Original Paper

A Neural Network Approach for Understanding Patient Experiences of Chronic Obstructive Pulmonary Disease (COPD): Retrospective, Cross-sectional Study of Social Media Content

Tobe Che Benjamin Freeman[1,2], BSc, DPhil; Raul Rodriguez-Esteban[1], PhD; Juergen Gottowik[1], PhD; Xing Yang[3], PhD; Veit Johannes Erpenbeck[4], MD; Mathias Leddin[1], PhD

[1]Roche Pharma Research and Early Development, Pharma Research and Early Development Informatics, Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd, Basel, Switzerland

[2]wordup development AG, CH-8006, Zurich, Switzerland

[3]Roche Pharma Research and Early Development, Pharma Research and Early Development Informatics, Roche Innovation Center Little Falls, F. Hoffmann–La Roche Ltd, Little Falls, NJ, United States

[4]Roche Pharma Research and Early Development, Immunology, Infectious Diseases and Ophthalmology Discovery and Translational Area, Roche innovation Center Basel, F. Hoffmann–La Roche Ltd, Basel, Switzerland

Corresponding Author:
Tobe Che Benjamin Freeman, BSc, DPhil
Roche Pharma Research and Early Development
Pharma Research and Early Development Informatics
Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd
Grenzacherstrasse 124
Basel, CH-4070
Switzerland
Phone: 41 793778595
Email: tobefreeman@gmail.com

Abstract

Background: The abundance of online content contributed by patients is a rich source of insight about the lived experience of disease. Patients share disease experiences with other members of the patient and caregiver community and do so using their own lexicon of words and phrases. This lexicon and the topics that are communicated using words and phrases belonging to the lexicon help us better understand disease burden. Insights from social media may ultimately guide clinical development in ways that ensure that future treatments are fit for purpose from the patient's perspective.

Objective: We sought insights into the patient experience of chronic obstructive pulmonary disease (COPD) by analyzing a substantial corpus of social media content. The corpus was sufficiently large to make manual review and manual coding all but impossible to perform in a consistent and systematic fashion. Advanced analytics were applied to the corpus content in the search for associations between symptoms and impacts across the entire text corpus.

Methods: We conducted a retrospective, cross-sectional study of 5663 posts sourced from open blogs and online forum posts published by COPD patients between February 2016 and August 2019. We applied a novel neural network approach to identify a lexicon of community words and phrases used by patients to describe their symptoms. We used this lexicon to explore the relationship between COPD symptoms and disease-related impacts.

Results: We identified a diverse lexicon of community words and phrases for COPD symptoms, including gasping, wheezy, mucus-y, and muck. These symptoms were mentioned in association with specific words and phrases for disease impact such as frightening, breathing discomfort, and difficulty exercising. Furthermore, we found an association between mucus hypersecretion and moderate disease severity, which distinguished mucus from the other main COPD symptoms, namely breathlessness and cough.

Conclusions: We demonstrated the potential of neural networks and advanced analytics to gain patient-focused insights about how each distinct COPD symptom contributes to the burden of chronic and acute respiratory illness. Using a neural network approach, we identified words and phrases for COPD symptoms that were specific to the patient community. Identifying patterns in the association between symptoms and impacts deepened our understanding of the patient experience of COPD. This approach can be readily applied to other disease areas.

XSL•FO
RenderX

Introduction

Online content made public by patients in blogs and on forum platforms provides detailed first person accounts of the lived experience of disease [1,2]. These communications from patients use a diverse vocabulary of words and phrases for disease symptoms [3]. Online content is conveyed in the patient's own voice and is contributed in the ecological context of day-to-day life [4], namely in the sharing of experiences with other members of the patient and caregiver community. Analysis of these online communications enables a patient-centric approach to understanding disease impact.

A systematic understanding of the language used by patients to describe their symptoms has important clinical implications, not least being the need to acquire accurate patient anamneses and respond to care needs [5]. Dreisbach et al [6] note that the use of normalized medical vocabularies supports a systematic approach to identify terms for clinical and subclinical symptoms. This approach enables the identification of community terms that, while not belonging to a traditional medical lexicon, denote respiratory dysfunction unambiguously.

Many researchers use interviews, focus groups, and patient advisory boards with a goal of observing patient experiences. These approaches enable direct observation of the patient; however, they tend to be a burden to patients [7]. Moreover, interviews and focus groups are generally limited to cohorts of just a few patients, and the results are qualitative in nature.

In contrast, machine learning and related computational techniques offer a means to analyze online content at scale. Current state-of-the-art approaches using neural network architectures are being deployed to map patient community terms onto controlled medical [8] and pharmaceutical vocabularies [3]. However, these approaches are anchored in a defined lexicon of scientific terms, thus compromising patient centricity. In a patient-centric approach, our understanding of disease should instead be anchored to patients' self-reported topics [7], as observed in the ecological context of daily life [4], and not exclusively anchored to expert medical thinking, as expressed in a scientific lexicon.

We address this limitation with a novel approach based on a neural network, specifically a word embedding [9], to identify words and phrases that patients with chronic obstructive pulmonary disease (COPD) use to describe their experiences of living with the disease. Unlike traditional neural network approaches, a word embedding is not trained on any specific set of scientific keywords [10,11].

We use the word embedding to identify a diverse lexicon of hundreds of COPD-related words and phrases from the context in which words appear in a text. Next, we use that lexicon to extract all mentions of words and phrases relating to COPD symptoms and disease impacts from a large corpus of social media text. Once extracted, we can analyze the relationship between COPD symptoms and disease impacts at scale.

The quantitative analysis of this diverse community lexicon reveals insights [6] about the lived experience of COPD. These insights can contribute positively to the development of effective medical treatments that are, from the patient's perspective, fit for purpose [12].

Methods

Ethics

This work is compliant with ethical guidelines for the collection and analysis of user-generated content on open internet platforms. Data were downloaded only from open health social networking sites and communities. No information from restricted data areas has been downloaded (ie, content that requires an ID or password for access). No aggregation or enrichment of data on an individual has been performed. Extracts used for exemplary purposes were carefully paraphrased to protect the privacy of individuals.

Data Availability

All social media content included in our analysis was sourced from open social networking sites and communities. Terms and conditions apply to the availability of the original social media data. The sources used in this study can be made available upon request. Example texts shown in this manuscript have been rephrased to prevent de-anonymization of the individuals included in our analysis.

Neural Network Methodology

We trained the neural network on a corpus of 1.1 million words sourced from 22 individual blogs and online forums (Multimedia Appendix 1). We used the skip-gram negative sampling variant of the word2vec neural network algorithm described by Mikolov et al [9] to discover community words and phrases for disease symptoms. Briefly, the neural network model was trained to predict context words that appear in close proximity with symptom keywords in the corpus text.

The resulting word embedding captured semantic and syntactic features of each unique word in the text corpus. Neighboring vocabulary items in the embedding will likely share semantic and syntactic features in common. We then used cosine similarity as a metric to probe the word embedding model for words and phrases that share common meanings. This makes it possible to build and expand a lexicon of community terms for each main COPD symptom type in a systematic and repeatable manner (Table S1 in Multimedia Appendix 1).

We started our search for community words and phrases for COPD symptoms with a small seed lexicon that included breathlessness, cough, and sputum. This seed lexicon was

sourced from MeSH terms from the US National Library of Medicine (NLM) [13] and from the NLM health information website for the layperson, MedlinePlus [14]. These 3 seed terms correspond to key pathophysiological manifestations of COPD, namely small airway fibrosis, emphysema, which refers to a destruction of the lungs' alveoli, and mucus hypersecretion [15-17].

We used the same approach to search for community words and phrases describing the impact of COPD on daily life. The seed terms for disease impacts include anxiety, depression, fatigue, pain, and exercise. We then scanned the entire corpus to detect posts in which COPD symptoms co-occur with mentions of disease-related impacts. Our analysis explored the relationship between specific symptoms and each of the main disease impact topics.

Results

Using the cosine similarity metric to probe the word embedding model, close neighbors of the symptom seed term breathlessness included gasping, wheezy, and the phrase pursed-lip (Table S1 in Multimedia Appendix 1). The phrase pursed-lip is noteworthy as it refers to a technique, called pursed-lip breathing, used in pulmonary rehabilitation. Specifically, pursed-lip breathing is used to manage anxiety associated with breathlessness [18]. Words and phrases neighboring the seed term sputum include

mucus-y, phlegm, clear mucus, and muck, as well as common misspellings of phlegm.

Probing the word embedding model with the seed term exercise, we found walk and the phrases low impact and difficulty exercising (Table S1 in Multimedia Appendix 1). These community terms are, as we might expect, for a relatively aged and exercise-limited patient cohort [19]. Manual inspection of individual excerpts from the corpus featuring symptom keywords further confirmed the relevance of these keywords (Table S4 in Multimedia Appendix 1).

Summing the number of mentions corresponding to each symptom lexicon across the entire corpus (Table S2 in Multimedia Appendix 1), the breathlessness lexicon was mentioned most frequently (mentioned in 10.49% [413/3938] of posts), followed by the lexicon for cough (270/3938, 6.86%) and, finally, mucus hypersecretion (159/3938, 4.04%).

Leveraging these distinct lexicons of symptoms and disease impacts (Table S3 in Multimedia Appendix 1), we were able to explore the relationship between specific symptoms and each of the main disease-impact topics. Figure 1 examines posts in which COPD symptoms co-occurred with mentions of disease-related impacts. The analysis shows that breathlessness was the symptom most frequently mentioned in association with the 4 main topics and impacts considered. The most frequent disease impact associated with COPD symptoms was fatigue, followed closely by self-reports of anxiety and depression.

Figure 1. Topics co-occurring with symptom mentions in the same post.
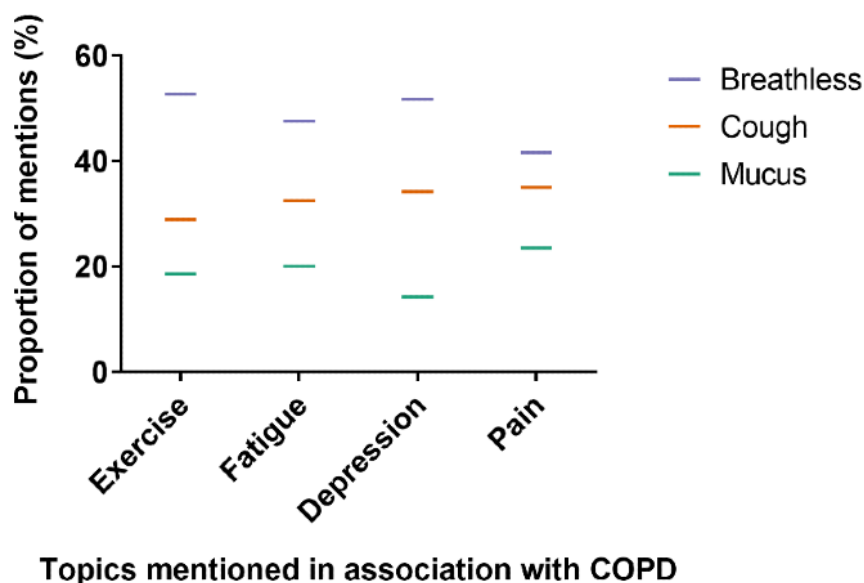
Breathlessness and cough followed a broadly similar trend, while the trend in the co-occurrence between mucus and the 3 disease severity levels was distinctive (Figure 2). The co-occurrence between mucus and mild severity was lower than that between mucus and moderate disease severity, inverting

the relationships observed for breathless and for cough. Taken together, it was apparent that there was an association between mucus and moderate disease severity that distinguished mucus from the symptoms breathlessness and cough.

Figure 2. Relationship between the symptoms of chronic obstructive pulmonary disease and disease severity.
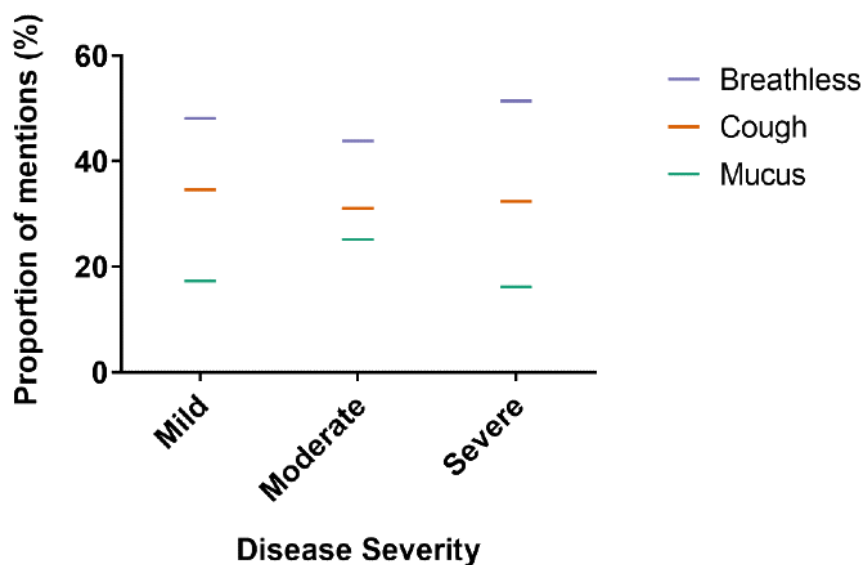
By applying principal component analysis (PCA), we visualized semantic relationships [20,21] between each symptom lexicon and a mapping of the psychological salience of these symptoms. PCA arranged data points corresponding to individual words and phrases on a 2D map [20] (see Multimedia Appendix 1 for further details). Our PCA results showed that words and phrases belonging to the 3 symptom lexicons were arranged in 3 distinct clusters on this map (Figure S1 in Multimedia Appendix 1).

By adding a lexicon of affective states such as feel depressed and be embarrassing to the PCA map, we could explore the psychological salience of these symptoms. The lexicon of affective states also appeared as a distinct cluster on the map and was positioned closest to the cough symptom cluster. The mucus cluster was displaced further away from the cluster of affective states than the cough cluster. Note, however, that the cough and mucus clusters were aligned along a single axis with respect to the cluster of affective states.

Discussion

Principal Findings

Our findings demonstrate the potential to deploy advanced analytics in the search for disease-related insights from hundreds of patients and many thousands of self-reports published online.

By probing a word embedding model trained on a corpus of online content contributed by COPD patients, we found a lexicon of community terms expressing a broad range of topics and meanings (Table S1 in Multimedia Appendix 1). Many terms found this way were related to COPD in a direct and intuitive fashion. And some terms revealed associations with unexpected, yet highly relevant topics (eg, pursed-lip) [18]. This term relates to the pursed-lip technique for managing anxiety associated with breathlessness.

The finding that breathlessness was the most frequently mentioned symptom accorded with medical consensus. As stated by the internationally recognized guidelines of the Global Initiative for Chronic Obstructive Lung Disease (GOLD)

[15,22], a decline in lung capacity, in combination with other disease-specific symptoms [23,24], forms the basis of a clinical diagnosis of COPD, and measurements of lung function and lung volumes are used to monitor disease progression [17].

In agreement with recent social media studies of COPD patients, our results highlight mucus hypersecretion as an important COPD symptom [25,26]. Compared with breathlessness and cough, mucus terms co-occur with mentions of moderate disease and co-occur less often with mild or severe disease. Similarly, when compared with breathlessness and cough, mucus symptoms were mentioned relatively less frequently when patients reported affective impacts of COPD such as depression.

These distinct associations relating to mucus hypersecretion were corroborated by a novel analysis using PCA to map the psychological salience of the 3 COPD symptoms. Relative to breathlessness and cough, mucus symptoms were mapped furthest from the affective impacts of COPD, suggesting that mucus has the weakest association with perceived affective impacts of the disease.

Mucus symptoms were mentioned at less than half the frequency that breathlessness was mentioned in the corpus. This finding is consistent with the GOLD report and reports indicating that not all COPD patients experience mucus hypersecretion as a symptom of their disease and that mucin concentrations are lower in COPD versus other obstructive lung diseases like cystic fibrosis or bronchiectasis [27]. And yet mucus hypersecretion is an important clinical factor in COPD. For example, mucus symptoms can motivate patients to take timely action against life-threatening respiratory infections [28]. Hypersecretion also drives cough symptoms and expectoration [15].

Without these advanced analytics, our insights about mucus symptoms would have been obscured by the overall dominance of breathlessness and cough symptoms mentioned in the corpus. Examining the co-occurrences between symptoms and disease impacts informed a deeper understanding of disease burden. The approach was able to quickly and accurately identify patient

populations whose experience was especially impacted by a particular symptom, adding greater potential for personalization.

This approach can ultimately guide clinical development in ways that ensure that future treatments are fit for purpose from the patient's perspective [12] and from the perspective of patients' perceived treatment needs.

Limitations

The forum content we included in the corpus had been posted anonymously and so we were unable to verify any bias arising from the demographics of forum contributors. Beyond the general guidance posted online by forum moderators, we could not explore biases introduced by a moderator removing posts from the forum.

We can expect a degree of clinical inaccuracy in the contributions posted by individuals who may not have formal medical training. Furthermore, the anonymity of social media makes it all but impossible to determine whether a post is authored by a genuine patient or caregiver or by someone merely posing as one. Taken together, any clinical interpretations we make from social media must take these uncertainties into account. However, because every post was manually reviewed, obviously fraudulent content from bots, scammers, and marketers was eliminated.

Despite limitations, the societal benefits that may be gained from large scale analysis of social media content are substantial, as researchers Gleibs et al [29] and Golder et al [30] have noted. The research community should ideally work closely with patients and health care advocates to ensure that people can continue to contribute to online forums and other social media platforms in a way that protects their privacy and ensures they are safe from potentially harmful misinformation.

Conclusions

Using a novel neural network approach, we demonstrate how online content can be a rich source of insights about the lived experience of COPD. Our findings demonstrate the potential of neural networks to gain a quantitative, patient-focused understanding about how each distinct COPD symptom contributes to the burden of chronic and acute respiratory illness. This approach can be readily applied to other disease areas in which there exists sufficient online content contributed by patients and caregivers.

Authors' Contributions

TF and RRE authored the paper. TF conducted the analysis. JG and XY sourced and prepared the corpus of content for downstream analysis. VJE and ML contributed to the analysis plan and manuscript writing.

Conflicts of Interest

VJE is an employee of F. Hoffmann—La Roche Ltd and holds stocks. JG and RRE are employees of F. Hoffmann—La Roche Ltd.

Multimedia Appendix 1

Further advanced analyses.
[DOCX File , 118 KB-Multimedia Appendix 1]

References

1. Gijsen V, Maddux M, Lavertu A, Gonzalez-Hernandez G, Ram N, Reeves B, et al. #Science: the potential and the challenges of utilizing social media and other electronic communication platforms in health care. Clin Transl Sci 2020 Jan;13(1):26-30 [FREE Full text] [doi: 10.1111/cts.12687] [Medline: 31392837]

2. Limsopatham N, Collier N. Normalising medical concepts in social media texts by learning semantic representation. Proc 54th Ann Mtg Assoc Computational Linguistics 2016:1014-1023. [doi: 10.18653/v1/P16-1096]

3. Tutubalina E, Miftahutdinov Z, Nikolenko S, Malykh V. Medical concept normalization in social media posts with recurrent neural networks. J Biomed Inform 2018 Aug;84:93-102 [FREE Full text] [doi: 10.1016/j.jbi.2018.06.006] [Medline: 29906585]

4. Taylor KI, Staunton H, Lipsmeier F, Nobbs D, Lindemann M. Outcome measures based on digital health technology sensor data: data- and patient-centric approaches. NPJ Digit Med 2020 Jul 23;3(1):97. [doi: 10.1038/s41746-020-0305-8] [Medline: 32715091]

5. van Rosse F, de Bruijne M, Suurmond J, Essink-Bot M, Wagner C. Language barriers and patient safety risks in hospital care. A mixed methods study. Int J Nurs Stud 2016 Feb;54:45-53. [doi: 10.1016/j.ijnurstu.2015.03.012] [Medline: 25840899]

6. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform 2019 May;125:37-46 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.02.008] [Medline: 30914179]

7.   Humphrey L, Willgoss T, Trigg A, Meysner S, Kane M, Dickinson S, et al. A comparison of three methods to generate a conceptual understanding of a disease based on the patients' perspective. J Patient Rep Outcomes 2017;1(1):9 [FREE Full text] [doi: 10.1186/s41687-017-0013-6] [Medline: 29757313]

8.   Weissenbacher D, Sarker A, Klein A, O'Connor K, Magge A, Gonzalez-Hernandez G. Deep neural networks ensemble for detecting medication mentions in tweets. J Am Med Inform Assoc 2019 Dec 01;26(12):1618-1626 [FREE Full text] [doi: 10.1093/jamia/ocz156] [Medline: 31562510]

9.   Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. ArXiv. Preprint posted online on January 16, 2013 2013:1 [FREE Full text] [doi: 10.3126/jiee.v3i1.34327]

10.   Mikolov T. Distributed representations of words and phrases and their compositionality. ArXiv. Preprint posted online on October 16, 2013 2013:1 [FREE Full text] [doi: 10.5860/choice.189890]

11.   Goldberg Y, Levy O. word2vec explained: deriving Mikolov et al's negative-sampling word-embedding method. ArXiv. Preprint posted online on February 15, 2014 2014:1 [FREE Full text]

12.   Patient-focused drug development: methods to identify what is important to patients guidance for industry. Food and Drug Administration. 2019. URL: https://www.fda.gov/media/131230/download [accessed 2021-10-08]

13.   NLM MeSH: chronic obstructive pulmonary disease. URL: https://www.ncbi.nlm.nih.gov/mesh?Db=mesh&Cmd=DetailsSearch&Term=%22Pulmonary+Disease,+Chronic+Obstructive%22%5BMeSH+Terms%5D [accessed 2021-10-08]

14.   NLM MedlinePlus: chronic obstructive pulmonary disease. URL: https://medlineplus.gov/copd.html [accessed 2021-10-08]

15.   Global Initiative for Chronic Obstructive Lung Disease (GOLD). URL: https://goldcopd.org/ [accessed 2021-10-08]

16.   Kessler R, Partridge MR, Miravitlles M, Cazzola M, Vogelmeier C, Leynaud D, et al. Symptom variability in patients with severe COPD: a pan-European cross-sectional study. Eur Respir J 2011 Feb;37(2):264-272 [FREE Full text] [doi: 10.1183/09031936.00051110] [Medline: 21115606]

17.   Miravitlles M, Worth H, Soler Cataluña JJ, Price D, De Benedetto F, Roche N, et al. Observational study to characterise 24-hour COPD symptoms and their relationship with patient-reported outcomes: results from the ASSESS study. Respir Res 2014 Oct 21;15:122 [FREE Full text] [doi: 10.1186/s12931-014-0122-1] [Medline: 25331383]

18.   Ubolnuar N, Tantisuwat A, Thaveeratitham P, Lertmaharit S, Kruapanich C, Chimpalee J, et al. Effects of pursed-lip breathing and forward trunk lean postures on total and compartmental lung volumes and ventilation in patients with mild to moderate chronic obstructive pulmonary disease: an observational study. Medicine (Baltimore) 2020 Dec 18;99(51):e23646 [FREE Full text] [doi: 10.1097/MD.0000000000023646] [Medline: 33371099]

19.   Elliott MW, Adams L, Cockcroft A, MacRae KD, Murphy K, Guz A. The language of breathlessness. Use of verbal descriptors by patients with cardiopulmonary disease. Am Rev Respir Dis 1991 Oct;144(4):826-832. [doi: 10.1164/ajrccm/144.4.826] [Medline: 1928956]

20.   Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. Science 2021 Jan 15;371(6526):284-288. [doi: 10.1126/science.abd7331] [Medline: 33446556]

21.   Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 2019 Jul 3;571(7763):95-98. [doi: 10.1038/s41586-019-1335-8] [Medline: 31270483]

22.   2021 GOLD reports: 2021 global strategy for prevention, diagnosis, and management of COPD. URL: https://goldcopd.org/wp-content/uploads/2020/11/GOLD-REPORT-2021-v1.1-25Nov20_WMV.pdf [accessed 2021-10-08]

23.   Barnes PJ, Burney PGJ, Silverman EK, Celli BR, Vestbo J, Wedzicha JA, et al. Chronic obstructive pulmonary disease. Nat Rev Dis Primers 2015 Dec 03;1:15076. [doi: 10.1038/nrdp.2015.76] [Medline: 27189863]

24.   Agustí A, Hogg JC. Update on the pathogenesis of chronic obstructive pulmonary disease. N Engl J Med 2019 Sep 26;381(13):1248-1256. [doi: 10.1056/nejmra1900475]

25.   Cook NS, Kostikas K, Gruenberger J, Shah B, Pathak P, Kaur VP, et al. Patients' perspectives on COPD: findings from a social media listening study. ERJ Open Res 2019 Feb 10;5(1):00128-02018 [FREE Full text] [doi: 10.1183/23120541.00128-2018] [Medline: 30775374]

26.   Patalano F, Gutzwiller FS, Shah B, Kumari C, Cook NS. Gathering structured patient insight to drive the PRO strategy in COPD: patient-centric drug development from theory to practice. Adv Ther 2020 Jan 09;37(1):17-26 [FREE Full text] [doi: 10.1007/s12325-019-01134-x] [Medline: 31707715]

27.   Ghosh A, Boucher RC, Tarran R. Airway hydration and COPD. Cell Mol Life Sci 2015 Oct;72(19):3637-3652 [FREE Full text] [doi: 10.1007/s00018-015-1946-7] [Medline: 26068443]

28.   Hewitt R, Farne H, Ritchie A, Luke E, Johnston SL, Mallia P. The role of viral infections in exacerbations of chronic obstructive pulmonary disease and asthma. Ther Adv Respir Dis 2016 Apr;10(2):158-174 [FREE Full text] [doi: 10.1177/1753465815618113] [Medline: 26611907]

29.   Gleibs IH. Turning virtual public spaces into laboratories: thoughts on conducting online field studies using social network sites. Anal Soc Iss Public Pol 2014 Jan 22;14(1):352-370. [doi: 10.1111/asap.12036]

30.   Golder S, Ahmed S, Norman G, Booth A. Attitudes toward the ethics of research using social media: a systematic review. J Med Internet Res 2017 Jun 06;19(6):e195 [FREE Full text] [doi: 10.2196/jmir.7082] [Medline: 28588006]

Abbreviations

COPD: chronic obstructive pulmonary disease
GOLD: Global Initiative for Chronic Obstructive Lung Disease
NLM: National Library of Medicine
PCA: principal component analysis

XSL•FO
RenderX