Variation in Responsiveness to Warranted Behaviour Change Among NHS Clinicians: a Novel Implementation of Change-Detection Methods in Longitudinal Prescribing Data

Journal:	*BMJ*
Manuscript ID	BMJ-2019-049437
Article Type:	Research
BMJ Journal:	BMJ
Date Submitted by the Author:	25-Feb-2019
Complete List of Authors:	Walker, Alex; University of Oxford Centre for Evidence-Based Medicine, Primary Care Health Sciences Pretis, Felix; University of Victoria, Department of Economics; University of Oxford, Institute for New Economic Thinking Powell-Smith, Anna; University of Oxford Centre for Evidence-Based Medicine, Primary Care Health Sciences Goldacre, Ben; University of Oxford, Primary Care Health Sciences
Keywords:	
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
Supplementary File B - manual_checking_graphs.zip	

Variation in Responsiveness to Warranted Behaviour Change Among NHS Clinicians: a Novel Implementation of Change-Detection Methods in Longitudinal Prescribing Data

Alex J Walker[1], Felix Pretis[2,3], Anna Powell-Smith[1], Ben Goldacre[1*]

*Corresponding author: ben.goldacre@phc.ox.ac.uk

[1]The DataLab
Nuffield Department of Primary Care Health Sciences
University of Oxford
Radcliffe Observatory Quarter
Woodstock Road
Oxford OX2 6GG
UK

[2]Department of Economics,
University of Victoria,
British Columbia,
Canada

[3]Institute for New Economic Thinking,
Oxford Martin School,
University of Oxford,
UK

Abstract

Objectives

Medicine aspires to be driven by evidence and data, against a background of constant change in evidence on effectiveness, safety, and cost-effectiveness for existing interventions; and the introduction of new interventions. While there is extensive literature on variation in care at a static time-point, there has been little work to quantify variation in speed of implementation for warranted changes, despite substantial anecdotal variation between clinicians and institutions in speed of adoption and change. We set out to develop tools which can measure speed of change at healthcare institutions, in order to give routine feedback on responsiveness, and to describe variation in delay to implementation across the health service.

Design

We repurposed an existing break-detection algorithm (a modified version of indicator saturation) to detect changes in prescribing behaviour. We use two example time-series to assess the method: expiry of the Cerazette patent, leading to cheaper generic desogestrel alternatives becoming available; and a change in antibiotic prescribing guidelines, favouring nitrofurantoin over trimethoprim for uncomplicated urinary tract infection (UTI).

Setting

English primary care prescribing data.

Participants

English general practices.

Outcome measures

We measured: the timing of the largest changes; the steepness of the change slope (change in proportion per month); and the magnitude of change.

Results

We found substantial heterogeneity between institutions in both timing and steepness of change. The range of time delay before a change was implemented was extremely large: interquartile range (IQR) 2-14 months for Cerazette (median 8), and IQR 5-29 months for UTI (median 18). There was similarly substantial heterogeneity in slope following a detected change: IQR 2-27% absolute reduction per month for Cerazette (median 9%); and IQR 1-7% for UTI (median 2%). When change was implemented, the magnitude of change exhibited substantially less heterogeneity: IQR 48-87% (median 70%) for Cerazette, and IQR 29-47% (median 38%) for UTI.

Conclusions

We were able to automatically and robustly detect changes in prescribing behaviour. There is very substantial variation in the speed with which individual NHS GP practices responded to change in cost and guidance. Detection of structural breaks using indicator saturation methods opens up new opportunities to improve patient care through audit and feedback by moving away from cross-sectional analyses, and automatically identifying institutions who respond rapidly, or slowly, to warranted changes in clinical practice.

What is already known on this topic

- Implementation of new evidence is critical in a well performing healthcare system.
- Speed of diffusion of innovations in healthcare is thought to vary, but previous work has focused on cross-sectional analysis, small samples, or narrative descriptions.

What this study adds

- Our method can automatically and robustly detect the timing and magnitude of changes in clinical behaviour across thousands of individual institutions.
- This method opens up new opportunities to improve patient care through audit and feedback, by moving away from cross-sectional analyses and automatically identifying institutions who respond rapidly, or slowly, to warranted changes in clinical practice.
- In two example measures of clinical behaviour we identify very substantial variation in the timing and slope for implementation of change between general practices.

Introduction

Medicine is characterised by the development of new interventions, and new information on existing interventions. This requires that clinical practice changes in response to updated evidence on effectiveness, safety, and cost. The “diffusion of innovation” is a longstanding area of research, originating with 1950s work on agriculture [1] and antibiotics [2]. Previous work has largely focused on narrative descriptions, discussing the nature of the innovation (its relative advantage, compatibility, complexity to implement); the channels through which the innovation is communicated; and the ‘social system’ that is involved in implementing the innovation [3]. Previous quantitative work has relied on the manual characterisation of individuals and organisations as either adopting, or not adopting, a new intervention [1,2]. Typically, the rate of adoption is variable over time, with a small number of initial ‘early adopters’, followed by a large number of institutions rapidly adopting the change, then a slower rate while ‘laggards’ adopt the change over a longer period [3].

Diffusion of innovations has received some attention in healthcare [4], but research to date has primarily focused on case studies [5], narrative descriptions of clinicians’ responses to change in guidance [6], interviews [7–9] and theoretical frameworks [10]. Previous quantitative work assessing implementation of new practices has typically relied on measuring change at the level of a whole population, using techniques such as interrupted time-series analysis [11,12]; or static measures of variation in care at a single point in time through “atlases of variation” and regression analyses [13–19].

Assessing variation between institutions in timing of implementation for new clinical behaviours requires a systematic and robust method to identify when institutions have made a change. As it is not feasible to manually review thousands of time-series charts to determine when meaningful change has occurred, this must be done computationally. Statistical methods for the detection of structural change (‘Break detection’) provide a robust method of detecting the timing of changes in time-series data without imposing an intervention or change date *a-priori* [20]. These techniques have previously been applied to a diverse range of applications, including economic- and climate-modelling [21,22].

We therefore set out to repurpose and adapt, for medical time-series data, statistical break-detection techniques based on indicator saturation. Here we report the deployment of these methods to assess variation in speed of adoption for two examples of warranted change in clinical practice: firstly, the move from branded to generic versions of the oral contraceptive desogestrel in 2012; and secondly, the change from trimethoprim to nitrofurantoin as first-line antibiotic for treating uncomplicated urinary tract infection (UTI) at various time-points after 2014.

Methods

Data

The monthly prescribing datasets, published by the NHS Business Services Authority, contain one row for each treatment and dose, in each prescribing organisation in NHS primary care in England, describing the number of prescriptions issued and the total cost. To extract data on standard general practices, we limited to institutions with setting code "4": general practices (GPs) [23], excluding all other organisations such as dentists, prisons and walk-in centres. Practices with incomplete time series, or those that did not vary during the time series, were excluded for that specific measure.

Desogestrel measure: We measured the total proportion of desogestrel prescriptions that were prescribed as branded Cerazette. A decrease in this proportion would correspond to an improvement in this measure. The time series for this measure ran from October 2010 to December 2015. This was in order to centre the data on the time period surrounding the patent expiry in December 2012.

Trimethoprim/nitrofurantoin measure: We measured the proportion of trimethoprim prescriptions as a proportion of total trimethoprim and nitrofurantoin prescriptions. A decrease in this proportion would correspond to an improvement in this measure. The time series for this measure ran from June 2013 to June 2018. This was done in order to centre the data on the time period surrounding the interventions: the change in antibiotic prescribing guidance in October 2014; followed by the introduction of a "Quality Premium" financial incentive, which was announced in October 2016 and implemented in April 2017

Detection

We used trend-indicator saturation [20], a modified version of indicator saturation [24] in each practice's time series to determine whether there was any statistically significant change in prescribing behaviour [24]. We formulated the detection of breaks as a model selection problem, where a time-series regression model of the prescribing behaviour is saturated with a full set of step-functions interacted with a linear time-trend. We select over these break functions at every point in time, removing all non-significant breaks at a chosen level of significance (in this case, p=0.000001) to tightly control the false-positive rate. The false-positive rate of detection is given by pT where T denotes the total number of observations in the sample. For a sample of T=100 observations, we therefore expect pT = 100 x 0.000001 = 0.0001 changes to be detected spuriously on average. Step-shift (or "cliff-like") changes in behaviour can be approximated by a single breaking trend with a high coefficient on the slope, while gradual, smooth-transition behaviour [25] can be approximated through a series of multiple broken linear trends with smaller slope coefficients.

To assess whether the methods for break detection were operating as expected, graphs of the time series were manually inspected, plotted along with the fitted regression model and detected changes. 100 randomly sampled graphs from each time-series were inspected in detail by two blinded researchers independently to ensure that the automatically detected break points overall reflected a true change in prescribing behaviour, with each giving a narrative description of any issues raised. All remaining graphs were rapidly reviewed to check for gross errors in automated detection.

Indicators of change

We generated three indicators to describe the response in prescribing behaviour of each practice.

Timing: The timing of a change in behaviour is measured as the start of the steepest negative (downward) shift in a time-trend of prescribing behaviour during the time series. This measure captures how long it takes a practice to begin to exhibit a substantial change in behaviour following a stimulus (in these examples: a medicine patent expiry, and a change in clinical guidance).

Slope: This is the steepness of the detected changes, and measures the pace of change within a practice (sudden, or gradual) once change has begun. Given that there may be more than one trend-shift detected, steepness is reported as the average slope of the steepest contiguous segment contributing at least 50% to the total level change from the point of the first detected shift that marks a change in behaviour, until the mean of the time series at the end of the sample.

Magnitude: The magnitude of change describes the extent to which each practice reduces the prescribing of the non-favoured medication. This is calculated by subtracting the mean proportion of "unfavourable prescribing" at the end of the time period from the proportion of "unfavourable prescribing" at the time of the largest detected change.

Data Sharing

Data management was carried out using SQL (in Google BigQuery), Python and R. Break detection was implemented using the R-package gets [20]. Complete code and data are provided online on Github (https://github.com/ebmdatalab/prescribing_change_metrics/tree/first_paper).

Patient and Public Involvement

We run OpenPrescribing.net, an openly accessible data explorer for all NHS England primary care prescribing data, which receives a large volume of user feedback from professionals, patients and the public. This feedback is used to refine and prioritise our informatics tools and research activities. Patients were not formally involved in developing this specific study design.

Results

Data

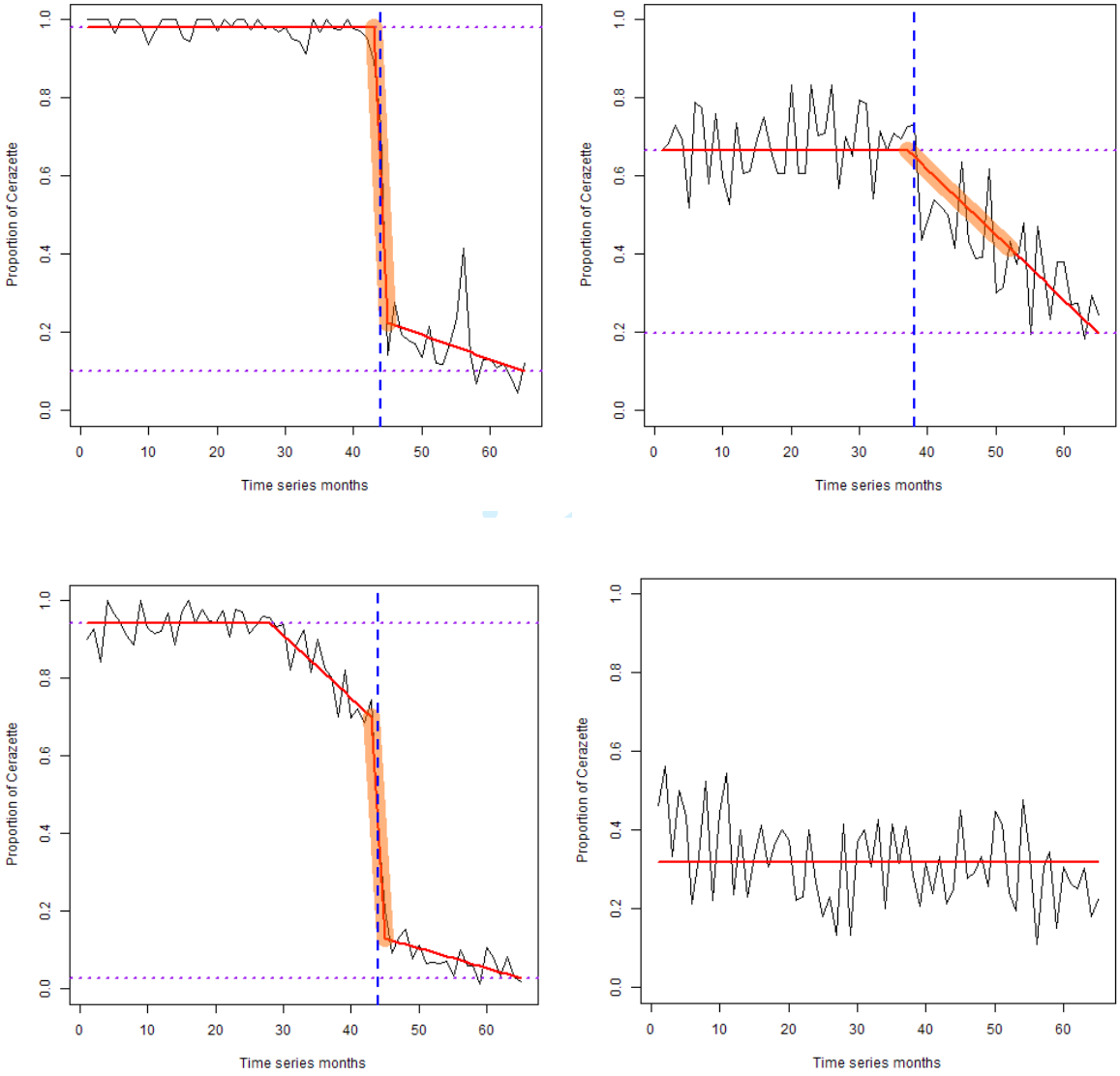
The total number of practices included in the analyses was 8,078 UK GPs. 2,034 practices were excluded from the Cerazette analyses, and 1,121 from the trimethoprim/nitrofurantoin analysis, due to incomplete time series. Missing values were either caused by small numbers leading to months where the denominator was 0, or by the practice not being open for part of the time series, for example due to closing. One practice was removed from the desogestrel measure due to every value being 1.0.

Detection

Figure 1 shows examples of practice time series for the desogestrel measure, illustrating the three indicators of change. The timings of detected breaks (the steepest substantial negative shift) are marked as a vertical dashed line. The segments over which the average slope is calculated are highlighted with orange shading. The magnitude of change is calculated as the difference between the dashed purple lines. Figure 1(a) shows a practice where a steep, cliff like change is detected, followed by a change to a more gradual decline, while Figure 1(b) shows a single, gradual detected change. Figure 1(c) shows a practice where an early gradual change is detected, followed by a steeper change: as above, for our descriptive analysis we report timing, slope, and magnitude for the breakpoint contributing to the largest change in practice. For the practice in Figure 1(d), no changes were detected that reached the necessary significance level (0.000001).

During the process of manual inspection of 200 randomly selected graphs a bug was found and fixed, whereby if the initial variance of the time series was very low (for example, if a practice prescribed 100% branded Cerazette for many months initially) the technique would become hypersensitive to change, leading to inappropriate detection. This was fixed by tweaking one of the parameters of the change detection algorithm (the maximum size of the block-partitioning). The algorithm was otherwise found to be operating as expected: out of 200 time-series reviewed, we found two cases of sub-optimal detection; and four cases of arguable / borderline suboptimal detection. The time-series examined, and manual-checking datasheet, can be seen in Supplementary Files A and B.

***Figure 1**: **Illustration of the methodology**. Proportion of Cerazette relative to total desogestrel prescribed (black solid line) of 4 representative general practices. Fitted-model and detected breaks using trend-indicator saturation (SIS interacted with linear trends) are shown in red. The commencement of the largest negative shift is marked with a vertical line of blue long-dashes, while additional breaks are indicated by changes in the slope of the red line. The measured slope is highlighted in orange, the pre-break level and final level at the end of the sample are marked by purple-dashed lines.*



Indicators of change

Table 1 summarises the detected heterogeneity in prescribing behaviour across all practices, for both measures, with summary statistics over the three estimated measures.

Table 1: Detected changes in prescription behaviour. Results across the three measures of change across all practices.

Prescribing measure	Metric	Timing (months from intervention)	Steepness (% change per month)	Magnitude (% change)
Desogestrel	Median	8	8.7	69.5
	Mean Abs. Deviation	12.3	21.1	24.6
	Interquartile range	2 – 14	2.0 – 26.6	48.3 – 87.3
UTI antibiotics	Median	18	2.1	37.8
	Mean Abs. Deviation	14.3	10.0	13.7
	Interquartile range	5 – 29	1.0 – 7.0	28.7 – 46.7

Timing measurements

For both measures, there is considerable heterogeneity between practices in the timing of their largest response to the warranted change in practice. The top panels of figures 2 and 3 show the distribution of the largest detected changes for each measure. Changes are detected across the whole range of the time series. Practices tended to respond more quickly, and with less variation for the desogestrel measure than the trimethoprim/nitrofurantoin measure. For the desogestrel measure, the largest peak in detected changes occurred a few months after the expiry of the Cerazette patent. In contrast, there are relatively few detected changes in the months following the UTI antibiotics guidance change, with the peak in detected changes not occurring until after the announcement of the Quality Premium financial incentive.

Figure 2: *The response of general practices to the patent expiry and subsequent price change for Cerazette/desogestrel. Top panel: histogram showing the number of practices with their largest detected downward change in each month. Second panel: bar chart showing the mean slope of detected change for all practices changing in that month. Third panel: bar chart showing the mean magnitude of detected change for all practices changing in that month. Bottom panel: line graph showing median Cerazette prescribing as a proportion of all desogestrel prescribing (solid line), along with deciles (dashed lines) and extreme percentiles (dotted lines).*

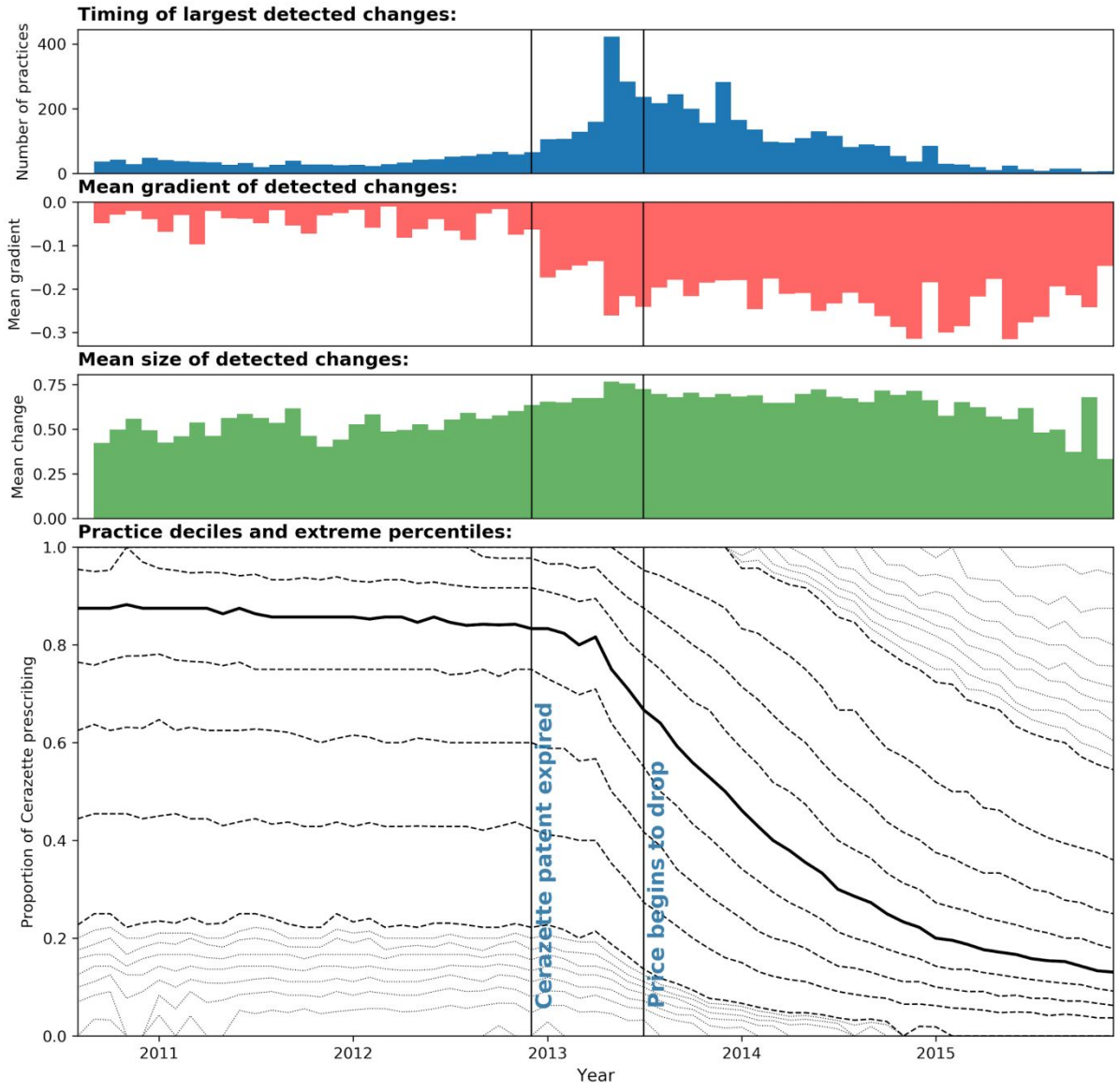
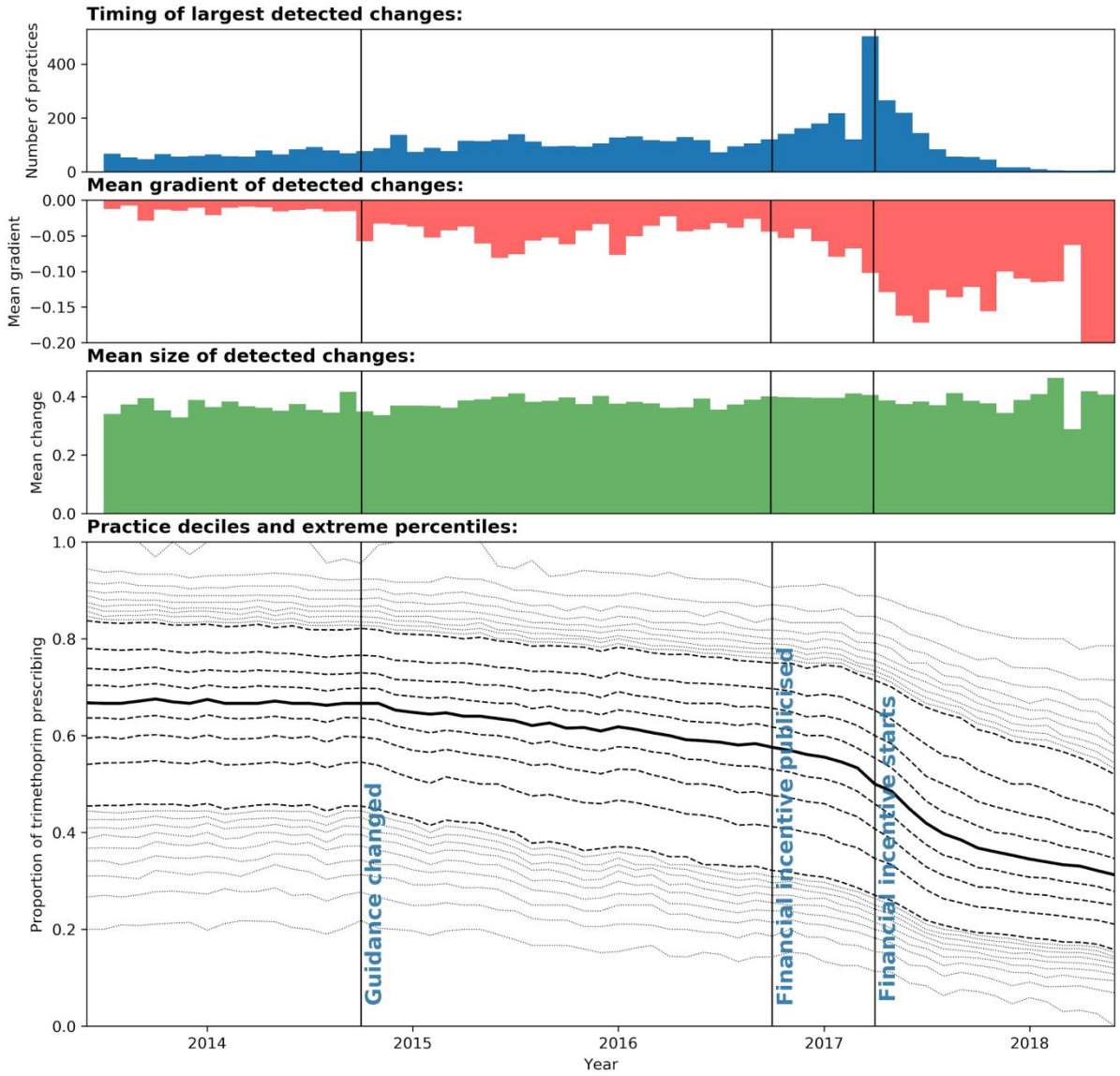


Figure 3: *The response of general practices to a change in antibiotic prescribing guidance (from trimethoprim to nitrofurantoin for uncomplicated UTIs). Top panel: histogram showing the number of practices with their largest detected downward change in each month. Second panel: bar chart showing the mean slope of detected change for all practices changing in that month. Third panel: bar chart showing the mean magnitude of detected change for all practices changing in that month. Bottom panel: line graph showing median trimethoprim prescribing as a proportion of all total trimethoprim and nitrofurantoin prescribing (solid line), along with deciles (dashed lines) and extreme percentiles (dotted lines).*



Slope measurements

The slope of detected change is also highly variable between practices (second panels, figures 2 and 3), especially for the desogestrel measure, where there is a greater than ten-fold difference in slope of change between the practice at the 25th centile and the 75th centile. For the desogestrel measure, steepness of change is substantially greater following the expiry of the Cerazette patent, indicating that those changing later typically do so more rapidly. The mean slope of detected change for the trimethoprim/nitrofurantoin measure is generally much lower, indicating slower change in practice; the mean slope only

substantially increases following the implementation of the Quality Premium financial incentive in April 2017.

Magnitude measurements

The level of heterogeneity in the magnitude of change is less than that for timing or slope of change (third panels, figures 2 and 3). It is somewhat variable over time for the desogestrel measure (figure 2), but uniform over time for the trimethoprim/nitrofurantoin measure (figure 3).

Discussion

Summary

The indicator saturation method was successfully implemented to detect meaningful changes in clinical practice. Among GP practices in the English health system we describe very substantial heterogeneity in the timing and slope of warranted changes in clinical practice following changes in price and clinical guidance.

Interpretation

The changes measured in this paper were highly warranted from a cost-effectiveness or clinical perspective, emphasised by the fact that most practices eventually made the change. However, the distribution of the measures of timing, slope of change - and to a lesser extent magnitude - show high variation and skewness. While a large proportion of practices exhibit a significant shift away from branded Cerazette in early 2013, a quarter do not show their most substantial change for 14 months (February 2014), with the slowest 10% changing at least a further 6 months later (September 2014), exposing the health system to substantial avoidable costs. The spread of timing of changes is even more dramatic for the trimethoprim/nitrofurantoin measure, with a quarter of practices not making their largest change until 29 months after the guidance was released, and 10% not changing until at least 32 months after release, exposing patients to suboptimal care. The slower dissemination of the antibiotic guidance could be due to it being a less clear change, with some clinical judgement involved, rather than “always prescribe the generic”, as with desogestrel.

This spread of response is not limited to the timing, but also the slope of the change. For example: the highest quartile of practices for “slope of response” reduced their proportion of branded Cerazette prescribing very swiftly, by at least 26% in a single month; while lowest quartile of practices for slope of response reduced branded prescribing very gradually, by less than 2% per month. There is also some indication (Figure 2, first and second panels) that practices implementing a change late tend to do so more rapidly than those who notice the need for change earlier: this is perhaps due to an increased sense of urgency for practices that have noticed later. Regardless of the heterogeneity in timing and slope of change, the relative uniformity in the magnitude of change suggests that once practices implement a change, they are able to do so effectively, with most practices ultimately implementing a large change in practice.

Strengths and weaknesses

Our data covers the complete prescribing data for all practices in England, not a sample. The underlying data are highly accurate, as they are based on prescription pharmacy claims used for very high tariff transactions within the health service, with all parties motivated to ensure complete and correct information. We account for confounding by indication by measuring the proportion of “all” prescribing that is “undesirable”, rather than the crude volume of “undesirable” prescribing (i.e. we measure Cerazette as a proportion of all Cerazette and generic desogestrel prescribing; and trimethoprim as a proportion of total trimethoprim and nitrofurantoin prescribing).

The indicator saturation approach to detect breaks successfully detected change in prescribing behaviour, and appears to be flexible across two very different applications: the

desogestrel measure has a single, unambiguous time point, after which prescribing generically is simply and unambiguously correct; the nitrofurantoin/trimethoprim guidance, by contrast, was communicated to clinicians through various different routes at different times, and is a change in practice that requires ongoing clinical judgement, as it may not always be correct to prescribe nitrofurantoin rather than trimethoprim for all patients.

We have focused our analysis on practises with complete time series. For practices where it is important to measure change despite missing values, this could be addressed through imputation; however practices with missing values tend to be smaller, with substantial noise in their prescribing data, and this presents additional challenges for detecting meaningful changes in clinical practice by any means, including manual categorisation.

We chose to present summary statistics for the steepest detected change, as that represents the most important and coordinated change. However, our break-detection approach could be used in different ways for different clinical and research problems, for example focusing on the first detected change in clinical practice, or the first change to reach a pre-specified threshold, depending on specific needs.

Findings in context

To our knowledge this is the largest study ever conducted on diffusion of change in medical practice. The largest previous study monitored 95 solo practitioners in Denmark, covering a population of 490,000 citizens; by comparison our study covers a population of 55 million citizens. More importantly, this previous study assessed only one crude outcome metric (time to first prescription of a new antibiotic); whereas we were able to harness novel computational methods to automatically detect more detailed changes in clinical practice, across a very large number of institutions (6,000 practices), and for more complex and generalisable clinical behaviours than “first ever prescription of a new medicine”.

The previous absence of computational techniques such as indicator saturation explains why most prior work on diffusion of change is either small-scale, or focused purely on narrative descriptions (as discussed in the Introduction): without automation, it is extremely labour-intensive to manually categorise whether, and when, a large number of institutions have modified their clinical practice in response to a warranted change.

Policy Implications

We identify two sets of policy implications from this work: the fact of substantial heterogeneity in response to warranted changes in practice; and the potential for better metrics and feedback to clinicians through the application of indicator-saturation-based break detection methodology to clinical data.

Variation In Speed Of Implementation

For both the prescribing measures studied in this analysis, we observed substantial heterogeneity in timing and slope of warranted change, but almost all practices ultimately exhibited very substantial changes in clinical practice. In lay terms: most practices changed their behaviour, but some changed much later than others; and some practices exhibited very rapid, coordinated change, while others changed only very gradually.

This heterogeneity is problematic: it exposes health systems to substantial avoidable costs, and exposes patients to suboptimal clinical care. While it may be unrealistic to expect all practices to respond immediately and adopt optimal prescription behaviour, the fact that some practices changed both early and rapidly suggests that rapid timely change is possible. Further work is required to explore the reasons for some practices being slow to implement prescribing changes. We have previously written on the importance - and comparative neglect - of systems to disseminate knowledge to clinicians and patients, and social structures to audit and assess the implementation of warranted changes in practice [6,26].

Novel Applications of Indicator Saturation

The automation of change detection also presents new opportunities for better use of data in audit and feedback on clinical practice, which has been shown in systematic review data to solicit modest but cost-effective improvements in clinical practice [27]. Such audits currently rely on a static snapshot of clinical practice. Indicator saturation methods raise the potential for more sophisticated metrics: for example, describing whether an individual clinician or institution tends to respond rapidly or slowly to changes in price, evidence, or safety across a range of different elements of clinical practice. This in turn may lead to better targeting of resource to support those who are responding slowly across a range of warranted changes.

Automated change detection also permits new approaches to interrogate which interventions are most impactful at soliciting change in clinical practice: for example, in Figure 3, the financial incentive is clearly associated with the largest number of practices initiating change in a single month.

These new methods may additionally help to distinguish between “warranted” and “unwarranted” variation in care, itself an ongoing challenge for all work on variation in clinical practice: specifically, whether observed variation is driven by variation in patients’ clinical needs and preferences (“warranted” variation); or variation in their clinicians’ knowledge, preferences, and service availability (“unwarranted” variation). A clinician presented with evidence that they are currently an outlier for a new desired change in clinical practice may argue that their patients are unusual and warrant clinical decisions that deviate from best practice guidelines; however, if indicator saturation methods show that there have been previous warranted changes in clinical practice that were ultimately implemented by this clinician, but three years later than their peers, then this is stronger evidence that current deviation from best practice is driven by the clinician’s knowledge or choices, rather than their patients’ needs or preferences.

Lastly, the potential to automate detection in timing and slope of change using indicator saturation presents an immediate opportunity to produce automated metrics on timing of change for individual clinicians and institutions. OpenPrescribing is an openly accessible service for detailed exploration of NHS England prescribing data by practice and by month, run by our team, with 104,000 users during the previous year. We are therefore now developing novel measures driven by indicator saturation to describe whether practices and CCGs overall tend to implement warranted changes in clinical practice earlier or later than their peers, for deployment and impact evaluation in a large pool of users.

Conclusions

Indicator saturation methods to detect structural breaks present substantial new opportunities to improve clinical practice by better identifying and reducing unwarranted variation in care.

Conflicts of Interest
BG has received research funding from the Laura and John Arnold Foundation, the Wellcome Trust, the Oxford Biomedical Research Centre, the NHS National Institute for Health Research School of Primary Care Research, the Health Foundation, and the World Health Organisation; he also receives personal income from speaking and writing for lay audiences on the misuse of science. AJW is employed on BG's grants for the OpenPrescribing project.

Funding
No specific funding was sought for this analysis. Work on OpenPrescribing is supported by the Health Foundation (ref 7599); the NIHR Biomedical Research Centre, Oxford; and by an NIHR School of Primary Care Research grant (ref 327). Funders had no role in the study design, collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Ethical approval
This study uses exclusively open, publicly available data; no ethical approval was required.

Contributorship
BG conceived the study. AJW, FP, APS and BG designed the methods. AJW collected and analysed the data with input from FP APS and BG. AJW, FP and BG drafted the manuscript. All authors contributed to and approved the final manuscript. BG supervised the project and is guarantor.

Transparency
The manuscripts guarantor affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

References

1 Ryan B, Gross N. Acceptance and diffusion of hybrid corn seed in two Iowa communities. *Res Bull* 1950;**29**:1.

2 Coleman JS, Katz E, Menzel H. *Medical innovation : a diffusion study*. Indianapolis: : Bobbs-Merrill Co 1966.

3 Rogers EM. *Diffusion of Innovations, 5th Edition*. Riverside, UNITED STATES: : Simon and Schuster 2003.

4 Berwick DM. Disseminating innovations in health care. *JAMA* 2003;**289**:1969–75.

5 Steffensen FH, Sørensen HT, Olesen F. Diffusion of new drugs in Danish general practice. *Fam Pract* 1999;**16**:407–13.

6 Croker R, Walker AJ, Goldacre B. Why did some practices not implement new antibiotic prescribing guidelines on urinary tract infection? A cohort study and survey in NHS England primary care. *J Antimicrob Chemother* Published Online First: 22 December 2018. doi:10.1093/jac/dky509

7 Jones MI, Greenfield SM, Bradley CP. Prescribing new drugs: qualitative study of

influences on consultants and general practitioners. *BMJ* 2001;**323**:378–81.

8 Prosser H, Almond S, Walley T. Influences on GPs' decision to prescribe new drugs-the importance of who says what. *Fam Pract* 2003;**20**:61–8.

9 Prosser H, Walley T. New drug uptake: qualitative comparison of high and low prescribing GPs' attitudes and approach. *Fam Pract* 2003;**20**:583–91.

10 Eccles M, Grimshaw J, Walker A, *et al.* Changing the behavior of healthcare professionals: the use of theory in promoting the uptake of research findings. *J Clin Epidemiol* 2005;**58**:107–12.

11 Curtis HJ, Walker AJ, Goldacre B. Impact of NICE guidance on tamoxifen prescribing in England 2011-2017: an interrupted time series analysis. *Br J Cancer* 2018;**118**:1268–75.

12 Walker AJ, Curtis HJ, Goldacre B. Impact of Chief Medical Officer Activity on Prescribing of Antibiotics in England - An Interrupted Time Series Analysis. *J Antimicrob Chemother*

13 Wennberg JE. *Tracking Medicine: A Researcher's Quest to Understand Health Care*. 1 edition. Oxford University Press, USA 2010.

14 Wise J. Health atlas shows large variations in care in England. *BMJ* 2010;**341**:c6809.

15 Curtis HJ, Croker R, Walker AJ, *et al.* Opioid Prescribing Trends and Geographical Variation in England 1998-2018 - A Retrospective Database Study. *Lancet Psychiatry*

16 Curtis HJ, Walker AJ, Mahtani KR, *et al.* Time trends and geographical variation in prescribing of antibiotics in England 1998–2017. *J Antimicrob Chemother* Published Online First: 15 September 2018. doi:10.1093/jac/dky377

17 Walker AJ, Curtis HJ, Bacon S, *et al.* Trends and variation in prescribing of low-priority treatments identified by NHS England: a cross-sectional study and interactive data tool in English primary care. *J R Soc Med* 2018;**111**:203–13.

18 Curtis HJ, Dennis JM, Shields BM, *et al.* Time Trends and Geographical Variation in Prescribing of Drugs for Diabetes in England 1998-2017. *Diabetes Obes Metab* 2018.

19 Walker AJ, Curtis HJ, Bacon S, *et al.* Trends, geographical variation and factors associated with prescribing of gluten-free foods in English primary care: a cross-sectional study. *BMJ Open* 2018;**8**:e021312.

20 Pretis F, Reade J, Sucarrat G. Automated General-to-Specific (GETS) Regression Modeling and Indicator Saturation for Outliers and Structural Breaks. *Journal of Statistical Software, Articles* 2018;**86**:1–44.

21 Pretis F, Mann ML, Kaufmann RK. Testing competing models of the temperature hiatus: assessing the effects of conditioning variables and temporal uncertainties through sample-wide break detection. *Clim Change* 2015;**131**:705–18.

22 Pretis F, Schneider L, Smerdon JE, *et al.* Detecting Volcanic Eruptions in Temperature Reconstructions by Designed Break-Indicator Saturation. *J Econ Surv* 2016;**30**:403–29.

23 NHS Digital. Number of Patients Registered at a GP Practice. 2017.http://content.digital.nhs.uk/gppatientsregistered (accessed 3 May 2017).

24 Castle J, Doornik J, Hendry D, *et al.* Detecting Location Shifts during Model Selection by Step-Indicator Saturation. *Econometrics* 2015;**3**:240–64.

25 Terasvirta T. Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models. *J Am Stat Assoc* 1994;**89**:208–18.

26 Walker AJ, Bacon S, Curtis H, *et al.* Six months on: NHS England needs to focus on dissemination, implementation and audit of its low-priority initiative. *J R Soc Med* 2018;:141076818808429.

27 Ivers N, Jamtvedt G, Flottorp S, *et al.* Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev* 2012;:CD000259.

Practice	**Alex comment**	**Ben comment**	**Resolve in discussion**	**action?**	**code: 0=non-issue 1=arguably sub-optimal detection 2=sub-optimal detection**
desogestrel					
B87031	two drops -prescribing is weird		agree could have been shallower slope but same initiation month. note moving p-value threshold changes this, but n detected changes in other practices drops substantially: trade-offs.		1
B83033		arguably better to detect change at 39 months and consider whole drop as one drop.	marginal. overall no issue here.		0
D81062	slightly early first change		not material to the results of this study, agree somewhat interesting!		0
E81009		very much borderline, and depends on use-case, but possibly better to consider it as one drop starting month 33.	marginal. overall no issue here.		0
E85746	I'd have drawn two flat lines with a step change		arguable, but noisy institution. marginal. overall no issue here.		0
F81114	same as B87031		agree could have been shallower slope and slightly earlier; as above, moving p-value threshold changes this, but n detected changes in other practices drops substantially: trade-offs.		1
F84717	slightly too early		arguable, noisy series. overall no issue here.		0
J84013	same as B87031		agree could have been shallower slope and slightly earlier; as above, moving p-value threshold changes this, but n detected changes in other practices drops substantially: trade-offs.		1
L83051		could argue better as one drop from month 47 but barely.	very marginal. overall no issue here.		0
L84045		could argue as one drop from month 33.	very marginal. overall no issue here. timing unaffected		0
trimethoprim					
E84077	detects a downward shift immediately after an increase, not really the algorithm misbehaving, but not a very meaningful change		agree, a drop seems fair, but problem with magnitude in particular.		2
E86030	as above	?off-topic but worth noting up-tick.	agree, consider in future whether to drop observations where a large drop follows a large rise.	feasibility test on this?	2
F83665		off-topic but as above wonder if in some use-cases the long second slope is more important.	agree, consider in future whether to prioritise the drop with the largest magnitude.		0
G82113	detected drop based on 2 months?? data followed by a slow increase		odd prescribing behaviour too.		1