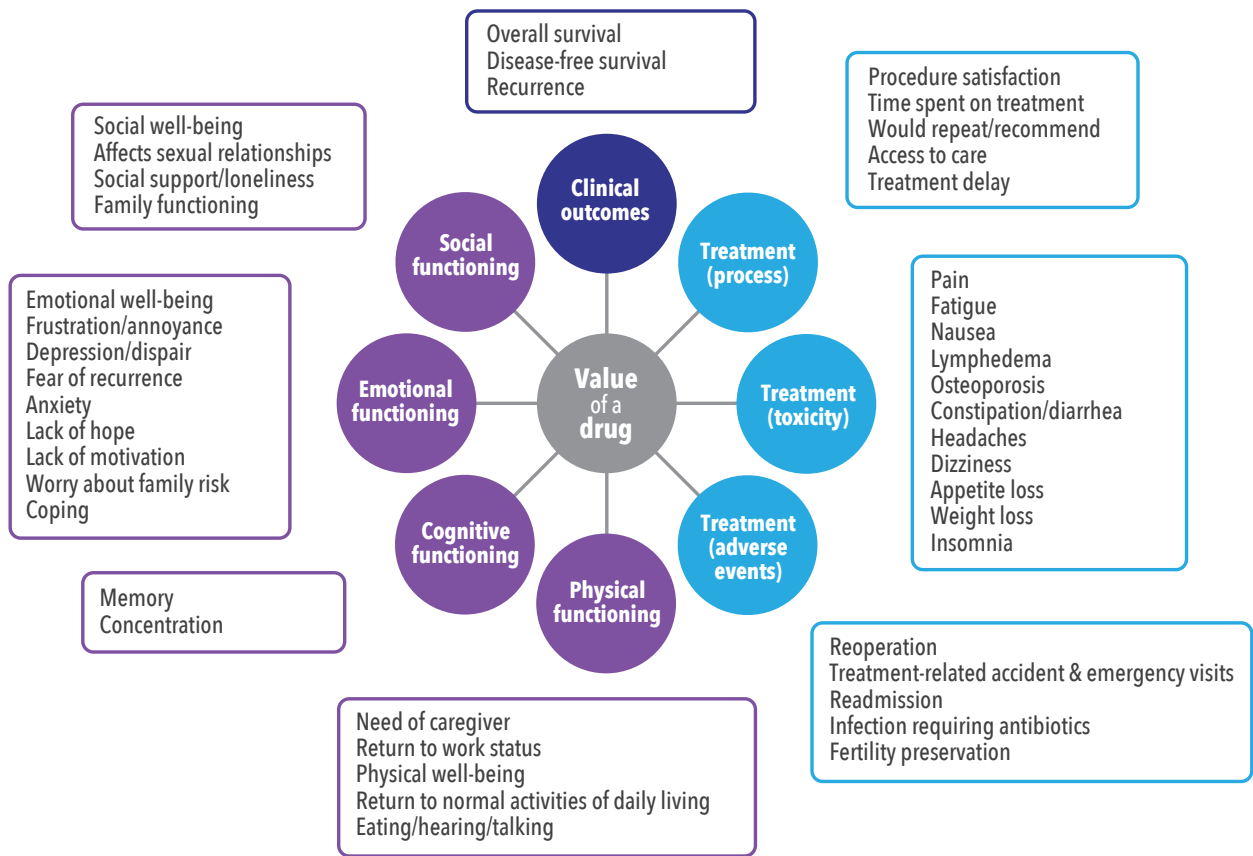
are clinical in nature, and can typically be leveraged from traditional sources such as those mentioned earlier; however, many of the other outcomes such as emotions or social functioning can be challenging or even impossible to capture using traditional sources. These represent areas where health-related social media can be used to complement traditional studies to provide a comprehensive picture of patient outcomes and experiences.

When Should Health-Related Social Media Be Considered?

It is important to think about when it is appropriate and impactful to use health-related social media. These are typically situations when the patient or caregiver perspective and experience are of interest. Health-related social media can be appropriate to gain these insights as the content is driven by patients and caregivers themselves and is therefore more likely to represent topics important to patients that may not have been considered clinically or by the research team. Furthermore, health-related social media can be particularly insightful in the case of emerging conditions where little is known about the patient experience and perspective, such as COVID-19. In the case of rare diseases where large groups of patients can be hard to find, health-related social media forums may be a place where patients from many locations come together. The incorporation of health-related social media should be considered when conducting studies investigating the following:

- **Unmet needs**
  - ▶ What elements patients and caregivers struggle with and research can help address

- **Burden of disease**
  - ▶ Economic and time impacts, impacts on work, resource utilization, social impacts, and more

- **Wider perspective of caregivers and family**
  - ▶ Caregivers often post on disease-specific social media forums, and patients also discuss the impacts of disease on their family

- **Issues important to patients**
  - ▶ Content is driven by the topics patients spontaneously mention and wish to discuss

- **Treatment experience**
  - ▶ Adverse events, holistic view of impacts on health-related quality of life, understanding of decision drivers for treatment choices, general opinions, perceptions and preferences about treatments

- **Populations and indications hard to find in traditional databases**
  - ▶ Rare disease, rapidly progressing conditions, new/emerging diseases (e.g., COVID-19)

Evidera has used health-related social media to inform patient preferences, study treatment patterns, perform sentiment analyses, gather patient and physician perceptions, enhance condition mapping to inform patient-reported outcome study design/instrument selection, analyze treatment decisions, and augment studies looking at safety by capturing adverse events.

How Do We Use Health-Related Social Media Data?

Once data is extracted from health-related social media sites, it needs to be analyzed in a way that produces useful insights that can be used to inform future clinical studies. There are several techniques that can be used to analyze the data (See Figure 3). **Natural language processing (NLP)** can be used to subset data to populations and discussions of interest, in addition to extracting frequently mentioned terms and topics. For example, Apache cTAKES is a natural language processing tool developed to extract clinical terms, such as disease symptoms or drugs, from text data. Custom lexicons can be derived using NLP to capture lay terms used by patients in addition to clinical terms. **Machine learning (ML)** has many applications but for social media data it can be particularly helpful to filter out noise (text that is not relevant to the study objective). Other applications of machine learning can also be explored based on specific study objectives. **Qualitative data analysis** is also important to social media data. The manual, in-depth analysis helps pull the full potential out of the data, allowing for deeper understanding of concepts and topics being discussed by patients.

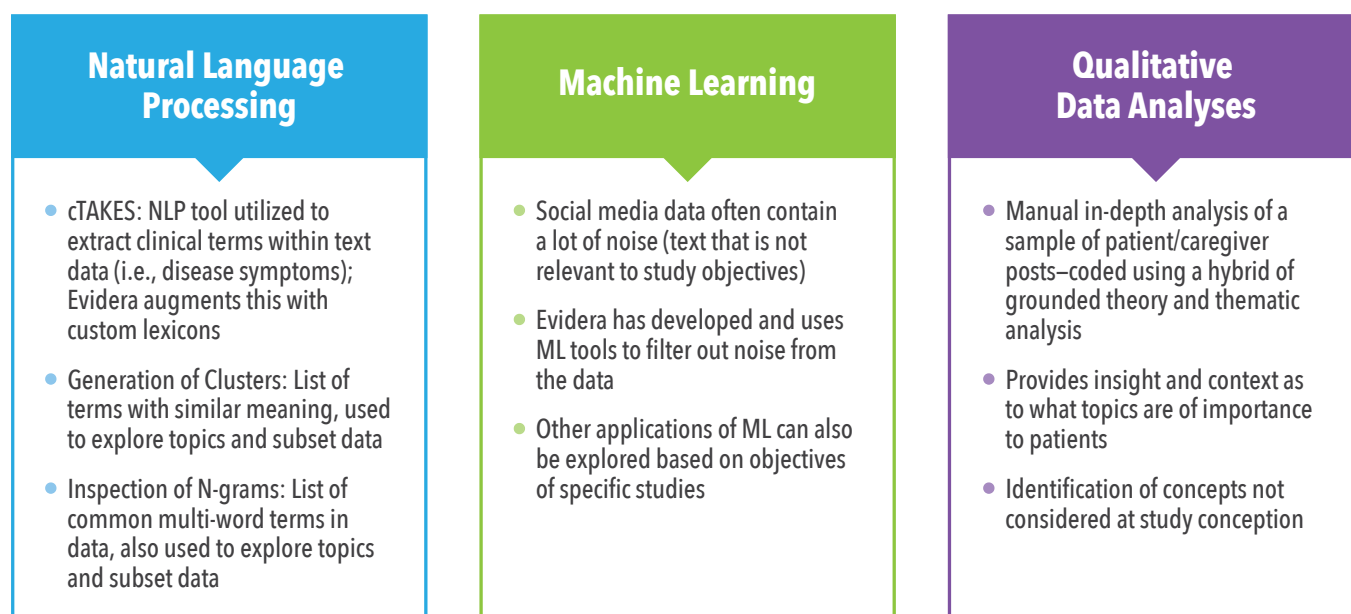## General Aspects of Social Media Analyses

We use posts from publicly available, condition-specific, social media forums to conduct studies. One consideration with health-related social media data is that the topics that may arise are not possible to specify a priori. Therefore, every social media study uses a scaled feasibility approach to mitigate risks and determine the appropriate sources and methods to use based on the specific research question.

Health-related social media posts can contain a large quantity of noise that is challenging to remove. Evidera has developed supervised machine learning algorithms to predict whether posts contain a true patient experience, and those posts are carried forward for analysis. A mixed-methods approach, combining quantitative and qualitative analyses, is often used to extract the most value out of the data as possible.

Health-Related Social Media and Ethics

Ethical considerations are extremely important when using any data, especially patient privacy and appropriate use of the data. While there are no specific guidelines on the use of health-related social media data, we believe it is important to follow specific rules to protect the privacy of patients and caregivers. For example, it is important to read

Figure 3. Common Types of Analyses

Natural Language Processing	Machine Learning	Qualitative Data Analyses
• cTAKES: NLP tool utilized to extract clinical terms within text data (i.e., disease symptoms); Evidera augments this with custom lexicons	• Social media data often contain a lot of noise (text that is not relevant to study objectives)	• Manual in-depth analysis of a sample of patient/caregiver posts–coded using a hybrid of grounded theory and thematic analysis
• Generation of Clusters: List of terms with similar meaning, used to explore topics and subset data	• Evidera has developed and uses ML tools to filter out noise from the data	• Provides insight and context as to what topics are of importance to patients
• Inspection of N-grams: List of common multi-word terms in data, also used to explore topics and subset data	• Other applications of ML can also be explored based on objectives of specific studies	• Identification of concepts not considered at study conception

the terms and conditions for each site and check whether there are any restrictions around extracting the content. We also look at the robot files, which may indicate whether elements of text can be programmatically retrieved. Any forum where text elements cannot be retrieved would not be extracted for a study. Additionally, only public, open-access forums should be used, as opposed to a closed forum that requires a login to view content, and researchers should not post to sites. This passive role is important for the integrity of the study and respect of patient privacy.

When using social media websites, seeking informed consent is often not feasible since it is not possible or practical to directly contact users. Per the ethics framework developed by the University of Sheffield,[4] steps should be taken to protect patient privacy and retain anonymity. All data should be de-identified and anonymized and posts or post content should not be reproduced verbatim or in a manner that allows the original post to be identified from study outputs.

Example Social Media Studies

The following section provides examples of how to utilize health-related social media to address different research needs.

EMERGING DISEASES
COVID-19

Population: Patients with breast cancer

Challenge: COVID-19 emerged in late 2019 and was declared a global pandemic by the World Health Organization in March 2020. To date there have been over 40 million cases and over 1 million deaths related to COVID-19 globally (as of October 20, 2020).[5] The aim was to understand the impact of COVID-19 on patients with other diseases, as well as their perceptions of COVID-19.

Approach: Extracted posts relating to COVID-19 from a large global breast cancer community and derived key themes and topics within the data using qualitative analyses.

Key Findings: The results of this study will be disseminated in late 2020, but we have already seen the prevalence of COVID-19 in health-related social media discussions. Real-time information and speed of access make social media a useful tool when information from other sources are lacking, such as with newly emerging diseases.

RARE DISEASES
Using Social Media and Advanced Analytics to Inform Study Design in AML and MDS[6]

Population: Patients with acute myeloid leukemia (AML) or myelodysplastic syndrome (MDS) ineligible for intensive chemotherapy

Challenge: The population is difficult to capture due to the rapidly progressing nature of the disease. The challenge was to understand patient preferences regarding end-of-life treatment and to attempt to uncover unmet needs.

Approach: Posts from three AML/MDS-specific forums were extracted, and NLP was used to obtain posts from patients who were ineligible for intensive chemotherapy. We then conducted a targeted qualitative review to extract the patient and caregiver insights.

Findings: Findings from this study were presented at the American Society of Hematology 2018 annual conference and have been published in a manuscript.[6] The study

Figure 4. Word Cloud AEs for Chemotherapy, Hormone Therapy and Targeted Therapy[8]

Chemotherapy

Hormone Therapy

Targeted Therapy

identified the desire of patients to be treated at home, suggested considerations for communicating information on treatment options, and highlighted the humanistic burden placed on patients and their caregivers.

TREATMENT PATTERNS
Extraction of Treatment Patterns from Health-Related Social Media Data

Population: Patients with metastatic renal cell carcinoma (RCC)

Challenge: To understand whether it is possible to accurately extract treatment patterns of patients with RCC in an automated manner from health-related social

media data using natural language processing, rule-based decisions, and machine learning.

Approach: Posts from metastatic RCC patients were extracted through a machine learning algorithm. Receipt of treatments of interest was identified using NLP and line of therapy was defined as the order in which the therapies of interest were administered.

Findings: While this work was exploratory, it showed that the patterns derived from the social media sources were within the range of estimates from the published studies for the majority of the treatments investigated. Findings from this study have been published in a manuscript.[7]

Figure 5. Heatmap of Top 25 Events Across Treatment Groups[8]

The most common AEs are shown on the horizontal axis and the treatment classes on the vertical axis. The frequency of associated mentions in posts are illustrated by the darkness of color, deep red having a high frequency of mentions and pale pink or white with fewest mentions. Overall, fatigue was the most commonly reported AE across chemotherapy and targeted treatment groups. Targeted therapies appeared to have a lower AE profile than either chemo or hormone-based therapies. Hormone therapy was strongly associated with flashes.
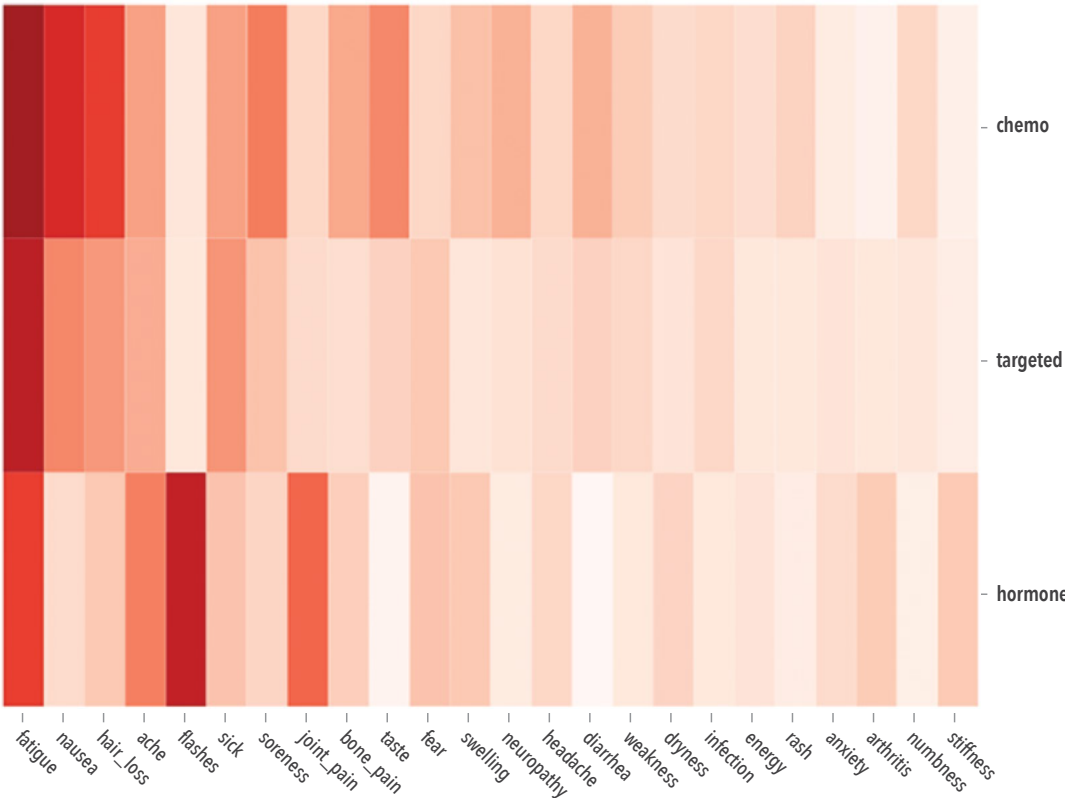
CAPTURING ADVERSE EVENTS
Adverse Event Profiling of Treatments for Breast Cancer

Population: Patients with breast cancer receiving chemotherapy, targeted therapy, or hormone therapy

Challenge: To identify symptoms and AEs from a large amount of unstructured text extracted from health-related social media forums to determine if this approach could provide novel information.

Approach: Posts were programmatically extracted from a large breast cancer community. After data cleaning and deidentification, AEs and symptom mentions were extracted using a lexical NLP approach, accounting for clinical and lay terms. Co-occurrences of treatment mentions, and symptom/AE mentions, were calculated for each treatment group (See Figures 4 and 5).

Findings: In addition to commonly reported symptoms and AEs, the study also uncovered less severe, or new and otherwise less frequently reported, symptoms/AEs that may have a significant impact on patients' quality of life. Supplementing traditional approaches through analysis of social media can generate additional insights and can enhance current approaches toward incorporating the patient perspective into healthcare research. Findings from this study will be presented at the International Society of Pharmacoeconomics and Outcomes Research (ISPOR) Europe conference in November 2020.

Conclusion

Health-related social media is a novel, growing, and constantly updated source of real-world data that has great value in uncovering patient experiences and perspectives. Outputs from health-related social media data can inform future research questions and its use can help provide a comprehensive understanding of treatment and disease outcomes, including outcomes not possible to capture in traditional sources of real-world data. ■

For more information, please contact Alison.Booth@evidera.com

REFERENCES

1. US Food and Drug Administration. Patient-Focused Drug Development: Collecting Comprehensive and Representative Input: Guidance for Industry, Food and Drug Administration Staff, and other Stakeholders. June 2020. Available at: https://www.fda.gov/media/139088/download. Accessed October 18, 2020.

2. US Food and Drug Administration. FDA Perspectives on Social Media for Postmarket Safety Monitoring. November 15, 2018. Available at: https://www.fda.gov/media/122897/download. Accessed October 17, 2020.

3. Cole A, Cubi-Molla P, Pollard J, Sim D, Sullivan R, Sussex J, and Lorgelly P. (2019). Making Outcome-Based Payment a Reality in the NHS. Available at: https://www.cancerresearchuk.org/sites/default/files/obp_final_report_pdf.pdf. Accessed October 17, 2020.

4. The University of Sheffield. Research Ethics Policy Note no. 14: Research Involving Social Media Data. Available at: https://www.sheffield.ac.uk/polopoly_fs/1.670954!/file/Research-Ethics-Policy-Note-14.pdf. Published 2018. Accessed April 8, 2020.

5. World Health Organization. Coronavirus Disease (COVID-19) Pandemic. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed October 20, 2020.

6. Booth A, Bell T, Halhol S, et al. Using Social Media to Uncover Treatment Experiences and Decisions in Patients with Acute Myeloid Leukemia or Myelodysplastic Syndrome Who are Ineligible for Intensive Chemotherapy: Patient-Centric Qualitative Data Analysis. *J Med Internet Res*. 2019 Nov 22;21(11):e14285. doi: 10.2196/14285.

7. Ramagopalan SV, Malcolm B, Merinopoulou E, McDonald L, Cox A. Automated Extraction of Treatment Patterns from Social Media Posts: An Exploratory Analysis in Renal Cell Carcinoma. *Future Oncol*. 2019 Nov;15(31):3587-3596. doi: 10.2217/fon-2019-0406. Epub 2019 Sep 4.

8. Pan S, Halhol S, Booth A, Cox A, Merinopoulou E. Profiling of Disease Symptoms and Adverse Events: Does Social Media Augment Traditional Approaches? Presented at ISPOR 21st Annual European Congress - 2018; November 10-14, 2018; Barcelona, Spain.