PERSPECTIVE OPEN

Social media interventions for precision public health: promises and risks

Adam G. Dunn[1], Kenneth D. Mandl[2,3,4] and Enrico Coiera[1]

Social media data can be used with digital phenotyping tools to profile the attitudes, behaviours, and health outcomes of people. While there are a growing number of examples demonstrating the performance of digital phenotyping tools using social media data, little is known about their capacity to support the delivery of targeted and personalised behaviour change interventions to improve health. Similar tools are already used in marketing and politics, using individual profiling to manipulate purchasing and voting behaviours. The coupling of digital phenotyping tools and behaviour change interventions may play a more positive role in preventive medicine to improve health behaviours, but potential risks and unintended consequences may come from embedding behavioural interventions in social spaces.

npj Digital Medicine (2018)1:47 ; doi:10.1038/s41746-018-0054-0

INTRODUCTION

In 2013, a series of new methods were published demonstrating the possibility of using Facebook 'likes' to predict aspects of personality and demographics.[1–3] These experiments showed the ease with which individuals can be profiled using the digital traces they leave behind online, and generated interest among academics as well as commercial and political organisations. Then in 2017, we saw experimental evidence that these tools can be wielded for the purposes of social manipulation,[4] as well as evidence that tools based on these methods were being deployed at unprecedented scales to manipulate voting in elections.[5] News about the way Cambridge Analytica accessed and used Facebook data remind us not only that our personal data can be leveraged to influence our behaviour, but also that the regulatory and ethical frameworks around those activities are underdeveloped.

To understand the role that similar approaches might play in preventive medicine, we examine studies in which social media data are used to predict or model health-related behaviours and outcomes. We then explore how these methods might be operationalised in the design of precision behavioural interventions, and how the effects of these interventions might be amplified or lead to unintended consequences when delivered in a networked public.

FROM CHARACTERISING POPULATIONS TO INDIVIDUAL PROFILING

Changes in the way people live and communicate have made it possible to access data about when people sleep,[6] when and where they exercise,[7,8] and track the information they engage with online.[9] Researchers have used these data in two ways: aggregated to identify signals of population-level outcomes, and at the individual level to predict personal attributes from linked data. Both forms rely on robust measures of behaviours, attitudes, or health outcomes, but the ways they are operationalised to change health behaviours are different.

Population-level studies that aggregate publicly-accessible data have demonstrated the capacity to model spatial variations in voting behaviours,[10] cardiovascular mortality,[11] and vaccine coverage.[12] Studies tend to use Twitter when larger volumes of data are required.[13] These types of studies are validated against traditional data sources including surveys, disease notifications, and census data. In most cases, data from social media platforms produce a biased representation of location or demography.[14,15] Accounting for biases in data are important in studies that conclude about incidence and prevalence without validating models against other data sources (especially social media studies that draw conclusions based on number of tweets).[16] Because studies examining associations between what can be observed on social media and health outcomes have so far been limited to high-prevalence conditions and behaviours like cardiovascular mortality and vaccine coverage,[11,12] it is not yet clear whether social media data can be used to reliably model rarer outcomes. Population-level studies can be operationalised to complement traditional public health surveillance with faster and less costly information; but tend to produce shallow information and are blunt instruments for designing communication interventions.

Individual-level studies that predict attitudes, behaviours, and health outcomes of people work differently, linking social media user data to validated survey instruments or health records, often using much smaller cohorts. An early example demonstrated the ability to predict major depressive episodes from Twitter data and used validated survey tools to establish diagnoses.[17] Mental health has become a common topic of focus,[18,19] though other attitudes, behaviours, and health outcomes have been studied in similar ways.[2,20] There are no barriers to extending these studies to other phenotypes.[21] While this approach can work with much smaller cohorts than population-level studies, their construction

[1]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia; [2]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115, United States; [3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, United States and [4]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, United States
Correspondence: Adam G. Dunn (adam.dunn@mq.edu.au)

Received: 4 May 2018 Revised: 22 August 2018 Accepted: 23 August 2018
Published online: 19 September 2018

and validation rely on the quality of the instruments used to measure attitudes, behaviours, and health outcomes of the participants. We expect that this approach will work across major social media platforms and make it possible to detect reasonable signals of suicidal ideation, the misuse of prescription drugs, problem gambling, unhealthy diets, vaccine hesitancy and refusal, and lifestyle factors associated with increased risks of cancer and cardiovascular disease.

DELIVERING INTERVENTIONS WITHIN A NETWORKED PUBLIC

Effective behaviour change interventions influence the attitudes people hold and the choices they make about their health or the health of their community. Traditional approaches might see a government or public health organisation address problems of vaccine coverage by conducting a survey on vaccine hesitancy to guide the design of a communication intervention; or use population level data about healthcare services to allocate more resources to locations with poorer access. Social media presents an unusual opportunity to identify and deliver personalised digital interventions in an integrated way,[22,23] and there is evidence that this form of personalised social manipulation can be effective.[4] When undertaken with individual consent from participants, such behaviour change approaches will live or die based on the merits of their effectiveness. It is a very different question to contemplate deploying such online personalised interventions at scale, without consent, and where targets of the intervention are unaware that they are being manipulated.

The challenges associated with delivering and evaluating population-level digital behaviour change interventions come from the networked nature of online social spaces. Borrowing from Tufekci,[24] a networked public refers to the complex interactions of people within a society, using communication technologies that facilitate the formation of communities (as structure), as well as the spread of information through those communities (as dynamics). When designing communication interventions to work in social spaces where people are concurrently consumers and broadcasters of information, interactions in the network may be potential confounders or part of the intervention.[22]

The first challenge is evaluation—we are only starting to grapple with the experimental designs needed to test such interventions in trials and in natural settings. Observational evidence shows that health behaviours can be partially explained by the health behaviours and outcomes of families and friends measured in egocentric social networks, including for behaviours related to obesity, smoking, and happiness.[25–27] Trials that can separate and control for the effects of social network structures are still relatively rare.[28] Early evidence from studies that insert software agents into artificially-constrained social network structures demonstrate the potential to drive collective behaviour change,[29] though recent work suggests the form of experiments that may test effectiveness in natural settings.[30]

The second challenge is implementation—social networks may supress or amplify the effects of behaviour change interventions in unpredictable ways. Interventions in this space must compete for attention in an information-rich environment where misinformation may spread faster.[31] Experiments in agent-based simulations and observational data from social media show that even where individuals have the capacity to discern between high and low quality information, an increased volume of information leads to an increased likelihood that low quality information will spread and persist.[32] Online social spaces may also amplify the effects of behaviour change interventions. For example, trials testing the effects of messaging interventions aimed at influencing vaccination attitudes often fail to show an effect on behaviour,[33,34] but this may be because they are tested on individuals in artificial environments rather than in the social spaces where information credibility and beliefs are socially constructed.[35,36]

RISKS AND UNINTENDED CONSEQUENCES

Backlash is a possible short-term consequence of the use of automated behaviour change tools. The public reaction to an interventional study where Facebook manipulated what users saw to determine its effects on mood was emblematic of what can happen when users discover that they have little control over the information they consume.[37] Increased use of these methods may represent an erosion of privacy and with it, a perceived threat to individual autonomy.

Medium term consequences might include driving unhealthy behaviours underground. Marlinspike,[38] in 2013, described how the perceived erosion of privacy that comes with public knowledge of expanded surveillance can create a chilling effect on behaviours, and the importance of privacy even for those who believe they have nothing to hide. Social media users routinely describe using pain drugs, stimulants, and alcohol online. When users discover that organisations are monitoring and manipulating their behaviour, they may adapt by obfuscating what they say or how they interact to avoid being targeted (e.g. when hate speech was targeted on Twitter, users started to use coded language). This would make social media a less reliable signal of behaviour.

Longer term risks may occur if the development of social manipulation methods outpaces the development of countermeasures, where users find ways to hide distinguishing features or limit what they share in the public domain. While there are legitimate reasons for developing and deploying automated behaviour change interventions on social media to improve health, new research efforts in the area could be adapted for use in commercial or political applications. This includes organisations unconstrained by the ethical standards required within academic environments. For example, Cambridge Analytica is suspected to have adapted research from social psychology in an attempt to manipulate voting patterns.[5]

RESEARCH BARRIERS AND OPPORTUNITIES

Given the capacity to scale precision behaviour change interventions to societal levels, clear governance structures are now needed to allow for their safe and ethical use. Within academia, ethics reviews will need to consider not only transparency and alignment with participant values but also the broader impact that reporting may have on society. The 2014 experiment in which Facebook modified timelines was an example where users gave consent by agreeing to the terms and conditions of use of the website but the balance of risks versus benefits of changing their timelines to manipulate their emotions may not have warranted.[37,39] Facebook is not alone. Large internet companies are known for continuously running large numbers of experiments (called A/B tests) without explicitly informing participants. But we typically do now view their impact in the same way because behaviours they seek to change are typically clicks and conversions rather than behaviours that may have direct health implications.

The immediate opportunity in the area comes from linking social media data to surveys and medical records, turning small but high-quality datasets into tools for predicting which individuals are most at risk of certain behaviours or outcomes at societal scales. Methods for iteratively refining predictive models to better target Facebook users are available,[40] and these are likely to improve identification further. The key difference is that social media also permits direct communication with people that have been traditionally hard to reach,[41,42] and to reach them well before they visit a clinic or hospital.

CONCLUSIONS

Questions remain about when it is appropriate to couple tools for digital phenotyping with targeted communication interventions to influence health behaviours. There is evidence that digital phenotyping tools have already been weaponised for political propaganda—we have gone from dropping pamphlets from planes to delivering tailored messages directly into the devices that dominate our attention. While the research area is still in its infancy, examples from outside published research leave little doubt that we can take advantage of social media to deliver fully automated, targeted, and cost-effective behaviour change interventions at scale. Despite the volume of health-related social media research published, only a handful of studies have demonstrably predicted health behaviours and outcomes for individuals. While there is a clear potential for their use in improving health outcomes, there are also risks that the adoption of new tools in the area may lead to a perceived threat to autonomy and backlash. Until researchers have the capacity to evaluate them in well-designed studies demonstrating that the benefits outweigh the risks, we recommend caution in their deployment in preventive medicine and public health.

ACKNOWLEDGEMENTS

The National Health and Medical Research Council (NHMRC) Project APP1128968. The NHMRC played no role in the preparation of the manuscript, nor the decision to submit the manuscript for publication.

AUTHOR CONTRIBUTIONS

A.G.D., K.D.M., and E.C. were responsible for the conception of the work, drafted the manuscript, and critically revised the manuscript for submission. All authors approve the completed version.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Schwartz, H. A. et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE* **8**, e73791 (2013).
2. Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl Acad. Sci.* **110**, 5802 (2013).
3. Youyou, W., Kosinski, M. & Stillwell, D. Computer-based personality judgments are more accurate than those made by humans. *Proc. Natl Acad. Sci.* **112**, 1036 (2015).
4. Matz, S. C., Kosinski, M., Nave, G. & Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proc. Natl Acad. Sci.* **114**, 12714 (2017).
5. Rosenberg, M., Confessore, N. & Cadwalladr, C. in D Baquet (ed.) How Trump consultants exploited the Facebook data of millions The New York Times. (The New York Times Company: New York City, New York, 2018) Retrieved from https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.
6. Althoff, T., Horvitz, E., White, R. W. & Zeitzer, J. Harnessing the web for population-scale physiological sensing: a case study of sleep and performance. In *Proceedings of the 26th International Conference on World Wide Web*. 113–122 (2017) https://doi.org/10.1145/3038912.3052637.
7. Althoff, T. et al. Large-scale physical activity data reveal worldwide activity inequality. *Nature* **547**, 336 (2017).
8. Althoff, T., White, W. R. & Horvitz, E. Influence of pokémon go on physical activity: study and implications. *J. Med. Internet Res.* **18**, e315 (2016).
9. Dunn, G. A., Leask, J., Zhou, X., Mandl, D. K. & Coiera, E. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. *J. Med. Internet Res.* **17**, e144 (2015).
10. Gebru, T. et al. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proc. Natl Acad. Sci.* **114**, 13108 (2017).
11. Eichstaedt, J. C. et al. Psychological language on twitter predicts county-level heart disease mortality. *Psychol. Sci.* **26**, 159–169 (2015).
12. Dunn, A. G. et al. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. *Vaccine* **35**, 3033–3040 (2017).
13. Colditz, J. B. et al. Toward real-time infoveillance of twitter health messages. *Am. J. Public Health* **108**, 1009–1014 (2018).
14. Mellon, J. & Prosser, C. Twitter and Facebook are not representative of the general population: political attitudes and demographics of British social media users. *Res. Polit.* **4**, 2053168017720008 (2017).
15. Tufekci, Z. Big questions for social media big data: representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*. 505–514 (2014).
16. Rothman, K. J., Gallacher, J. E. J. & Hatch, E. E. Why representativeness should be avoided. *Int. J. Epidemiol.* **42**, 1012–1014 (2013).
17. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. In *7th International Conference on Weblogs and Social Media* Vol. 13, 1–10 (AAAI, Boston, 2013).
18. De Choudhury, M., Counts, S., Horvitz, E. J. & Hoff, A. Characterizing and predicting postpartum depression from shared Facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* 626–638, 10.1145/2531602.2531675 (2014).
19. De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G. & Kumar, M. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* 2098–2110, 10.1145/2858036.2858207 (2016).
20. Markovikj, D., Gievska, S., Kosinski, M. & Stillwell, D. Mining facebook data for predictive personality modeling. In *Proceedings of the 7th international AAAI conference on Weblogs and Social Media, Boston, MA, USA* 23–26 (2013).
21. Torous, J. et al. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *npj Digit. Med.* **1**, 15 (2018).
22. Coiera, E. Social networks, social media, and social diseases. *BMJ* **346**, f3007 (2013).
23. Michie, S., Yardley, L., West, R., Patrick, K. & Greaves, F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. *J. Med. Internet Res.* **19**, e232 (2017).
24. Tufekci, Z. *Twitter and Tear Gas: The Power and Fragility of Networked Protest* (Yale University Press: New Haven & London, 2017).
25. Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. *New Engl. J. Med.* **357**, 370–379 (2007).
26. Fowler, J. H. & Christakis, N. A. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *BMJ* **337**, a2338 (2008).
27. Christakis, N. A. & Fowler, J. H. The collective dynamics of smoking in a large social network. *New Engl. J. Med.* **358**, 2249–2258 (2008).
28. Centola, D. Social media and the science of health behavior. *Circulation* **127**, 2135–2144 (2013).
29. Shirado, H. & Christakis, N. A. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* **545**, 370 (2017).
30. Mønsted, B., Sapieżyński, P., Ferrara, E. & Lehmann, S. Evidence of complex contagion of information in social media: an experiment using twitter bots. *PLoS ONE* **12**, e0184148 (2017).
31. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146 (2018).
32. Qiu, X., Oliveira, F. M., Sahami Shirazi, D., Flammini A, A. & Menczer, F. Limited individual attention and online virality of low-quality information. *Nat. Human. Behav.* **1**, 0132 (2017).
33. Nyhan, B., Reifler, J., Richey, S. & Freed, G. L. Effective messages in vaccine promotion: a randomized trial. *Pediatrics* **133**, e835 (2014).
34. Brewer, N. T., Chapman, G. B., Rothman, A. J., Leask, J. & Kempe, A. Increasing vaccination: putting psychological science into action. *Psychol. Sci. Public Interest* **18**, 149–207 (2017).
35. Westerman, D., Spence, P. R. & Van Der Heide, B. Social media as information source: recency of updates and credibility of information. *J. Comput.-Mediat. Commun.* **19**, 171–183 (2014).
36. Coman, A., Momennejad, I., Drach, R. D. & Geana, A. Mnemonic convergence in social networks: the emergent properties of cognition at a collective level. *Proc. Natl Acad. Sci.* **113**, 8171 (2016).
37. Editorial Expression of Concern.Experimental evidence of massive scale emotional contagion through social networks. *Proc. Natl Acad. Sci.* **111**, 10779 (2014).

38. Marlinspike, M. *We Should All Have Something To Hide*. https://web.archive.org/web/20171228164145/https://moxie.org/blog/we-should-all-have-something-to-hide/(2013).

39. Kramer, A. D. I., Guillory, J. E. & Hancock, J. T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl Acad. Sci.* **111**, 8788 (2014).

40. Yom-Tov, E., Shembekar, J., Barclay, S. & Muennig, P. The effectiveness of public health advertisements to promote health: a randomized-controlled trial on 794,000 participants. *npj Digit. Med.* **1**, 24 (2018).

41. Whitaker, C., Stevelink, S. & Fear, N. The use of facebook in recruiting participants for health research purposes: a systematic review. *J. Med. Internet Res.* **19**, e290 (2017).

42. Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V. & Stillwell, D. Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. *Am. Psychol.* **70**, 543–556 (2015).