Table 2 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2017 in the section 'Education and Consumer Health Informatics'. The articles are listed in alphabetical order of the first author's surname.

Section Education and Consumer Health Informatics
▪ Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. "When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets. JMIR Public Health Surveill 2016 Oct 24;2(2):e162.
▪ Freedman RA, Viswanath K, Vaz-Luis I, Keating NL. Learning from social media: utilizing advanced data extraction techniques to understand barriers to breast cancer treatment. Breast Cancer Res Treat 2016 Jul;158(2):395-405.
▪ Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, McIver DJ, Rozenblum R, Wright A, Bourgeois FT, Greaves F. Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf 2016 Jun;25(6):404-13.
▪ Kondylakis H, Koumakis L, Hänold S, Nwankwo I, Forgó N, Marias K, Tsiknakis M, Graf N. Donor's support tool: Enabling informed secondary use of patient's biomaterial and personal data. Int J Med Inform 2017 Jan;97:282-92.
▪ Massey PM, Leader A, Yom-Tov E, Budenz A, Fisher K, Klassen AC. Applying Multiple Data Collection Tools to Quantify Human Papillomavirus Vaccine Communication on Twitter. J Med Internet Res 2016 Dec 5;18(12):e318.

TK. Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning. Addict Behav 2017 Feb;65:289-95.
5. Sarker A, O'Connor K, Ginn R, Scotch M, Smith K, Malone D, et al. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. Drug Saf 2016 Mar;39(3):231-40.
6. Rastegar-Mojarad M, Liu H, Nambisan P. Using Social Media Data to Identify Potential Candidates for Drug Repurposing: A Feasibility Study. JMIR Res Protoc 2016 Jun 16;5(2):e121.
7. Kavuluru R, Sabbir AK. Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on Twitter. J Biomed Inform 2016 Jun;61:19-26.
8. Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, Hanson CL. Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality. JMIR Ment Health 2016 May 16;3(2):e21.
9. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. "When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets. JMIR Public Health Surveill 2016 Oct 24;2(2):e162.
10. Freedman RA, Viswanath K, Vaz-Luis I, Keating NL. Learning from social media: utilizing advanced data extraction techniques to understand barriers to breast cancer treatment. Breast Cancer Res Treat 2016 Jul;158(2):395.
11. Massey PM, Leader A, Yom-Tov E, Budenz A, Fisher K, Klassen AC. Applying Multiple Data Collection Tools to Quantify Human Papillomavirus Vaccine Communication on Twitter. J Med Internet Res 2016 Dec 5;18(12):e318.
12. Nguyen QC, Li D, Meng HW, Kath S, Nsoesie E, Li F, et al. Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. JMIR Public Health Surveill 2016 Oct 17;2(2):e158.
13. Torii M, Tilak SS, Doan S, Zisook DS, Fan JW. Mining Health-Related Issues in Consumer Product Reviews by Using Scalable Text Analytics. Biomed Inform Insights 2016 Jun 20;8(Suppl 1):1-11.
14. Song J, Song TM, Seo DC, Jin JH. Data Mining of Web-Based Documents on Social Networking Sites That Included Suicide-Related Words Among Korean Adolescents. J Adolesc Health 2016 Dec;59(6):668-73.
15. Wilbanks JT, Topol EJ. Stop the privatization of health data. Nature 2016;535(7612):345-8.
16. Kondylakis H, Koumakis L, Hänold S, Nwankwo I, Forgó N, Marias Ket al. Donor's support tool: Enabling informed secondary use of patient's biomaterial and personal data. Int J Med Inform 2017 Jan;97:282-92.

Correspondence to:
Luis Fernandez-Luque, PhD
Qatar Computing Research Institute
Hamad Bin Khalifa University
Qatar Foundation
HBKU Research Complex, Doha, Qatar
E-mail: lluque@qf.org.qa

Appendix: Content Summaries of Selected Best Papers for the 2017 IMIA Yearbook in the Section "Education and Consumer Health Informatics"

Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A

"When 'Bad' is 'Good'": Identifying Personal Communication and Sentiment in Drug-Related Tweets

JMIR Public Health Surveill 2016 Oct 24;2(2):e162

Although several studies have reported on the development of automated approaches to analyse tobacco and e-cigarette-related tweet content, and to identify adverse effects associated with the medical use of pharmaceutical drugs, there have been very few attempts to apply automated content analysis techniques to analyse drug abuse–related tweets. This lack of research is partially related to the fact that drug-related content adds another layer of ambiguity and difficulty in the development of automated techniques because of the pervasive use of slang terminology and implied meanings. For the words that suggest a particular sentiment, traditional approaches that use sentiment lexicons may not perform well, and machine learning techniques, trained using manually coded data, could increase the accuracy of sentiment identification in drug-related tweets. The purpose of this study was to describe the development and performance of machine learning classifiers to automatically identify tweets by the type of communication (personal, official/media, or retail) and sentiment (positive, negative, or neutral) expressed in cannabis- and synthetic cannabinoid–related tweets. To reach a sample size of 4,000 tweets for the manually-labelled data set for machine learning, more than 8,000 tweets were manually reviewed and filtered using QDA Miner. The tweets for manual coding were extracted from the pool of 15,623,869 tweets that were collected by eDrugTrends between May and November 2015. The sample of 4,000 manually-labelled tweets was split into two subsamples, 1,000 were used to train a source classifier, and 3,000 were allocated for sentiment classification. The most discriminative unigram and bigram features that were identified by chi-square test reflect thematic groups as pertinent to sentiment categories: "want," "love," "need" for positive, in contrast to "don't," "shit," "fake" for negative tweets. But the sentiment classifier tended to incorrectly classify tweets that expressed an opposing opinion to negative thoughts or actions related to cannabis use or its legalization. Furthermore, humorous and sarcastic tweets were also more difficult to classify correctly by the classifier. The identification of sentiment in personal, user-generated tweets is more relevant for drug abuse epidemiology research

than an approach that includes media- and business-related tweets.

Freedman RA, Viswanath K, Vaz-Luis I, Keating NL

Learning from social media: utilizing advanced data extraction techniques to understand barriers to breast cancer treatment

Breast Cancer Res Treat 2016 Jul;158(2):395-405

To date, most studies examining barriers to care for diverse populations have been conducted within registry- or claims-based cohorts. Additional smaller studies using surveys, focus groups, and medical records are often limited to a single geographic area or institution and may not necessarily generalize across diverse populations. Furthermore, most surveys have structured formats and are subject to recall bias. Social media has been recognized as a potential source of patient data often underrepresented in studies using conventional research methodologies, emerging thus as a rich, yet largely untapped, resource for understanding what patients are candidly saying about their experiences and treatments. The purpose of this study was to utilize machine learning to identify key issues and themes that patients with breast cancer were sharing online, focusing on the barriers to treatment. Postings from a 365-day period, ending on January 31, 2015, on message boards, blogs, topical sites, content sharing sites, and social networks were examined. 3,200,128 unique posts that discussed breast cancer were identified. The analyses were limited to the 1,024,041 (32 %) posts about treatment. When possible, a phase of treatment (pre-diagnosis, diagnosis, assessment, decision to treat, or treatment) was identified by tagging posts based on cues for a user's current situation through topical keywords and relevant self-reported experiences yielding 627,381 posts. Among these posts, overarching themes and treatment barriers were assigned for 387,238 (62% of 627,381). Organizational barriers generally increased from pre-diagnosis (6% of posts) to diagnosis (13%) and remained high during assessment (28%), decisions to treat (21%), and treatment (29%). Sociocultural barriers decreased over the treatment trajectory (24% of posts in the pre-diagnosis phase to 18–20% of posts about treatments) as did psychological barriers (43% to 19–25%). Situational barriers remained relatively constant over the treatment trajectory and were reported in a quarter of posts. For emotional barriers, most conversations reported fears, anxiety, denial, and depression. The most common belief-related sentiments were spiritual/religious (41%), although other prominent themes included misinformation (30%) and preferences/perceptions (29%). The most common physical concerns expressed were side effects (40%), followed by physical limitations (31%) and body changes (29%). Resource concerns included posts about insurance (49%), costs (33%), and logistics of treatment (18%). Dominant concerns raised within posts about healthcare perception barriers included poor communication (36 %), trust (22%), accessibility of services (21%), and negative experiences (21%). Among posts related to relationship barriers, the most dominant issues included problems with intimacy (35%), friends (34%), and children (31%). Duration and process barriers were categorized as issues with the regimens prescribed (41%), duration of treatment (23%), effects of the after treatment (19%), and complexity of care (17%). In 9,465 posts, users suggested that they refused recommended treatments that were recommended for their breast cancer. With this new type of "social intelligence" for research, mining the vast repository of unstructured big data for insight into patients' concerns and experiences, the authors learned about barriers to care for a large and diverse population of users.

Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, McIver DJ, Rozenblum R, Wright A, Bourgeois FT, Greaves F

Measuring patient-perceived quality of care in US hospitals using Twitter

BMJ Qual Saf 2016 Jun;25(6):404-13

Experiences and perception of patients receiving healthcare as well as the necessity for healthcare stakeholders to measure and report outcomes are usually based on structured questionnaires. Limitations of these surveys include significant time lag between an outcome and a report of that outcome, and low response rates. As Twitter is actively used by one out of five adults, the authors sought to identify and analyse the content of posts sent to hospitals as a novel real-time measure of quality, supplementing traditional survey-based approaches. Hawkins, *et al.*, assessed the use of Twitter as a supplemental data stream for measuring patient-perceived quality of care in US hospitals and for comparing patient sentiments about hospitals with established quality measures. A machine learning approach was used to classify tweets associated with patient experiences.

Of the tweets directed to 2,349 US hospitals having an account on Twitter, over the period 1 October 2012 to 30 September 2013, 404,065 were analysed. Sentiment of patient experience was calculated for these tweets using natural language processing (the open source Python library TextBlob). A total of 11,602 tweets were manually categorised into patient experience topics, including food, money, pain, general, room condition, time, communication, discharge, medication instructions, side effects. Finally, 297 hospitals, representing 111 unique Twitter accounts with at least 50 patient-experience tweets were surveyed to understand how they use Twitter to interact with patients. The authors focused on the percentage of patients who rated a hospital at the highest levels on a validated scale of quality of care. The second validated measure of quality of care was the Hospital Compare 30-day hospital readmission rate calculated from the period 1 July 2012–30 June 2013 (https://www.medicare.gov/hospitalcompare/search.html). Roughly half of the hospitals in the US have a presence on Twitter (50.2%). Of the 297 surveyed hospitals, half responded and all confirmed that they closely monitor social media and interact with users. Of the tweets directed toward these hospitals, 34,725 (9.4%) were related to patient experiences, covering diverse topics. The top three topics discussed were: time management, money concerns, and communication with staff. Analyses limited to hospitals with at least 50 patient-experience tweets revealed that they were more active on Twitter, more likely to be below the national median of Medicare patients ($p<0.001$) and above the national

median for nurse/patient ratio (p=0.006), and to be a non-profit hospital (p<0.001). After adjusting for hospital characteristics, they found that Twitter sentiment was not associated with Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) ratings; however, having a Twitter account was associated with HCAHPS score, although there was a weak association with 30-day hospital readmission rates (p=0.003). The authors showed that monitoring Twitter provides useful, unsolicited, and real-time data that might not be captured by traditional feedback mechanisms. Tweets describing patient experiences in hospitals cover a wide range of patient care aspects and can be identified using automated approaches. The authors recommended that patients, researchers, and policy makers also attempt to utilise this data stream to understand the experiences of healthcare consumers and the quality of care they receive.

Kondylakis H, Koumakis L, Hänold S, Nwankwo I, Forgó N, Marias K, Tsiknakis M, Graf N

Donor's support tool: Enabling informed secondary use of patient's biomaterial and personal data

Int J Med Inform 2017 Jan;97:282-92

The purpose of this paper was to study the current practices for obtaining consent for biobanking and the legal requirements for reusing the available biomaterial and data in EU. The authors present a novel modular IT tool named "Donor's Support Tool" in order to ensure that patients actively provide and update their consent according to applicable national laws, thus enabling the secondary use of data and biomaterial. The legal landscape for the secondary use of biomaterial and data in the European Union is complex. There is no harmonized European regulation that covers both the processing and use of biosamples and associated personal or clinical data at the same time. Different regimes apply to each EU member. At present, the use of personal data enjoys the more harmonized framework. Informed consent is one of the best-known elements of medical ethics and bioethics, and is widely utilized in clinical practice and clinical research. But there are various types of consent: the consents that applies to a specific purpose or research study, the consent that is partially restricted to a domain of purposes or types of research studies, the consent that is multi-layered, wherein consent can apply to a number of unnamed or unspecified purposes or studies, or the broad consent which applies to any purpose or research study, named or unnamed. In clinical trials, only the specific consent is allowed, while different approaches, ranging from specific to broad, or even simply 'presumed' consent, could be applied in the processing of human tissues among EU member states. Similarly, for personal data processing, multiple approaches could possibly apply in the EU member states. The EU Clinical Trial Regulation (EU No 536/2014, https://ec.europa.eu/health/human-use/clinical-trials/regulation_en) requires that consent for a participation in a clinical trial be in a written form. National data protection laws usually require an explicit and written consent for the processing of sensitive data, except in the UK and Austria where no specific formal requirement has been set up. Regarding the identification and the authentication of the consent subject, even if a qualified electronic signature is desired, the usage of such signatures is not widespread among the European population. Transforming the legal requirements into information technology requirements, the authors designed and implemented the IT platform enabling citizens to actively provide and update their consent in real time. The three modules (personal information management system, donor's generation module, and donor's decision module) place participants at the heart of decision-making and allow individuals to tailor and manage their own consent preferences. Comparisons with six other relevant approaches are provided: SecureConsent, Mytrus, Educonsent, iMedConsent, FORCS e-consent and Consentir. The system was also tested by the University College of London using retrospective data.

Massey PM, Leader A, Yom-Tov E, Budenz A, Fisher K, Klassen AC

Applying Multiple Data Collection Tools to Quantify Human Papillomavirus Vaccine Communication on Twitter

J Med Internet Res 2016 Dec 5;18(12):e318

The purpose of this study was to quantify Human Papilloma Virus (HPV) vaccine communication on Twitter, specifically focusing on (1) sentiment, (2) side effects, and (3) prevention and protection, and to describe a novel methodology using two data collection methods to analyse Twitter data. Two methods were used to collect and validate Twitter data related to HPV vaccination. From August 1, 2014 to July 31, 2015, 305,517 and 258,102 tweets were collected respectively using a prospective or a retrospective data collection method. Only English-language tweets were included. A corpus of 1,470 manually coded tweets was used to develop a machine learning classifier for each variable in the codebook. Binary variables were classified using a linear classifier (Moore-Penrose pseudo-inverse), while a decision tree was applied to variables with more than two categorical responses. A total of 193,379 English-language tweets were collected, classified, and analysed between August 1, 2014 and July 31, 2015. Over 88.64% (191,515/216,060) of the final dataset included the keyword search term HPV, and nearly 34.91% (75,433/216,060) included HPV vaccine. Associated words varied with each keyword, with HPV being associated with personal words such as "I", "me", and "have", and #HPV being associated with January (cervical cancer awareness month), prevent, and learn. Positive sentiment toward the vaccine was the largest type of sentiment in the sample, with 75,393 positive tweets (38.99% of the sample). Many more users participated in positive sentiment than in negative sentiment (36,283 vs 24,010 users, respectively). There is also an important relationship between tweet sentiment and tweet content: many more tweets that were classified as positive mentioned information about prevention or protection, whereas tweets classified as negative included a much greater discussion about side effects. This can be important information for health promotion and communication campaigns, specifically in terms of tailoring a message and joining a particular conversation.