Predicting inspection outcomes using 'patient voice'

Alex Griffiths and **Meghan Leaver** explore the online world to identify good and bad care

Regulators, it seems, are always being asked to do more with less. Politicians and regulators alike have frequently asserted that this will be possible by 'making better use of data' to target resources effectively. As regulators' risk models come under greater scrutiny however, there is a growing realization of their limitations; the aggregation of administrative data, for example waiting times, mortality rates, and staff turnover in the English National Health Service (NHS), has systematically failed to identify poorly performing Hospital Trusts (Francis, 2013; Griffiths et al., 2016). With the demands on regulators, or the constraints on their budgets, unlikely to go away anytime soon, what can regulators do?

Research conducted at **carr** in cooperation with the LSE's Department of Psychological and Behavioural Science provides one possible solution. Following recommendations by the 2013 Francis Inquiry that patient voice be better monitored in the NHS to avoid a repeat of the scandal at Mid Staffs, we sought to investigate whether the vast amount of disparate feedback posted online could help identify good and bad care, and help regulators prioritize their interventions. It can.

Over the past year we have gathered more than 1.5 million tweets, Facebook posts and comments posted on dedicated patient feedback websites directly concerning NHS hospitals and the Trusts that they comprise. By automatically identifying, classifying and scoring relevant information on a universal scale, and then combining those pieces of information, we have been able to form a 'collective judgement' for each hospital on any given date. There is a strong, statistically significant relationship between the collective judgement on the start date of inspections by the Care Quality Commission (CQC) and the ratings awarded at the end of those inspections. This is true for both individual NHS hospitals and the larger 'Trust' groupings to which they belong.

A key question at this point is how can data generated by people with no clinical expertise produce a meaningful judgement that matches that of the large number of professional inspectors, onsite analysts, 'experts by experience' and clinicians that constitute a CQC inspection team? The answer to that question, oddly, comes from the 1906 West of England Fat Stock and Poultry Exhibition. There, the statistician Francis Galton came across a competition to guess the butchered weight of a live ox. There were 800 competitors, most of whom were not experts in cattle or butchery, submitting their guesses on numbered cards. With the competition over and the weight of the butchered ox determined as 1,198 lbs, Galton borrowed the 800 entry slips to analyse the guesses. Much to his surprise, the average of those guesses was 1,197 lbs, essentially perfect.

What Galton had stumbled upon was what is now referred to as the 'Wisdom of Crowds'. The phenomenon means that, under the right circumstances, groups can be remarkably insightful. This can be the case even if the majority of people within a group are not especially well informed or rational (Surowiecki, 2004). Whilst we as individuals seldom have all the necessary facts to make an accurate assessment, and are subject to numerous heuristics and biases, when our individual assessments are aggregated in the right way, our collective assessment is often highly accurate.

Although the theory behind identifying poor quality care with patient feedback is simple, the practicalities are not. Whilst Galton had a manageable number of entry slips, all featuring an estimate that related solely to the competition he was interested in on a specific date, we faced the equivalent of an unknown but vast number of entry slips, only some of which contain relevant information, relating to hundreds of competitions over a number of years. There have, therefore, been a number of practical challenges to overcome.

The first challenge has been accessing, updating and storing the data with each source presenting its own unique problems. The second challenge lay in sorting the relevant from the irrelevant comments. The majority of tweets mentioning hospitals relate not to quality of care they provide, but to recruitment, public information and self-promotion. One in every 1,500 tweets concerns cake. After a significant investment of time and expertise and the use of sufficiently equipped hardware, we developed an algorithm which identifies with 95.9% accuracy whether a tweet concerns one of four aspects of care quality, or does not concern the quality of care at all. Without this advanced automation, it would not be possible to consistently and economically identify the relevant patient feedback, and hence derive a meaningful collective judgement on the quality of care.

The third challenge has been how to extract meaning from the data. Continuing with the Twitter example, we knew which tweets related to specific aspects of the quality of care at a given hospital or Trust, but not whether it was positive or negative, or to what extent. Again, due to the ever growing volume of data, this scoring cannot realistically be done by humans. A second algorithm was therefore developed to read and score the data from disparate sources on the same scale. Only then, with the relevant comments identified and scored on a unified scale, could the weighted, moving average of their sentiment be calculated to derive a 'collective judgement' on the quality of care at each hospital on any given date.

What, then, does the successful, automated use of patient voice mean for regulators? The key message is that, under the right circumstances, high volume, third party data can succeed where traditional administrative data has proved ineffective and help to target limited resource. Furthermore, it may allow regulators to not only better target their resources towards individual regulatees, but focus more precisely on specific areas of concern within those regulatees. Whereas administrative data tends to be reported at the level of overarching hospital trust, university or energy supplier for example – large, diverse groupings which can contain significant internal variation in quality – 'crowds' may be willing and able to target activity at a more granular level such as hospital, academic department or business area. Moreover, without the ability to trigger inspections ourselves we are unable to test the potential of declining collective judgements to identify and prevent problems before they become more serious, and reacting quickly to collective judgement may also serve to prevent, rather than simply identify, poor performance. The use of high volume, third party data may therefore have significant benefits for overburdened regulators.

The findings also raise a number of secondary questions for regulators and their own capacities. Firstly, as ever increasing volumes of decentralized information become available, effective risk monitoring and resource prioritization may require fewer analysts pouring over spreadsheets, but a smaller number of more highly skilled data scientists instead. Secondly, if regulators fail to set the trend in this area, they may face being delegitimized by private sector organizations stealing a march on the effective identification of regulatory risks. Thirdly, when 'service users' can, as a whole, successfully identify poor care, even in a field as complex as acute healthcare, regulators may face a tougher challenge convincing others of their value.

References

Francis, R. (2013). Report of the Mid Staffordshire NHS Foundation Trust public inquiry. London: The Stationery Office.

Griffiths, A., Beaussier, A-L., Demeritt, D. and Rothstein, H. (2016) 'Intelligent monitoring? Assessing the ability of the Care Quality Commission's statistical surveillance tool to predict quality and prioritise NHS hospital inspections.' BMJ Quality & Safety, qualitysafety.bmj.com/content/early/2016/04/15/bmjqs-2015-004687

Surowiecki, J. (2004) The Wisdom of Crowds. London: Little, Brown.

Alex Griffiths is the QUAD Research Officer in **carr**, **Meghan Leaver** is a doctoral researcher in the Department of Psychological and Behavioural Science at the LSE.