European Psychiatry 38 (2016) 15–21

Contents lists available at ScienceDirect

European Psychiatry

journal homepage: http://www.europsy-journal.com

Original article

Novel psychoactive substances: An investigation of temporal trends in social media and electronic health records

A. Kolliakou [a,*], M. Ball [a], L. Derczynski [b], D. Chandran [a], G. Gkotsis [a], P. Deluca [c], R. Jackson [d], H. Shetty [e], R. Stewart [a]

[a] Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[b] Department of Computer Science, University of Sheffield, Sheffield, UK
[c] National Addiction Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[d] Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[e] NIHR Biomedical Research Centre, South London and Maudsley NHS Trust, London, UK

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form 9 May 2016
Accepted 16 May 2016
Available online

Keywords:
Novel psychoactive substances
Mephedrone
Electronic health records
Social media
Public health monitoring

ABSTRACT

Background: Public health monitoring is commonly undertaken in social media but has never been combined with data analysis from electronic health records. This study aimed to investigate the relationship between the emergence of novel psychoactive substances (NPS) in social media and their appearance in a large mental health database.
Methods: Insufficient numbers of mentions of other NPS in case records meant that the study focused on mephedrone. Data were extracted on the number of mephedrone (i) references in the clinical record at the South London and Maudsley NHS Trust, London, UK, (ii) mentions in Twitter, (iii) related searches in Google and (iv) visits in Wikipedia. The characteristics of current mephedrone users in the clinical record were also established.
Results: Increased activity related to mephedrone searches in Google and visits in Wikipedia preceded a peak in mephedrone-related references in the clinical record followed by a spike in the other 3 data sources in early 2010, when mephedrone was assigned a 'class B' status. Features of current mephedrone users widely matched those from community studies.
Conclusions: Combined analysis of information from social media and data from mental health records may assist public health and clinical surveillance for certain substance-related events of interest. There exists potential for early warning systems for health-care practitioners.

1. Introduction

The need to interpret and act upon information from large-volume media such as Twitter is well recognised in business and politics, and increasingly appreciated in health research. For example, interactions on social media have enabled researchers to study health-related attitudes and behaviours in relation to tobacco smoking in Twitter [1] and Facebook [2], as well as identifying user social circles with common medical experiences [3], medical malpractice [4], HIV prevention [5] and pharmaco-vigilance [6] in Twitter. Wikipedia usage has also been utilised to estimate the prevalence of influenza-like illness in the United States in near real-time [7]. Similarly, the size, coverage and longitudinal nature of electronic health records (EHR), as well as their potential for data linkage offer unprecedented opportunities for big data analytics. Healthcare is thus emerging as part of a worldwide network of developing technologies [8] and with the arrival of Web 2.0, the relationship between patients and healthcare providers is rapidly changing. Web 2.0 [9] is a term popularised in 2004 and marks the beginning of a new wave of online activity–one that moves away from passive viewing of content and emphasizes interaction between users as creators of the very content that is being accessed. Communication now transcend geographical, cultural and language barriers to allow the exchange of information before it reaches the clinical room and a suspected flu outbreak could be trending on Twitter within hours, long before specialists have had an opportunity to properly examine it [10].

* Corresponding author. Biomedical Research Centre Nucleus, PO92, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, London SE5 8AF, UK. Tel.: +00 44 0 20 32 28 85 61.
E-mail address: anna.kolliakou@kcl.ac.uk (A. Kolliakou).

http://dx.doi.org/10.1016/j.eurpsy.2016.05.006
0924-9338/

The number of novel psychoactive substances (NPS), commonly referred to as 'Legal highs*', has been growing steadily in the last couple of decades with 100 not-previously-reported substances identified, in Europe, in 2015 [11]. The Internet is a primary source of information about NPS and the rapid rate with which they appear, as well as the uncertainty over the actual 'branding' and composition of these substances, pose substantial challenges for healthcare providers. Online user-generated content is increasingly becoming essential to providing an informed and up-to-minute portrayal of positive and negative effects, subjective experiences and availability of NPS [12]. As part of the PHEME project (www.pheme.eu), whose wider aim is to explore social media veracity and rumours, we investigated the temporal relationship between the emergence of NPS in social media and their appearance in a large mental health EHR covering an inner urban catchment area.

2. Method

2.1. Electronic health record data resource

Mental healthcare data were collected using the South London and Maudsley (SLaM) Biomedical Research Centre (BRC) Case Register [13]. The SLaM NHS Foundation Trust provides comprehensive mental health services to a geographic catchment area of over 1.2 million residents in four south London boroughs, making it one of the largest mental healthcare organisations in Europe. A single electronic health record has been used by all SLaM teams since April 2006. The Clinical Record Interactive Search (CRIS) application, developed in 2007–2008, extracts anonymised data from structured fields as well as unstructured free text from case notes and correspondence [13], which are particularly valuable in mental health research. The free text fields are used by health-care professionals to record clinical information over the course of care ranging from diagnoses and mental state examinations to daily nursing entries and treatment plans. CRIS contains over 250,000 de-identified patient records, including over 20 million text documents (growing at a rate of 170,000 per month) and has supported a number of studies [14–19]. The SLaM Case Register has ethical approval as a database for secondary analysis (Oxford REC C, reference 08/H0606/71 + 5) and a service-user led oversight committee provides governance for projects utilising these data [20].

To ascertain references to NPS in the clinical records, we searched case note, correspondence and discharge summary text fields for the following keywords: spice, methoxetamine, AMT, Benzo Fury, Piperazines (BZP, TFMPP, DBZP and mCPP), mephedrone, 2-DPMP, Salvia divinorum, morning glory, 2C-B, MDAI, MDPV, bromodragonfly, kanna, 4-Acetoxy-Met, naphyrone and 'Legal high*'. A list of the related terms commonly associated with these substances was also produced for searches. Table 1 shows the full list of search terms together with the number of retrieved documents containing the search term, the number of retrieved documents checked for actual mention of NPS and the number of true references (mentions truly related to the term of interest; e.g., kitty cat as referred to mephedrone and not to the animal) within the checked documents. Due to the low frequency at which NPS were mentioned in the clinical record, it was considered impractical to explore all associations with social media mentions and subsequent analyses focused solely on mephedrone, the most commonly referenced agent. Mephedrone [4methylmethcathinone, or 1-(4-methylphenyl)-2-methylaminopropan-1-one] is a phenethylamine and cathinone derivative, which typically mimics stimulant effects produced by amphetamines such as cocaine. It is widely available to purchase over the Internet usually in the form of white crystalline powder and is most commonly ingested intranasally or orally [21].

2.2. Twitter data resource

Twitter is a micro-blogging platform, which as of the second quarter of 2015, averaged 304 million monthly active users [22]. With 500 million tweets on a typical day (5700 per second) and a wealth of text, graph, image and video interaction, Twitter is one of the largest social media sources [23]. For our study, we accessed tweets archived from a Twitter feed licensed to the University of Sheffield from July 2009 to September 2014 inclusive. These comprise a random 10% sample of all tweets [24] and are kept in hourly or daily files. The sample was searched for terms related to mephedrone by using Aho-Corasick [25] search first to

Table 1
Search list for 'Legal highs*' and related terms in the clinical record.

Keywords	Related terms	Documents retrieved/ documents checked	True references
Legal high*	Plant food, mdat, eric 3, dimethocaine, bath salts	173/173	143
Spice	Spice silver, spice gold, spice diamond, bliss, blaze, genie, zohai, jwh-018, jwh-073, jwh-250, yucatan fire, moon rocks, k2, red x dawn, fake weed, x, tai high hawaiian haze, spice, mary joy, exodus damnation, ecsess, devil's weed, clockwork orange, bombay blue extreme, blue cheese, black mamba, annihilation, amsterdam gold	639/300	12
Mephedrone	Subcoca-1, 4-mmc, kitty cat, miaow miaow, meow meow, m-smack, m-cat	491/250	213
Methoxetamine	M-ket, mex, kmax, special m, ma, legal ketamine, minx, jipper, kwasqik, hypnotic, panoramix, magic, lotus, roflcoptr, rhino ket, mxe, moxy, mket, mexy, mexxy	27/27	2
AMT	2-(1h-indol-3-yl)-1-methyl-ethylamine, indopan, amt freebase, alpha-methyltryptamine	1/1	1
Benzo Fury	White pearl, benzo fury, 6-apdb, 6-apb, 5-apdb, 5-apb, apb	17/17	12
Piperazines (BZP, TFMPP, DBZP and mCPP)	The good stuff, smileys, silver bullet, rapture, pep twisted, pep stoned, pep love, pep, party pills, nemesis, legal x, legal e, happy pills, frenzy, fast lane, exodus, euphoria, esp, cosmic kelly, bzp, bolts extra strength, blast, benzylpiperazine, a2	49/49	3
2-DPMP	Vanilla sky, purple wave, ivory wave, desoxypipradrol, d2pm, 2-diphenylmethylpyrrolidine	0/0	0
Salvia divinorum	Mexican magic mint, holy sage, eclipse, salvinorin a	86/40	15
Morning glory	Pearly gates	27/27	2
2C-B, 2C-T-7	Nexus	23/23	1
MDAI	–	0/0	0
MDPV	–	5/5	5
Bromodragonfly	–	0/0	0
Kanna	Sceletium tortuosum, mesembrine	0/0	0
4-AcO-Met	4-acetoxy-met, metacetin	0/0	0
Naphyrone	Nrg	38/38	8

reduce the number of records processed in detail. The remaining records were then de-serialised, language identification used where available to filter out non-English tweets, and the terms were searched for in just the tweet text field. All the above were performed on a Sun Grid Engine cluster. The resulting data were read in Microsoft Excel format as in one tweet per line. Annotating of the tweets was performed in two phases. First, a proportion of the extracted tweets were manually double-annotated based on whether they were a true, false or unclear reference to mephedrone. An inter-annotator agreement analysis using the Kappa statistic was performed to determine consistency between the two annotators and a collective resolution of the disagreements from the first round of annotating was implemented to reach 100% agreement.

To develop an automatic application for identifying genuine mentions of mephedrone in the remaining tweets, a natural language processing (NLP) approach was taken. This involved applying an algorithm that was able to determine if the semantic meaning of the text was a reference to mephedrone or not (e.g. kitty cat as a reference to mephedrone and not to the animal). The algorithm was developed using a subset of the tweets already annotated. These tweets were analysed and the linguistic patterns that indicated a true reference to mephedrone were determined, which were then used to create identification rules implemented on General Architecture for Text Engineering software (GATE) [26,27]. GATE also supported the rapid deployment of these applications over the larger set of tweets retrieved. Rules were tested over another 'gold standard' subset of tweets already annotated.

The mephedrone-related Twitter query produced 145,578 tweets of which 5000 were double-annotated (available at: https://figshare.com/articles/Mephedrone_annotations_for_Twitter/1613832) with an inter-annotator agreement kappa statistic of 0.861 (95% CI 0.846, 0.877). Following arbitration of disagreements, the final annotations were as follows: 903 positive, 3932 negative and 165 unclear. The NLP algorithm was developed through the use of 2400 annotated tweets (training set). The rules created to identify the linguistic patterns indicating a positive reference to mephedrone were then tested on another 2400 annotated tweets (gold standard set) using GATE. The development of the GATE application was successful in identifying true instances of mephedrone in the tweets with a precision score (positive predictive value) of 0.988 and a recall score (sensitivity) of 0.896. The application (available at: http://dx.doi.org/10.5281/zenodo.35376) was then deployed over the complete dataset of 145,578 tweets retrieved between 2009 and 2014–7044 were identified as true instances of mephedrone.

2.3. Google Trends data resource

Google Trends [28] is a public web facility that can show users how often a particular term has been entered on Google Search relative to the maximum volume of searches over a certain time-period. It also has a useful function whereby it can show news stories related to the search-term overlaid on the results chart and can potentially demonstrate how events occurring worldwide affect search-term popularity. We used the Google Trends interface to search for mephedrone-related terms (Table 1) and calculate the relative number of times mephedrone was entered as a keyword in Google search.

2.4. Wikipedia data resource

Wikipedia, launched in 2001, is a multilingual, open-access, web-based library based on a model of editable content provided by anonymous volunteers [29]. It consists of over 37 million pages and attracts around 374 million unique visitors every month, making it one of the largest reference websites and surpassing all other encyclopaedias in both size and coverage. We collected and analysed page view statistics [30] for the English Wikipedia between 10/12/2007 and 01/10/2014 and the page 'mephedrone' [31]. After processing, a report was generated with regards to the number of visits on a daily basis for the target entry-page, from 10/10/2008 until 01/10/2014.

2.5. Further analysis

In order to compare temporal trends in mephedrone mentions on Twitter, Google Trends, Wikipedia and the EHR, we performed a two-fold normalisation of the data. Firstly, there had occasionally been technical issues resulting in less than 10% of all tweets being captured each month. Firstly, there had occasionally been technical issues resulting in less than 10% of all tweets being captured each month. These could be due to choked data–glitches in the agreed streaming rate, which sometimes caused the data to drop from a 10% of global tweets to a 1% feed. This was a general effect at data collection level and could be isolated and normalised for by increasing estimates 10-fold during the given period. Another reason could be missing data due to issues such as reboots and network disconnections which led to some hourly or even daily parts of the data being either missed or irretrievable. We worked to recover data that gave errors, discarding broken records and measuring the rate of known good data collection. This was achieved by first identifying what the problem was (missing or choked data) and, then, based on the start and end times of the problems, determining that either 0 or 10% of the expected amount had been captured during the given interval. In turn, this allowed us to calculate the effective observed rate per hour, per day and per month. Accordingly, knowing what the actual collected amount was for a period, we could scale observations made in that period. The sampling strategy on all feeds is fair, as we know from Kergl et al. [24]; and so while this method introduces some uncertainty, the sample bias remains of the same nature. To control for the effect this may have had on the true number of mephedrone mentions in Twitter, we divided the number of positive mephedrone tweets by the number of total tweets captured for each month. Secondly, since data were measured on different scales for the three data sources, we normalised values between 0 and 1 (representing the maximum and minimum frequencies within the observation period) to allow comparisons on a common scale. Data were then overlaid to generate a graph showing the relative timing of references to mephedrone in the clinical record, mephedrone mentions in Twitter and mephedrone-related searches in Google and Wikipedia.

Finding ourselves in a unique position to conduct a post-hoc analysis specific to mephedrone references, we sought to establish the wider context within which mephedrone appeared in the clinical record and, in particular, identify the characteristics specific to the case notes where current mephedrone use was reported. TextHunter [32], an NLP software tool for extracting generic concepts from free text in clinical research, was used to extract information on mephedrone mentions in the clinical record. All terms related to mephedrone (Table 1) were searched for and all sentences containing these terms were annotated as a positive (i.e. patient reports mephedrone use), negative (i.e. patients reports no mephedrone use) or irrelevant (i.e., patient reports mephedrone use by a fried) reference to mephedrone use. All positive references were further annotated based on present or past mention of mephedrone use. As there could be multiple references to mephedrone in one patient record, the document where use was first reported was selected for each patient. Data on demographic status (age, gender and ethnicity), primary diagnosis

and primary service contact for all such patients aged 18-65 were utilised. Date of birth and type of primary service contact were extracted at the time-point closest to the first mention of current mephedrone use and most recent primary diagnosis was obtained from records between 1st January 2007 and 31st September 2014. All data were analysed using Stata V.13 [33].

3. Results

Mapping of mephedrone presence in the four information sources is shown in Figs. 1 and 2. In the first figure, all sources show peaks in mephedrone mentions in early 2010 shortly before its reclassification to a 'class B' drug under UK law in April 2010, followed by virtual disappearance on Twitter, longer-term low-grade activity on Google Trends and a steady rise in mentions in the mental health record. A steady decrease was noted in Wikipedia visits, with the exception of spikes associated with two news-related events in June 2012 and January 2014. Pre-2010, activity was evident on Google Trends and Wikipedia but not substantial in the clinical record or Twitter. Fig. 2 provides a more detailed representation of the emerging mentions during the 01/01/2009-01/08/2010 period. A rise in activity on Google Trends and Wikipedia was the first observation, preceding any rise in Twitter activity by 3–4 months. Although mephedrone mentions on Twitter, Google Trends and Wikipedia simultaneously reached their peak in the first quarter of 2010, this was marginally preceded by a spike in mephedrone references in the clinical record. Activity sharply declines in all four sources by the middle of the year.

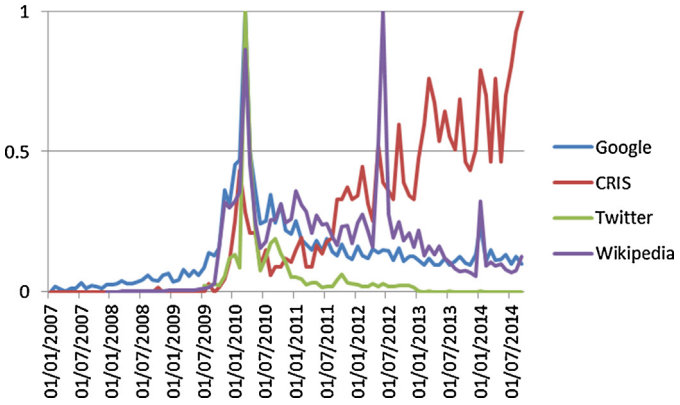

Fig. 1. Absolute values at 1: Google trends – 290, CRIS – 67, Twitter – 1319, Wikipedia – 355,789.

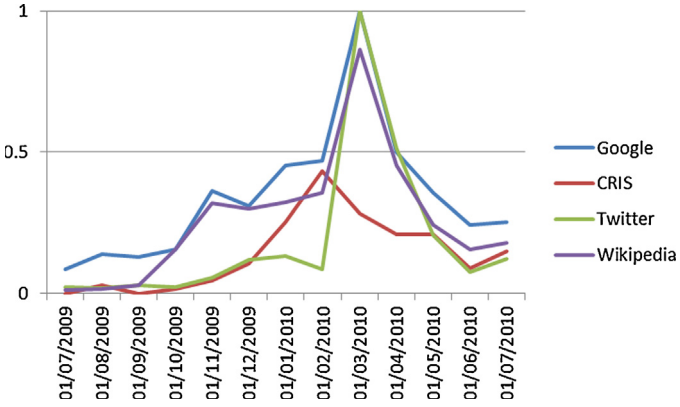

Fig. 2.

We retrieved 2799 sentences containing mephedrone-related terms in the clinical record database. Of these, 2578 were annotated as a positive occurrence of the drug and current use was implied in 2187 of these 2578 sentences. Following restriction of documents to the first mention of current mephedrone use for each patient, 468 records were returned. Median age in this case sample was 30 (IQR = 18–55), 84.0% were male, and 65.2% were from a white ethnic background (black 11.7%; other 12.4%; not stated 10.7%). Primary ICD-10 diagnoses at the time of mephedrone use being first recorded were as follows:

- 40.0% disorder due to drug use;
- 7.3% depressive disorder;
- 7.0% 'other disorder';
- 7.0% personality disorder;
- 6.2% anxiety disorder;
- 6.0% developmental disorder;
- 4.9% disorder to due to alcohol use;
- 4.9% psychotic disorder;
- 3.6% schizopreniform disorder;
- 2.6% bipolar disorder;
- 0.9% eating disorder;
- 9.6% (not available).

Finally, 28.8% were under the care of addiction services at the time of first mention of mephedrone use (18.4% liaison psychiatry; 12.0% A&E; 12.0% general mental health services; 5.8% psychosis services; 3.2% HIV services; 9.0% other; 10.9% not stated).

4. Discussion

We initially sought to investigate the temporal relationship between the appearance of a range of NPS in social media and their occurrence in a mental health EHR representing an urban catchment area. However, a key initial finding was that there were generally very few references to NPS in the clinical record. Although we cannot infer prevalence in a clinical sample relying on recorded mentions of use, our findings from rates of NPS reference in the clinical record mirror those of studies reporting low rates of NPS use among participants in night-time economy (NTE) in south London [34] as well as New York City [35] and in two online populations [36,37]. Despite the growing number of NPS identified by early warning systems, European reports also suggest that lifetime prevalence of NPS use remains low in most countries [11]. Mephedrone was the most commonly referenced NPS in the EHR and, similarly, the preference for its use over other NPS has been documented by the Crime Survey for England and Wales (CSEW) [38], in respondents from NTE in London and Lancashire [39,40] and in regular clubbers [41]. A steady decline in mephedrone use has been reported by the CSEW with last year use of mephedrone among adults aged 16–59 being 1.3% in 2010/11, then falling to 1% in 2011/12 and to 0.5% in 2012/13 between before stabilising at 0.6% in 2013/2014 [42]. In our study, however, we observed a steady increase in the number of mephedrone mentions in the EHR over the last 3 years. As concern about NPS grows, it is only natural that clinical practice will transform to keep pace with these rapid developments in public and policy approach and healthcare professionals might therefore be more inclined to ask about mephedrone use and record its presence or absence. In addition, experience of harm or dependence from mephedrone use may be causing an increased number of presentations to specialist national health services.

Although some research exists on NPS use in mental health service users, to the best of our knowledge, ours is the first study to report on the profile of recorded current users in a secondary

mental healthcare setting. Demographic characteristics were largely compatible in distribution with previous findings from general population surveys. Users were primarily male [35,42–44] and the median age was 30; slightly higher than the mid-to-late 20s reported in general population samples [35,36,44,45]. This could be explained by the prominence of young clubbers in the surveys as well as the exclusion in our study of mephedrone users younger than 18, which will have raised the lower end of the age range. Lastly, the majority of our sample was, similar to other research, of white ethnic background [35,36]. Previous investigations of NPS use in people with mental disorders have suggested high levels of use by those diagnosed with a psychotic or bipolar disorder [46,47]. Since we extracted the primary diagnosis closest to the time of recorded use, diagnostic instability for mental disorders [48] could account for the majority of our sample having been given a diagnosis of a disorder due to drug use. The progression of illness, emergence of new information and variability in diagnostic instruments [48] might have also contributed to this discrepancy.

Exploring the temporal relationships between the four different data sources with regard to mephedrone occurrences, we found a similarity in trends between mentions of mephedrone in Twitter, mephedrone-related searches on Google Trends and visits in Wikipedia and references to mephedrone in the mental healthcare record. Notably, markedly increased mentions in all four data sources occurred around the time mephedrone was designated a 'class B' drug in the UK in April 2010 indicating the influence of key events and news stories in driving public interest as well as clinical attention. This corresponds to an increase accesses to TOXBASE, the UK's online service run by the National Poison Information Service (NPIS), around February/March 2010 related predominantly to mephedrone. Since that peak, mephedrone accesses have reduced but remain relatively high to other NPS [42]. As Measham et al. [49] further discuss, this also illustrates the rapid rise in popularity and availability of mephedrone from the summer of 2009 in the United Kingdom. In similar health-related research, Duh [50] also reported an overall trend across AskaPatient.com and Google Trends on the frequency and pattern of adverse events for atorvastatin with peaks during popular media coverage demonstrating the resemblance in trends between Google Trends and other online platforms. Our study has taken this comparison a step further to show that social media information-seeking and information-sharing behaviour and data from clinical communication platforms were largely comparable. Although the implications of this observation are not yet clear, an 'infoveillance' [51] approach to substance use appears at least theoretically promising for gathering information on levels of internet activity–particularly with respect to emerging drugs whose low rates of use would normally require a very large sample to identify their emergence and where relevant reports are delayed by traditional methods of data collection. Social media analysis in healthcare monitoring is in itself promising and continues to gather momentum; however, combined analysis can only be achieved when rates of reference in the EHR are sufficiently high to support meaningful comparisons. Murphy et al. [52] conducted a study mining Twitter and Google Trends data for *Salvia divinorum* (an NPS with hallucinogenic properties) and compared these to general population use data. Although they found that information sharing about *S. divinorum* on Twitter may be associated with actual self-reported use in the general population, they failed to observe a consistent trend between the three data sources. They argued that the rarity of *S. divinorum* use, unwillingness to discuss or search for information related to illegal behaviours, as well as the slow diffusion rate of emerging drugs of abuse may have influenced the results. Inarguably, what both our and their findings emphasise is that using Twitter data to explore behavioural trends might not be suitable for all kinds of health issues; particularly those which are frequently undisclosed [52]. However, research conducted in online forums where users can reveal and exchange drug-using information without concerns of identity exposure, has achieved great methodological success in gathering a wealth of data on the growing market of NPS [12]. Research in either source encounters the caveats of unregulated, user-generated content which can be inaccurate or misleading but also has the potential to advise general and clinical knowledge and education and training, particularly in the absence of traditional sources of information for rare substances [12].

Although our observations were exploratory, it is noteworthy that the largest spike of activity related to mephedrone in the online sources, in mid-2010, did not precede but actually succeeded a substantial spike in mephedrone mentions in the clinical record in the beginning of 2010. It is intuitive to assume that social media analysis has been mostly successful because it relates trends to the general population of which online users are regarded as largely representative; that is, it informs events in the group that provided the data in the first place. It is not known to what extent mental health service users utilise online sources and so combining social media analysis with clinical record data might not follow similar patterns. Also, we concentrated on social media and particularly Twitter, as a representative medium of trending news and events. A key event, such as mephedrone's reclassification, will have been circulating on other mainstream media to which patients and healthcare professionals might have already been exposed. This is a good example of the well-known model of social media whereby it expresses real-world discussions and events as a latent variable; that is, it reports a mediated version of real-world activities.

A strength of this study was the utilisation of 'big data' both from online media and mental health records. Twitter and Google Trends are widely established as sources for gaining understanding in a variety of public health topics [50] and the English Wikipedia has emerged as an important source of online health information in relation to other providers [53]. They offer an abundance of information that can be extracted, categorised and analysed for health research, providing an insight into online attitudes and behaviours that may replicate general population and clinical trends. Another strength was the successful implementation of NLP tools to reliably ascertain true references of mephedrone in Twitter, which limits potential bias arising from manual identification. Clinical data were collected from a large mental healthcare provider rather than from participants selected specifically for NPS research, thus maximising the generalisability of our results. It is important to remember, however, that EHR data were derived from a specific urban area of south London and we did not attempt to use geo-location parameters to limit the extraction of social media data; mephedrone trends indeed appeared to transcend geographical restrictions. Data quality from clinical records will also be influenced by the accuracy and timeliness with which clinicians ask and record information as well as the information itself communicated by service users and this might have accounted at least in part for the very limited information on other NPS. A representative, random sample of Twitter data were analysed for this project, although caution has to be exercised when inferring population-level attitudes and opinions from this source [54]. Around 23% of online adults are estimated currently to use Twitter but the service has seen a significant increase since 2013 in certain demographic groups: men, those aged 65 and older and those from white ethnic backgrounds [55], resulting in concerns over the level of public identity that is represented in this medium. A further limitation inherent in drug use research is reporting bias arising from people's lack of knowledge of what substances they are actually using, particularly since NPS are

closely associated with online purchasing and high levels of assertive marketing through unregulated websites [56]. Finally, it is important to bear in mind that aggregated social media and patient data were used for these analyses and no attempt was made to link specific social media accounts to specific clinical records. Thus, it is not possible to conclude from coinciding trends on social media and the clinical record whether these reflected awareness or not of the social media activity by those patients in whom NPS use was recorded.

Our study has demonstrated the potential for combined analysis of information from online media and data from mental health records in exploring the temporal relationship of NPS emergence. From the point of view of public health monitoring and as a promising tool for early warning systems for health practitioners, information-seeking behaviour in sources such as Google and Wikipedia may forecast increased online activity and clinical interest for emergent NPS. In addition, steady increase in online activity does in the mid-term precede spikes in particular term occurrence in the clinical record. Peaks and troughs in online chatter are heavily influenced by trending news and events putting clinicians at an advantage to collect extensive and useful information from a number of media sources in a timely manner. There is tremendous capacity for progress in infoveillance for public health monitoring and the first steps towards an automatic warning system for health services are well underway.

Disclosure of interest

The authors declare that they have no competing interest.

Acknowledgements

This research was partially funded by the European Union/EU under the Information and Communication Technologies (ICT) theme of the 7th Framework Programme for R&D (FP7), grant PHEME (611233). The NIHR Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust and King's College London provided core support in the development and delivery of the study.

References

[1] Myslin M, Zhu SH, Chapman W, Conway M. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013;15:e174. http://dx.doi.org/10.2196/jmir.2534.
[2] Struik LL, Baskerville NB. The role of Facebook in Crush the Crave, a mobile- and social media-based smoking cessation intervention: qualitative framework analysis of posts. J Med Internet Res 2014;16(7):e170. http://dx.doi.org/10.2196/jmir.3189.
[3] Hanson CL, Cannon B, Burton S, Giraud-Carrier C. An exploration of social circles and prescription drug abuse through Twitter. J Med Internet Res 2013;15(9):e189. http://dx.doi.org/10.2196/jmir.2741.
[4] Nakhasi RJ, Passarella SG, Bell MJ, Paul M, Dredze PJ. Provost, Malpractice and Malcontent: Analysing Medical Complaints in Twitter, AAAI Technical Report. Information Retrieval and Knowledge Discovery in Biomedical Text. Johns Hopkins University; 2012.
[5] Young SD, Jaganath D. Online social networking for HIV education and prevention: a mixed-methods analysis. Sex Transm Dis 2013;40(2):162–7.
[6] Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform 2015;54:202–12.
[7] McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol 2014;10(4):e1003581. http://dx.doi.org/10.1371/journal.pcbi.1003581.
[8] Ohno-Machado L. Focusing on the patient: mHealth, social media, electronic health records, and decision support systems. J Am Med Inform Assoc 2014;21(6):953.
[9] O'Reilly T. What is Web 2. 0 - Design patterns and business models for the next generation of software; 2005, http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html.
[10] Li J, Cardie C. Early stage influenza detection from Twitter; 2013 [ar Xiv: 1309.7340v3].
[11] EMCDDA. 2016 EU Drug Markets Report: In-depth Analysis. European Monitoring Centre for Drugs and Drug Addiction; 2015, http://www.emcdda.europa.eu/publications/eu-drug-markets/2016/in-depth-analysis.
[12] Davey Z, Schifano F, Corazza O, Deluca P, on behalf of the Psychonaut Web Mapping Group e-Psychonauts. Conducting research in online drug forum communities. J Ment Health 2012;21:386–94.
[13] Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry 2009;9:51. http://dx.doi.org/10.1186/1471-244X-9-51.
[14] Chang CK, Hayes R, Broadbent M, Fernandes AC, Lee W, Hotopf M, et al. All-cause mortality among people with serious mental illness (SMI), substance use disorders, and depressive disorders in southeast London: a cohort study. BMC Psychiatry 2010;10:77. http://dx.doi.org/10.1186/1471-244X-10-77.
[15] Hayes RD, Chang CK, Fernandes A, Broadbent M, Lee W, Hotopf M, et al. Associations between substance use disorder sub-groups, life expectancy and all-cause mortality in a large British specialist mental healthcare service. Drug Alcohol Depend 2011;118:56–61.
[16] Hayes RD, Chang CK, Fernandes AC, Begum A, To D, Broadbent M, et al. Functional status and all-cause mortality in serious mental illness. PLoS One 2012;7:e44613. http://dx.doi.org/10.1371/journal.pone.0044613.
[17] Wu CY, Chang CK, Hayes RD, Broadbent M, Hotopf M, Stewart R, et al. Clinical risk assessment rating and all-cause mortality in secondary mental healthcare: the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) Case Register. Psychol Med 2012;42:1581–90.
[18] Wu CY, Chang CK, Robson D, Chen S-J, Hayes RD, Stewart R. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. PLoS One 2013;8:e74262. http://dx.doi.org/10.1371/journal.pone.0074262.
[19] Patel R, Jayatilleke N, Jackson R, Stewart R, Mcguire P. Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach. Lancet 2014;383:S16. http://dx.doi.org/10.1016/S0140-6736(14)60279-8.
[20] Fernandes AC, Cloete D, Broadbent MTM, Hayes RD, Chang C-K, Roberts A, et al. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. BMC Med Inform Decis Mak 2013;13:71.
[21] Winstock AR, Mitcheson LR, Deluca P, Davey Z, Corazza O, Schifano S. Mephedrone, new kid for the chop? Addiction 2010;106:154–61.
[22] Statista. Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2015 (in millions). The Statistics Portal; 2015, http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.
[23] Twitter. New Tweets per second record, and how! Twitter Engineering; 2013, https://blog.twitter.com/2013/new-tweets-per-second-record-and-how.
[24] Kergl D, Roedler R, Seeber S. On the endogenesis of Twitter's Spritzer. Gardenhose sample streams. Proc Adv Soc Networks Analysis Mining 2014;1:357–64. http://dx.doi.org/10.1109/ASONA.M.2014.6921610.
[25] Aho AV, Corasick MJ. Efficient string matching: an aid to bibliographic search. Commun ACM 1975;18(6):333–40.
[26] Cunningham H, Maynard D, Bontcheva K, on behalf of the GATE group. Text processing with GATE (Version 6). University of Sheffield; 2011.
[27] Bontcheva K, Derczynski L, Funk A, Greenwood MA, Maynard D, Aswani N, et al. In: An Open-Source Information Extraction Pipeline for Microblog Text; 2013.
[28] Google. Insights into what the world is searching for–the new Google Trends, Google Inc.; 2012.
[29] Wikipedia; 2015, https://en.wikipedia.org/wiki/Wikipedia:About.
[30] Wikipedia; 2015, http://stats.grok.se/.
[31] Wikipedia. Mephedrone; 2015, https://en.wikipedia.org/wiki/Mephedrone.
[32] Jackson R, Ball M, Patel R, Hayes RD, Dobson RJB, Stewart R. TextHunter—a user friendly tool for extracting generic concepts from free text in clinical research. Proc Am Med Informatics Assoc 2014;1:729–38. http://dx.doi.org/10.13140/2.1.3722.9121.
[33] Statacorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP; 2011.
[34] Wood DM, Hunter L, Measham F, Dargan PI. Limited use of novel psychoactive substances in South London nightclubs. Q J Med 2012;105:959–64.
[35] Kelly BC, Wells BE, Pawson, Leclair A, Parsons JT, Golub SA. Novel psychoactive drug use among younger adults involved in US nightlife scene. Drug Alcohol Rev 2013;32:588–93.
[36] Johnson PS, Johnson MW. Investigation of "Bath Salts" use patterns within an online sample of users in the United States. J Psychoactive Drugs 2014;46(5):369–78.
[37] Global Drug Survey. The Global Drug Survey 2015 findings; 2015, http://www.globaldrugsurvey.com/the-global-drug-survey-2015-findings/.
[38] Home, Office. Drug misuse declared finding from the 2011 to 2012 Crime Survey for England and Wales (CSEW), 2nd ed., London: Home Office; 2012, https://www.gov.uk/government/statistics/drug-misuse-declared-findings-from-the-2011-to-2012-crime-survey-for-england-and-wales-csew-second-edition.
[39] Measham D, Wood DM, Dargan PI, Moore K. The rise in legal highs: prevalence and patterns in the use of illegal drugs and first- and second-generation "legal highs" in South London gay dance clubs. J Subst Use 2011;16(4):263–72.
[40] Measham F, Moore K, Østergaard J. Emerging Drug Trends in Lancashire: Night Time Economy Surveys. Phase One Report. Lancaster University/LDAAT; 2011.
[41] Mixmag. Global Drug Survey; 2011, http://issuu.com/mixmagfashion/docs/drugsurvey.

[42] Home, Office. Drug misuse: findings from the 2013/14 Crime Survey for England and Wales. Home Office; 2014, https://www.gov.uk/government/publications/drug-misuse-findings-from-the-2013-to-2014-csew/drug-misuse-findings-from-the-201314-crime-survey-for-england-and-wales.
[43] Mixmag. Global Drug Survey; 2012, http://issuu.com/mixmagfashion/docs/drugs_survey_2012_2.
[44] Carhart-Harris RL, King LA, Nutt DJ. A web-based survey on mephedrone. Drug Alcohol Depend 2011;118(1):19–22.
[45] Winstock A, Mitcheson L, Ramsey J, Davies S, Puchnarewicz M, Marsden J. Mephedrone: use, subjective effects and health risks. Addiction 2011;106(11):1991–6.
[46] Lally J, Higaya E-E, Nisar Z, Bainbridge E, Hallahan B. Prevalence study of head shop drug usage in mental health services. Psychiatrist Online 2013;37:44–8.
[47] Martinotti G, Lupi M, Acciavatti T, Cinosi E, Santacroce R, Signorelli MS, et al. Novel psychoactive substances in young adults with and without psychiatric comorbidities. Biomed Res Intern 2014;1:1–7.
[48] Baca-Garcia E, Perez-Rodriguez MM, Basurte-Villamor I, Fernandez del Moral AL, Jimenez-Arriero MA, Gonzalez de Rivera JL, et al. Diagnostic stability of psychiatric disorders in clinical practice. Br J Psychiatry 2007;190:210–6.
[49] Measham F, Moore K, Newcombe R, Welch Z. Tweaking, bombing, dabbing and stockpiling: the emergence of mephedrone and the perversity of prohibition. Drugs Alcohol Today 2010;10(1):14–21.
[50] Duh MS. Monitoring online signals: implications of internet data in pharmacovigilance. Health Care Bulletin; 2014, http://www.analysisgroup.com/health-care-bulletins/fall-2014/monitoring-online-signals/.
[51] Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyse search, communication and publication behaviour on the Internet. J Med Internet Res 2009;11(1):e11. http://dx.doi.org/10.2196/jmir.1157.
[52] Murphy J, Kim A, Hagood H, Richards A, Augustine C, Kroutil L, et al. Twitter feeds and Google search query surveillance: can they supplement survey data collection?. Bristol UK: Sixth International Conference of the Association for Survey Computing; 2011, https://www.rti.org/pubs/twitter_google_search_surveillance.pdf.
[53] Laurent MR, Vickers TJ. Seeking health information online: does Wikipedia matter? J Am Med Inform Assoc 2009;16:471–9.
[54] Bruns A, Stieglitz S. Twitter data: what do they represent? Inf Technol 2014;56(5):240–5.
[55] Duggan M, Ellison NB, Lampe C, Lenhart A, Madden M. Demographics of key social networking platforms. Pew Research Center; 2015, http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/.
[56] Littlejohn C, Baldacchino A, Schifano F, Deluca P. Internet pharmacies and online prescription drug sales: a cross-sectional study. Drugs Educ Prev Policy 2005;12:75–80.