BMJ Open

Can process mining automatically describe care pathways of patients with long-term conditions in UK primary care? A study protocol

Ian Litchfield,[1] Ciaron Hoye,[2] David Shukla,[1] Ruth Backman,[1] Alice Turner,[3] Mark Lee,[4] Phil Weber[5]

To cite: Litchfield I, Hoye C, Shukla D, *et al*. Can process mining automatically describe care pathways of patients with long-term conditions in UK primary care? A study protocol. *BMJ Open* 2018;**8**:e019947. doi:10.1136/bmjopen-2017-019947

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-019947).

Received 4 October 2017
Revised 4 October 2018
Accepted 1 November 2018

For numbered affiliations see end of article.

Correspondence to
Dr Ian Litchfield;
i.litchfield@bham.ac.uk

ABSTRACT

Introduction In the UK, primary care is seen as the optimal context for delivering care to an ageing population with a growing number of long-term conditions. However, if it is to meet these demands effectively and efficiently, a more precise understanding of existing care processes is required to ensure their configuration is based on robust evidence. This need to understand and optimise organisational performance is not unique to healthcare, and in industries such as telecommunications or finance, a methodology known as 'process mining' has become an established and successful method to identify how an organisation can best deploy resources to meet the needs of its clients and customers. Here and for the first time in the UK, we will apply it to primary care settings to gain a greater understanding of how patients with two of the most common chronic conditions are managed.

Methods and analysis The study will be conducted in three phases; first, we will apply process mining algorithms to the data held on the clinical management system of four practices of varying characteristics in the West Midlands to determine how each interacts with patients with hypertension or type 2 diabetes. Second, we will use traditional process mapping exercises at each practice to manually produce maps of care processes for the selected condition. Third, with the aid of staff and patients at each practice, we will compare and contrast the process models produced by process mining with the process maps produced via manual techniques, review differences and similarities between them and the relative importance of each. The first pilot study will be on hypertension and the second for patients diagnosed with type 2 diabetes.

Ethics and dissemination Ethical approval has been provided by East Midlands–Leicester South Regional Ethics Committee (REC reference 18/EM/0284). Having refined the automated production of maps of care processes, we can explore pinch points and bottlenecks, process variants and unexpected behaviour, and make informed recommendations to improve the quality and efficiency of care. The results of this study will be submitted for publication in peer-reviewed journals.

Strengths and limitations of the study

- ► This is the first time process mining has been applied to primary care in the UK and it offers a valuable, quantified approach for rapidly and reliably understanding the pathways of patients across large numbers of general practices with the potential to benefit both patient care and optimise service use.
- ► Because healthcare data are notoriously unstructured and clinical processes complex, varied and long running, we will use an iterative approach to data preparation, mining and visualisation, combining machine learning and expert review.
- ► The study is set in four practices of contrasting characteristics to help determine best practice in the use of process mining across the varied primary care setting.
- ► Using orthodox process mapping exercises alongside the data-driven process mining approach means we can identify the differences and similarities of the maps produced by both techniques and refine the process mining algorithms as necessary.

INTRODUCTION

In the UK, primary care is seen as the optimal context for delivering care to an ageing population with a growing number of long-term conditions.[1 2] To do this, it must integrate teams of doctors, nurses and allied staff within high-quality processes.[3] However, if care delivery is to be optimised, a more precise understanding of existing care processes and their consequences is required. In this way, existing systems can be amended and improved based on robust evidence.

In attempting to understand the interaction between health service and patient, numerous improvement methodologies have been employed, among them process mapping, a technique which involves gathering extensive qualitative data from a broad range of service providers and users to map individual processes. First employed in the manufacturing sector[4] to understand the flow of materials and resource that converted raw material into an end product, creating similar process maps in healthcare settings

is logistically challenging and labour intensive, requiring the simultaneous input of a range of clinical and non-clinical staff of varying seniority alongside an equally diverse mix of patients, all with experience of a particular service. Though a relatively effective means of describing a care process or pathway at a single location, such maps are limited by their subjective nature and a lack of quantitative evidence to describe the frequency that specific parts of the process are followed and by which groups of patients. A more precise understanding of these processes built on quantitative evidence and conducted comparatively quickly across a number of practices will enable senior managers and commissioners to make evidence-based decisions on scheduling activities and the allocation of staff and resources. This will help ensure patients enjoy timely and appropriate care and will also support more effective allocation of limited resources to help meet growing demand.

The need to understand and optimise organisational performance is not unique to healthcare. In many industries such as telecommunications, manufacturing and finance, a methodology known as 'process mining'[5–8] has become an established and successful automated method to quickly identify the processes used by an organisation for dealing with its clients and customers. It uses data routinely collated by an organisation's IT systems, containing details on activities, timing and resource, to enable its business processes and organisational structures to be described both visually in the form of a flow chart and formally using mathematical representations.[9–11] It also enables these discovered processes to be objectively compared against management supposition, external requirements for how the processes should operate, or different processes at similar organisations, to find out which configuration of staff and resource most effectively produces the required outcomes.[12 13] By using relevant criteria, the advantages and disadvantages of various configurations can support recommendations for optimising future allocation of resources.[14]

This study is a first step towards employing process mining techniques to understand the complexity of primary care delivery in the UK. We will develop novel algorithms that will automatically produce process models to help senior practice staff and commissioning groups gain a deeper understanding of existing processes of delivering care. To prove that the concept of process mining in primary care works, we will compare the results of our automated process mining with those resulting from orthodox process mapping techniques by comparing the pathways produced by both methods for patients with hypertension or type 2 diabetes mellitus (T2DM), at four practices in the West Midlands.

KNOWLEDGE REVIEW

Process mining in healthcare

There is extensive evidence of how in industry process mining has highlighted inefficiencies in existing organisational processes, for example where quieter sections of the pathway are over-resourced or pinch-points where demand exceeds capacity[15–17]; provided informed simulation of new scenarios, for instance how reallocating resources might affect run-time or process outcome[18]; or identified where tasks could be undertaken by an alternative member of staff such as those with a more appropriate level of seniority or skill set.[19]

More recently, there is a growing body of work applying process mining, also known as careflow mining[20] in this context, to healthcare,[21] although to date few of these studies have been based in the UK.[22] Much of this existing research describes explorative case studies applying process mining in specific secondary care contexts.[23] Previous work applying process mining techniques to emergency care,[24] patients with cardiovascular disease,[25] oncology,[21 23 26 27] T2DM,[28 29] stroke[30] and sepsis[31] have demonstrated that by collating information on the care processes of individual patients with a particular condition, distinctive care pathways can be determined.[28 32] To do this, process mining has used records of the various sequences of events encountered in a care pathway such as consultations, laboratory tests, diagnoses and procedures,[20 23 33–36] alongside related information such as the job title of the healthcare professionals involved at each step.[37] Care processes discovered in this way have provided a well-founded evidence base for investigating patterns of behaviour, testing process improvement and ultimately their effect on patient outcomes.[14 38 39] Previous work by Weber *et al* has employed a principled machine learning theoretical approach,[40] and there is evidence that applying computational optimisation, search or clustering techniques can guide mining and generalisation, or simplify the resulting process models.[41–43]

Process mining in primary care has been studied less frequently than in secondary care, and there have been calls for further research in this setting.[29 44] This context is characterised by a particularly heterogeneous environment consisting of multiple sites that can vary significantly in size, demographics and staff profile with related data potentially sourced from several different systems. Applications to UK contexts and data are particularly rare,[21 23] and to the best of our knowledge, this study will be the first to apply process mining exclusively to primary care datasets in the UK. Our work will focus on processes for treatment of T2DM and hypertension (HT). Process mining has been applied to T2DM,[11 28 45] but only a related technology (Association Rule Mining) has been applied to HT.[46]

Mining complex processes: the 'spaghetti effect'

Routinely collected healthcare data can lack structure and include recording errors, manual data entry or variable levels of detail. The underlying processes are dynamic, complex, multidisciplinary, evolve as medical evidence develops and are frequently ad hoc.[43] In the case of chronic illness, the patients' interaction with the health service lasts for years. Taken together, these characteristics give rise to the problem of so-called 'spaghetti'

models of un-interpretable complexity which contain so many nodes and interconnections that no useful structure or information can be inferred.[20 45 47]

To mitigate for the 'spaghetti effect', a number of techniques are available in each of four aspects of the process of producing a process model, that is, data preparation, data selection, mining and visualisation. At the data preparation stage, aggregation and clustering[36 42] can be used to group low-level events into more abstracted 'event types'.[20] Repeated events may be grouped by time interval[34] or pruned using some other threshold measure.[20] Various methods have also been used to interpret events more accurately[20 48] which can clarify interactions between them, or to group related activities.[11 30] At the data selection stage, the issue of multiple process variants can be dealt with by clustering traces. To achieve this, a number of approaches have proved effective be they data-driven, that is, unsupervised machine learning,[43 49 50] or knowledge-driven (supervised) allocation of traces to process variants.[51–53]

In the process mining phase, probabilistic methods have successfully used a representation which allows the mined model to be interpreted at different levels of aggregation,[54 55] though it is also possible to mine a hierarchical model directly.[56] Nguyen *et al*'s approach was to break the process into 'stages' assuming inherent high-level structure[57]; it is also possible to restrict mining to certain parts of the process, to address specific questions,[23] or to use the extra information provided by clinical results or timing to guide the structure of the mined model.[20 24 58] Ultimately, once models have been produced, visualisation can facilitate interactive control of the level of detail (eg, [10 11 59]), inclusion of expert knowledge[41] or visual effects such as heat maps.[45]

Mining the pathways of chronic disease can also lead to entangled process models where long-running processes have the potential to introduce complex cyclical models as similar sequences of events are repeated with variation as the disease progresses, and difficulty establishing the scope to be included in the mined process. As such cases become apparent, they can be investigated using process analytics methods such as identification of frequent sequences of activities[20 23 29] and change or concept drift detection[60–63] to intelligently extract the significant subvariants of the process.

METHODS AND ANALYSIS

Our study is the first time in the UK that process mining has been used in primary care settings to describe the care processes used by individual practices. We will use process mining techniques to automatically produce process models, describing the pathways used to manage patients with HT. We will also produce process maps at the same practices using orthodox methods and compare and contrast the two.

RESEARCH QUESTION

The overarching aim of our study is to determine whether process mining techniques can be applied to primary care in the UK with its challenges of scale and diversity and to describe best practice in doing so. This includes how they might complement and augment orthodox process mapping methodologies. We plan to meet this aim by fulfilling three key objectives, each corresponding to one of the three phases of the study.

First, we will develop methods and algorithms for creating models of the care processes for treating patients in individual general practitioner (GP) practices using the data routinely collated by each practice within their clinical management system. Second, we will use traditional process mapping exercises involving patients and staff to manually produce maps of HT care processes at the same practices. Third, we will compare and contrast process maps produced via the two different techniques, and compare with staff and patients at the practices where they were derived. We will then repeat the process for patients with T2DM. This will allow us to develop a framework to optimise the use of process mining to automatically describe complex care pathways in primary care. The study will begin in June 2018 and last for 12 months.

RESEARCH DESIGN

Phase I: care process discovery and presentation

In this project, we will use the comprehensive dataset held by the clinical management system (CMS) of each practice. This contains various coded information on patient contact with the service including consultations, diagnoses, prescriptions and laboratory tests. Our initial task is to develop process mining algorithms to determine processes used in a single practice in the management of HT using the standard process mining methodological approach[43 63] which entails data selection and extraction, clustering (aggregating), mining and visualising. In doing so, we will identify the relevant variables needed to define the care of our target patients, identify the corresponding events and select the relevant records. The development of these algorithms will be iterative, and each iteration will be reviewed by our clinical expert (DS) and informatics lead (CH). Once finalised at one practice, these algorithms will then be used to automate the production of process models for the treatment of patients with HT at a further three practices. The graphical presentation of the care process will be based on business process model notation (BPMN).[64] Further details on the application of process mining techniques to healthcare data are contained in the Research methodologies section.

Phase II: creating process maps

We will use proven process mapping techniques to produce process maps that describe the roles of various individuals, and the flow of materials and information required to support care for patients with the target condition. These maps will be developed following process mapping exercises conducted with groups of staff and patients at each practice.[65]

Phase III: comparison of process mining with process mapping

In the final phase, we will present the mined and mapped processes derived from each practice to focus groups consisting of patients and staff from that practice. These focus groups will allow us to explore any differences between the models and the maps, their relative importance and how these algorithms can be further refined.[8 47 66 67] In the future, any comparison with intended pathways may be automated using process conformance methods[8 12] to accurately measure compliance using metrics.

Patient and public involvement

The motivation for the study of using routinely recorded data to improve the efficiency and quality of healthcare processes came from the Clinical Commissioning Group (CCG) who were faced with the task of meeting increasing demand with limited resources yet not having a tool that could readily provide them with detailed information about service provision and use this would require. The concept was discussed with a patient representative with expertise in computer science and who worked as a practice manager so was able to comment on both technological aspects of the work and the potential benefits to both service providers and patients of being able to better understand existing care processes.

RESEARCH METHODOLOGIES

Here, we offer more detail on the three key methodologies we will be using: process mining, process mapping and focus groups.

Process mining

Data requirements

Process mining uses so-called event logs routinely recorded by an organisation's IT systems to learn a model of a business (or clinical) process which indicates what activities can take place, what order they occur in, which sequences of activities may take place simultaneously, or are mutually exclusive, or are repeated. The event log at a minimum records 'events' of a business 'activity' taking place, the time it occurred and what 'case' it belongs to. The concepts of process mining are summarised in table 1 alongside examples from the healthcare environment.

A 'case' collects all activities belonging to a specific instance of the process. In industry, this might be a given invoice or insurance claim. The sequence of recorded events making up a case is known as a process 'trace'. In healthcare, a case will include all events relating to an individual patient and their contact with their practice, possibly restricted to a particular context of interest such as a medication review. Events may also record who the patient was in contact with (eg, practice nurse or GP) and the action undertaken (eg, prescription of medication, blood tests ordered), as described by the SNOMED codes.[68] This dataset allows the production of process models or maps containing information on the patient, clinician, action and location.

Mining processes

Recently, a standard methodology for process mining has emerged which focuses on data preparation, selection and visualisation[63] which we follow in this study. This means we will prepare then inspect log files, apply mining algorithms (to analyse the flow of activities, performance and organisational aspects), present and report results.

Where process mining is being used for the first time in a specific healthcare environment, the recommended approach is exploratory. Initially, we will use existing algorithms (eg, [9 10 69]) as a starting point, explore optimal settings of their so-called 'tuning parameters' (eg, [9 10 40 59]), then refine them as necessary to account for specific characteristics of our data and clinical processes. The setting of tuning parameters

Table 1 Process mining concepts

Concept	Description	Healthcare example
Process	Structured set of activities and connections relating to patients' interactions with a general practice	Patient's regular medication review
Activity	A specific piece of work	Measuring patient's blood levels
Event	An instance of an activity occurring at a specific time	Measuring patient Smith's HbA1c levels at 14:00 1 January 2018
Case	A given instance of a process (eg, for a specific patient)	Medication review for patient Smith
Trace	The recorded events evidencing the activities of a given case	Register, review meds, prescribe drug A, refer for lifestyle advice
Timestamp	Date and time an event occurred	
Resource	Materials, staff or other assets required by an activity	Healthcare assistant with specialist phlebotomy skills
Supplementary information	Additional data may be used to enhance or enrich the process	GP name, practice location, medication dosage

GP, general practitioner.

BOX 1 Steps in developing process mining algorithms

The development of algorithms to discover process models is iterative and involves the following four steps:

1. Apply basic process mining algorithms including Alpha,[91] Heuristics Miner,[9] Inductive Miner[69] and Fuzzy Miner[10] to obtain initial results.
2. Enhance algorithms to enable use of timing and other data to refine the displayed process to optimise the correctness and usefulness of the first iteration maps. Develop clear visualisations based on Weber *et al*'s[13] and Muller and Rogge-Solti's work[75] suitable for clinicians to understand which aspects of the process they focus on, for example, excluding or highlighting detail as required.
3. Review the process maps with experienced stakeholders for explanations of any anomalies, the required level of detail and the ease of use of the algorithms, process representations and visualisations.
4. Refine the data selection and process mining algorithms using knowledge gained in step 3 to produce correct and applicable process maps and trusted automated process mining algorithm. Steps 2 and 3 are then repeated as necessary.

and development of the algorithms is based on iterative interaction with experts at each stage to validate results.[22 31] This allows the identification of problems and limitations arising from (1) erroneous assumptions in interpreting the data; (2) errors in recording the data, indicating a need for further data cleaning; (3) complexity arising from changes to policy or organisational structures during the period covered by the data collected; or (4) process behaviour missed due to, for example, mining from too little data.[7 70] In this way, we can tune the data selection and refine the process mining algorithms, leading to a new and clearer model of the care process, a step that may be repeated several times (see box 1). Once developed, process mining algorithms and tools can be applied to additional datasets held at similar sites to produce process maps in a matter of minutes (eg, [59]). To mitigate any 'spaghetti effect', our focus will be on data pre-processing, data-driven event and trace clustering,[43 71 72] and limited interactive control of the final visualisation, supported by principled machine learning approaches and expert review.

Presentation

In presenting the mined processes to stakeholders, we will use BPMN, a user-friendly and widely accepted graphical language which has previously been used for modelling clinical processes.[50 73–75] A highly simplified example of how a mined process might appear is shown in figure 1 relating to a hypothetical excerpt from the mined process for T2DM. This example illustrates how there may be evidence in a mined model of several variants of underlying process (outlined here by the dashed boxes), as well as unknown activities (represented by the filled boxes in the diagram) which indicate 'noise' in the data.

Process mapping

In the UK and elsewhere, healthcare providers are increasingly relying on process improvement methodologies to streamline production, increase efficiency and minimise waste.[76–78] These methodologies require that existing systems of service provision are thoroughly understood,[79] process maps graphically represent the material and information flows that transform an unhealthy patient into a healthy one.[80] The process is frequently depicted as a series of steps using specified shapes, symbols and colours to provide information on the type of action, the individuals involved and any associated values including metrics such as cycle or wait times. The process maps that result ultimately help identify which inputs and tasks have the greatest impact on the desired output or any areas of waste and delay and so can inform action plans that generate and implement solutions.[81]

Each process mapping exercise involves clinical and non-clinical staff of varying seniority alongside a representative range of patients. They typically take around 2 hours and involve the use of a large sheet of paper containing a horizontal timeline.[65] Participants are then asked to note specific events within the care process (such as booking an appointment or a patient review) and apply these at relevant points across the timeline to create a graphic representation of the process.

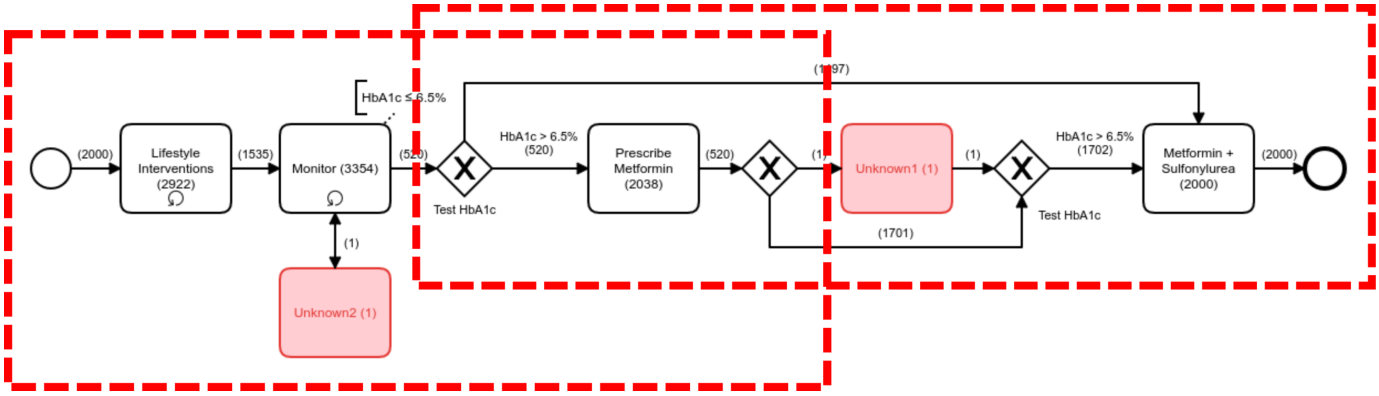

Figure 1 Simplified example of process model from the first iteration of mining from data relating to part of the process for type 2 diabetes mellitus treatment, illustrating common complicating factors (multiple underlying process variants, noisy data) requiring refinement to the mining algorithms and data interpretation.

Table 2 Main file types of CMS data

Variable	Content
Patient demography	1. Practice ID. Patient ID, age, gender, registration date, date left practice and date of death 2. Patient postcode linked area-based socioeconomic, ethnicity, rurality and environmental indices
Clinical data	1. Read coded diagnoses and symptoms, referrals to hospitals and specialists and some free text. Location and date of these events 2. Laboratory results, measurements entered by the practice (blood pressure, weight, tobacco consumption, etc). Date of these events
Prescribing	Prescriptions written by the practice, date issued, formulation, strength, quantity and dosage
Vaccinations	Immunisations carried out at the practice
Consultations	Date, time and duration of consultation
Staff	Role and gender of staff who entered the above data
Practice	Practice ID. Patient list size, linked to number of GPs whole time equivalent, geographical location, Clinical Commissioning Group

CMS, clinical management system; GP, general practitioner.

Focus groups

Focus groups were chosen as the primary method of data collection as the interaction between participants can serve to challenge any over-idealised statements and produce realistic accounts of what people actually do.[82] They also offer an opportunity for participants to reflect and test ideas rather than formulate ideas on the spot and the uninhibited discussion can remind participants and generate new thoughts.[83]

A focus group will be conducted at each practice and will consist of between six and eight participants[84] reflecting a range of clinical and non-clinical staff and patients with first-hand experience of delivering and receiving care for the relevant condition. The groups will be digitally recorded and transcribed verbatim.

Data management and analysis

We will use data from the CMSs of which there are three predominant in the UK: EMIS-Web, SystmOne and InVision. The data they contain are routinely collected and collated and contain information on patient demography, clinical data, time and duration of consultation as well as information on practices and staff (see table 2). Events (prescriptions, referrals, appointments, etc) are recorded with a date, although in some cases more information may be specified. We expect this granularity to be adequate for process mining since we are dealing with patient interactions with a practice over a period of time. If multiple activities are found to occur on the same day, it may be possible to disambiguate these for example by referring to location or practitioner involved.

Data will be selected for a minimum of 24 months to ensure coverage of treatment life cycles (typically up to 12 months). This will include data for an estimated 4000 patients. While individual patients may interact with the services for far longer, using data collected over this time period, considering the number of patients listed at each practice, and the prevalence of the target conditions, we expect to include examples of all variants of treatment patterns.[11 20 23] The pseudonymised data will be sourced via the CCG via the author CH. The data will contain events relating to many processes (eg, treatment of different morbidities for patients at a particular practice), so in order to produce a clear and meaningful process model, it is necessary to focus on events related to the underlying processes, in the first instance for treating patients with HT. We will therefore select patients diagnosed with this condition and identify the relevant data that capture their care, adapting pseudonymisation tools previously used by the CCG in providing similar data for use with BLISS project.[85] In May 2018, the new general data protection regulation comes into effect repealing the previous Data Protection Directive 95/46/EC of 1995. Though built on similar principles, there are nevertheless additional protective measures for personal data used in health-based research and we will ensure that our data permissions reflect the new regulation.[86] We will also account for the recommendations of the Review of data security, consent and opt-outs published by the National Data Guardian.[87]

We will interpret the variables to select (1) event IDs (actions), case IDs (patient IDs) and timestamps relating to HT; and (2) associated data including activity durations, locations, clinicians, test results and medication. Once we have identified the metadata needed, we will extract the relevant information from the CMS. To facilitate this data extraction, we will write code that selects the relevant patient records and fields relating to HT from the databases; this will be pseudonymised and stored on a secure server hosted by the University of Birmingham. Once we have constructed process models and maps for HT, we will repeat the process for patients with T2DM.

Settings and participants

Birmingham Solihull Clinical Commissioning Group (BSOL CCG)

The study will be conducted with BSOL CCG which has the fourth largest population of all CCGs in England with 95 member practices. They are a clinically led organisation, with an annual budget of £1 billion commissioning services for a population of around 710 000 offering fully

integrated, sustainable health and social care and the potential for a large and diverse study.

Recruitment

Coauthor CH is digital lead at BSOL CCG and will assist in identifying and recruiting practices purposively selected to demonstrate maximum variance in terms of characteristics that include size of patient list, socioeconomic environment and number of GPs. There will be four practices involved in the study, and these will be visited in person by a member of the study team where the broader aims of the study and the role and implications of involvement of the individual practices will be discussed with practice staff. Patients will be recruited through clinical staff and via posters in practice waiting rooms to raise awareness of the work and invite their participation. Where possible, other means of communication such as text messages from the practice to patients will be used. Each patient participant will be provided with an information leaflet and consented by a member of the study team.

Process mapping groups will consist of at least one individual from each of the following job categories: General Practitioner, Practice Nurse, Health Care Assistant, Receptionist and Practice Manager. Patients with HT will be invited to join purposively selected to include different ethnic, age and gender groups. The moderator will seek the experiences of both groups of how the current management of HT proceeds.

For the final phase, focus groups will be convened consisting of between six and eight participants from each practice invited to attend from the previous process mapping exercises or recruited using the methods described above.

DISCUSSION

Process mining allows the automatic collation, linkage, analysis and use of routinely collected data. Its use will strengthen the alignment between data analysis and decision-making processes around effective resource use.

Because the CMS dataset contains information on the frequency with which different parts of the process are followed, and the resources involved at each stage, we can describe the weight of traffic across a process and highlight bottlenecks or other areas where resources can be usefully reallocated. We have some experience of using these datasets, the data they contain forms the basis of pseudonymised datasets used in other examples of primary care research[88] and we have also successfully used the data held on the CMS in exploring prescribing behaviours in multimorbid patients in primary care.[89]

The aim of process mining is not merely to gain insights into processes but to use such intelligent analysis to streamline them[90] and to improve patient outcomes.[14 38 39] In the future, it is expected that these models can be used to simulate new scenarios where activities are scheduled differently or resources have been reallocated. This will mean senior practice managers and commissioners can explore the effects of reallocation of resources before introducing any changes in reality.

The study will use pseudonymised and aggregated patient data. The encryption key for this data will be held securely at each practice so that anonymisation is preserved and patient-identifiable data are not stored on University of Birmingham servers. For the focus groups, full informed consent will be obtained by a member of the research team with a Good Clinical Practice certificate, Research Passport, letter of access and any other associated approvals prior to starting. After the focus group has been conducted, participants will have up to 2 weeks to withdraw their data prior to analysis. All data for this study will be held securely, either in a locked cabinet in a secure access building, or on University computers behind a firewall and with appropriate encryption, on backed up servers.

ETHICS AND DISSEMINATION

Our work will be of interest to all those interested in making evidence-based decisions on resource allocation including GP partners, practice managers, commissioning groups and government organisations. As our algorithms will be the first to systematically analyse healthcare processes in primary care in the UK, our findings are expected to be of significant relevance to the service delivery, informatics and process improvement academic communities. A favourable ethical opinion was provided by East Midlands–Leicester South Research Ethics Committee (REC reference 18/EM/0284).

We will publish peer-reviewed articles in high-impact healthcare and informatics journals as well as generic trade journals such as *Practice Management* and *The Pulse* to disseminate our findings to health service managers in primary care. This process will be bolstered by an online presence using a bespoke website and social networking pages such as Facebook and Twitter to promote and disseminate our work to the wider public.

Our findings will be presented at national and international healthcare conferences focused on the quality and safety of healthcare and process mining and at bioinformatics conferences. The impact of our work is enhanced by the close partnership with BSOL CCG and their commitment to explore the use of process mining to inform strategic guidance and recommendations for the optimal allocation of resources across the CCG, and within individual practices appropriate to the needs and preferences of their patients.

Author affiliations
[1]Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK
[2]Digital Transformation, Birmingham Solihull Clinical Commissioning Group, Birmingham, UK
[3]University Hospitals Birmingham NHS Foundation Trust and Institute of Applied Health Research, University of Birmingham, Birmingham, UK
[4]School of Computer Science, College of Engineering and Physical Sciences, University of Birmingham, Birmingham, UK

[5]School of Engineering and Applied Science, System Analytics for Innovation, Aston University, Birmingham, UK

Contributors IL, PW and CH were responsible for the conception of the work and the design of the study. IL led the drafting of the article with input from PW, CH and DS. ML, AT, CH, DS and RB all provided critical revisions. The final version was drafted by IL and PW and approved by AT, RB, ML, CH and DS.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Barnett K, Mercer SW, Norbury M, *et al*. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 2012;380:37–43.
2. Bodenheimer T, Lorig K, Holman H, *et al*. Patient self-management of chronic disease in primary care. *JAMA* 2002;288:2469–75.
3. Royal College of General Practitioners. GP forward view: interim assessment. 2017 http://www.rcgp.org.uk/-/media/Files/Policy/2017/RCGP-GP-Forward-View-Interim-assessment-2017.ashx?la=en.
4. Rother M, Shook J. *Learning to see: Value stream mapping to create value and eliminate MUDA*. Lean Enterprise Institute: Brookline, MA, 2003.
5. Process mining. We use cookies, just to track visits to our website, we store no personal details. http://www.processmining.org/ (accessed July 2017).
6. van der Aalst WMP, Adriansyah A, de Medeiros AKA, *et al*. Process mining manifesto. In: Florian D, Kamel B, Schahram D, eds. *Business Process Management Workshops. BPM 2011. Lecture Notes in Business Information Processing*. Berlin: Springer, 2012:164–94.
7. Weber P. *A framework for the analysis and comparison of process mining algorithms. PhD thesis*: University of Birmingham, 2014.
8. van der Aalst WMP. *Process mining: discovery, conformance and enhancement of business processes*. New York: Springer, 2011.
9. Weijters A, Ribeiro JTS. *Flexible Heuristics Miner (FHM). Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2011, part of the IEEE Symposium Series on Computational Intelligence*. Paris, France, 2011:310–7.
10. Gunther CW, van der Aalst WMP. *Fuzzy mining—adaptive process simplification based on multi-perspective metrics. International Conference on Business Process Management*. Brisbane, Australia: BPM 2007, Business Process Management, 20072007:328–43.
11. Klimov D, Shknevsky A, Shahar Y. Exploration of patterns predicting renal damage in patients with diabetes type II using a visual temporal analysis laboratory. *J Am Med Inform Assoc* 2015;22:275–89.
12. Rozinat A, van der Aalst WMP. Conformance checking of processes based on monitoring real behavior. *Information Systems* 2008;33:64–95.
13. Weber P, Taylor PN, Majeed B, *et al*; *Comparing complex business process models. IEEE international conference on industrial engineering and engineering management, IEEM 2012*. Hong Kong, China, 2012.
14. Schonenberg H, Weber B, van Dongen F, *et al*. *Supporting flexible processes through recommendations based on history. LNCS 5240*: Springer, 2008:51–66.
15. Mans R, van der Aalst WMP, Vanwersch RJB. Process mining in healthcare—evaluating and exploiting operational healthcare processes. *Springer briefs in business process management*: Springer International Publishing, 2015.
16. Suriadi S, Ouyang C, van der Aalst WMP, *et al*. Event interval analysis: why do processes take time? *Decis Support Syst* 2015;79:77–98.
17. Adriansyah A, Buijs JC. A mining process performance from event logs. In: La Rosa M, Soffer P, eds. *Business Process Management Workshops. BPM 2012. Lecture Notes in Business Information Processing*. . Berlin: Springer, 2012:132. 217–8.
18. van der Aalst WMP, Schonenberg MH, Song M. Time prediction based on process mining. *Inf Syst* 2011;36:450–75.
19. Low WZ, vanden Broucke SKLM, Wynn MT, *et al*. Revising history for cost-informed process improvement. *Computing* 2016;98:895–921.
20. Dagliati A, Sacchi L, Zambelli A, *et al*. Temporal electronic phenotyping by mining careflows of breast cancer patients. *J Biomed Inform* 2017;66:136–47.
21. Rojas E, Munoz-Gama J, Sepúlveda M, *et al*. Process mining in healthcare: a literature review. *J Biomed Inform* 2016;61:224–36.
22. Baker K, Dunwoodie E, Jones RG, *et al*. Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. *Int J Med Inform* 2017;103:32–41.
23. Caron F, Vanthienen J, Vanhaecht K, *et al*. Monitoring care processes in the gynecologic oncology department. *Comput Biol Med* 2014;44:88–96.
24. Fernandez-Llatas C, Valdivieso B, Traver V, *et al*. *Using process mining for automatic support of clinical pathways design*. In: Fernandez-Llatas C, García-Gómez JM, eds. *Data mining in clinical medicine, no. 1246*. New York: Springer, 2015:79–88.
25. Fernandez-Llatas C, Bayo JL, Martinez-Romero A, *et al*. Interactive pattern recognition in cardiovascular diseases management. A process mining approach. *Proceedings of the IEEE international conference on biomedical and health informatics*. Las Vegas: EEUU, 2016.
26. Sacchi L, Segagni D, Dagliati A, *et al*. Mining careflow patterns in data warehouses of breast cancer patients. *Proc American Medical Informatics Association Annual Symposium (AMIA 2013)*. Washington DC, USA, 2013.
27. Kurniati AP, Hall G, Hogg D, *et al*. *Process mining in oncology using the MIMIC-III dataset*. Bandung, Indonesia: Data and Information Science (ICoDIS).
28. Dagliati A, Sacchi L, Cerra C, *et al*. *IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2014*. Valencia, Spain, 2014:240–3.
29. Lismont J, Janssens AS, Odnoletkova I, *et al*. A guide for the application of analytics on healthcare processes: a dynamic view on patient pathways. *Comput Biol Med* 2016;77:125–34.
30. Montani S, Striani M, Quaglini S, *et al*. *Knowledge-based trace abstraction for semantic process mining. LNCS 10259 LNAI:267–271, 2017, Artificial Intelligence in Medicine—16th Conference on Artificial Intelligence in Medicine*: AIME, 2017.
31. Mannhardt F, Blinde D. Analyzing the trajectories of patients with sepsis using process mining. CEUR 1859:72–80, 2017, Joint Proc Radar Tracks at the 18th BPMDS 2017 &c., co-located with the 29th CAiSE. 2017.
32. Meyer G, Adomavicius G, Johnson PE, *et al*. A machine learning approach to improving dynamic decision making. *Information Systems Research* 2014;25:239–63.
33. Fernandez-Llatas C, Lizondo A, Monton E, *et al*. Process mining methodology for health process tracking using real-time indoor location systems. *Sensors* 2015;15:29821–40.
34. Alharbi A, Bulpitt A, Johnson O. *Improving pattern detection in healthcare process mining using an interval-based event selection method*: Lecture Notes in Business Information Processing, 2017:88–105.
35. Huang Z, Dong W, Ji L, *et al*. Discovery of clinical pathway patterns from event logs using probabilistic topic models. *J Biomed Inform* 2014;47:39–57.
36. Mannhardt F, Leoni M, Reijers HA, *et al*. From low-level events to activities—a pattern-based approach. BPM, 14th International Conference, Proc. *LNCS* 2016;9850:125–41.
37. Suriadi S, Mans RS, Wynn MT, *et al*. Measuring patient flow variations: a cross-organisational process mining approach lnbip 181:43-58, 2014, Asia Pacific Business Process Management—2nd Asia Pacific Conference, AP-BPM. 2014.
38. Peleg M, Soffer P, Ghattas J. *Mining process execution and outcomes—position paper*: BPM 2007 International Workshops, 2008:395–400.
39. Lakshmanan GT, Rozsnyai S, Wang F. *Investigating clinical care pathways correlated with outcomes*. LNCS: BPM 2013. Proceedings.
40. Weber P, Bordbar B, Tiňo P. A principled approach to mining from noisy logs using Heuristics Miner. *Proc IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2013:119–26.
41. Canensi L, Leonardi G, Montani S, *et al*. Multi-level interactive medical process mining conference on Artificial Intelligence in Medicine in Europe (AIME). 2017;260:2017.
42. Prodel M. *Process discovery, analysis and simulation of clinical pathways using healthcare data. PhD thesis*, 2017.
43. Rebuge A, Ferreira DR. *Business process analysis in healthcare environments: a methodology based on process mining, information systems*. : Elsevier, 2012:37: 99–116.

44. Zhou Z, Wang Y, Li L. Process mining based modeling and analysis of workflows in clinical care—a case study in a Chicago outpatient clinic. *Proc 11th IEEE Int'l conf Networking, Sensing and Control, ICNSC 2014*, 2014:590–5.
45. Fernandez-Llatas C, Martinez-Millana A, Martinez-Romero A, *et al*. Diabetes care related process modelling using Process Mining techniques. Lessons learned in the application of Interactive Pattern Recognition: coping with the Spaghetti Effect. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2015:2127–30.
46. Shin AM, Lee IH, Lee GH, *et al*. Diagnostic analysis of patients with essential hypertension using association rule mining. *Healthc Inform Res* 2010;16:77–81.
47. van der Aalst WMP. Process mining: discovering and improving spaghetti and lasagna processes. *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2011, part of the IEEE Symposium Series on Computational Intelligence 2011*. Paris, France, 20112011:13–20.
48. Lu X, Dirk F, van den Biggelaar F, *et al*. Handling duplicated tasks in process discovery by refining event labels. *LNCS* 2016;9850:90–107.
49. Delias P, Doumpos M, Grigoroudis E, *et al*. Supporting healthcare management decisions via robust clustering of event logs. *Knowledge-Based Systems* 2015;84:203–13.
50. Lu F, Zeng Q, Duan H. Synchronization-core-based discovery of processes with decomposable cyclic dependencies. *ACM Trans Knowl Discov Data* 2016;10:1–29.
51. Caron F, Vanthienen J, Vanhaecht K, *et al*. A process mining based investigation of adverse events in care processes. 2015;43:16–25.
52. Bose RP, Chandra J, Van Der Aalst WMP. *Analysis of patient treatment procedures. LNCS 99:165–166, 2012, BPM*. International Workshops, Revised Selected Papers, 2011.
53. Stefanini A, Aloini D, Dulmin R, *et al*. *Linking diagnostic-related groups (DRGs) to their processes by process mining*: BIOSTEC 2016, Proc. HealthInf, 2016:438–43.
54. Zhang Y, Padman R, Patel N. Paving the COWpath: learning and visualizing clinical pathways from electronic health record data. *J Biomed Inform* 2015;58:186–97.
55. Blum T, Padoy N, Feußner H, *et al*. Workflow mining for visualization and analysis of surgeries. *IntJ of Com Assisted Radiology and Surgery* 2008;3:379–86.
56. Bose RP, Chandra J, Verbeek E, *et al*. *Discovering hierarchical process models using ProM. LNBIP 107:33–48, 2012, IS Olympics: Information Systems in a Diverse World—CAiSE Forum*, 2011. Selected Extended Papers.
57. Nguyen HH, Dumas M, Hofstedeter AHM, *et al*. *Mining business process stages from event logs. In 29th CAiSE*. Essen, Germany, 2017.
58. Kaymak U, Mans R, Van De Steeg T, *et al*. *IEEE International Conference on Systems, Man and Cybernetics*, 2012:1859–64.
59. Gunther CW, Rozinat A. Disco: Discover Your Processes. BPM(Demos). 2012:40–4.
60. Weber P, Tiňo P, Bordbar B. *Process mining in non-stationary environments. ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges, Belgium: ESANN, 2012.
61. Bose RP, van der Aalst WMP, Zliobaite I, *et al*. Handling concept drift in process mining. *Proc. CAiSE*, 2011:391–405.
62. Hompes BFA, Buijs J, van der Aalst WMP, *et al*. Detecting changes in process behavior using comparative case clustering. LNBIP 244:54–75, Proc. Data-Driven Process Discovery and Analysis—5th IFIP WG 2.6 International Symposium, SIMPDA 2015, Revised Selected Papers.
63. Bozkaya M, Gabriels J, Werf J. *Process diagnostics: a method based on process mining. In: International Conference on Information, Process, and Knowledge Management, 2009*: eKNOW'09, IEEE, 2009:22–7.
64. OMG. Business Process Model and Notation (BPMN) Technical Report formal/2011-01-03, OMG. 2011.
65. The King's Fund. Patient and family centred care toolkit. https://www.kingsfund.org.uk/projects/pfcc/process-mapping (accessed Jul 2017).
66. Adriansyah A, Munoz-Gama J, Carmona J, *et al*. Alignment based precision checking. *Business process management workshops*: Springer, 2013:137–49.
67. van Eck ML, Lu X, Leemans SJJ, *et al*. *PM 2: a process mining project methodology. advanced information systems engineering—27th International conference. CAiSE*. Stockholm, Sweden, 20152015:297–313.
68. SNOMED International. SNOMED International determines global standards for health terms, an essential part of improving the health of humankind. http://www.snomed.org/snomed-ct (Accessed July 2017).
69. Leemans SJJ, Fahland D, van der Aalst WMP. Discovering block-structured process models from event logs—a constructive approach. In: Colom JM, Desel J, eds. *Application and theory of petri nets and concurrency—34th international conference. PETRI NETS 2013*. . Milan, Italy: Springer, 2013:2013. 311–29.
70. Weber P, Bordbar B, Tino P. A framework for the analysis of process mining algorithms. *IEEE Trans Syst Man Cybern* 2013;43:303–17.
71. Greco G, Guzzo A, Pontieri L, *et al*. Discovering expressive process models by clustering log traces. *IEEE Transactions on Knowledge and Data Engineering* 2006;18:1010–27.
72. Diogo R. *Approaching process mining with sequence clustering: experiments and findings. In Proc. Business Process Management, 5th International Conference, BPM 2007*. Brisbane, Australia: LNCS 4714:360–374, Springer, 2007.
73. Scheuerlein H, Rauchfuss F, Dittmar Y, *et al*. New methods for clinical pathways—Business Process Modeling Notation (BPMN) and Tangible Business Process Modeling (t.BPM). *Langenbecks Arch Surg* 2012;397:755–61.
74. Rolon E, Aguilar ER, Garcia F, *et al*. Process modeling of the health sector using BPMN: a case study. *Proc First International Conference on Health Informatics. HEALTHINF 2008: Funchal, Portugal*, 2008:173–8.
75. Muller R, Rogge-Solti A. *BPMN for healthcare processes. 3rd Central-European workshop on services and their composition, services und ihre komposition. ZEUS 2011*. Karlsruhe, Germany, 2011:65–72.
76. Lummus RR, Vokurka RJ, Rodeghiero B. Improving quality through value stream mapping: a case study of a physician's clinic. *Total Quality Management & Business Excellence* 2006;17:1063–75.
77. NHS Institute for Innovation and Improvement. Improvement leaders' guide. *Process mapping, analysis and redesign: general improvement skills*. NHS England, 2005.
78. Teichgräber UK, de Bucourt M. Applying value stream mapping techniques to eliminate non-value-added waste for the procurement of endovascular stents. *Eur J Radiol* 2012;81:e47–e52.
79. McLaughlin N, Rodstein J, Burke MA, *et al*. Demystifying process mapping: a key step in neurosurgical quality improvement initiatives. *Neurosurgery* 2014;75:99–109.
80. Chen ET, Eder M, Elder NC, *et al*. Crossing the finish line: follow-up of abnormal test results in a multisite community health center. *J Natl Med Assoc* 2010;102:720–5.
81. Baker M, Taylor I. *Making hospitals work*. Herefordshire: Lean Enterprise Academy, 2009.
82. Morgan DL. Future directions in focus group research. *Successful Focus Groups. London: Sage* 1993.
83. Gill P, Stewart K, Treasure E, *et al*. Methods of data collection in qualitative research. *British Dental Journal* 2008;204:291–5.
84. Silverman D. *Doing qualitative research: a practical handbook*. London: Sage, 2000.
85. University of Birmingham. Birmingham Lung Improvement StudieS (BLISS). 2017 http://www.birmingham.ac.uk/research/activity/mds/projects/HaPS/PHEB/BLISS/index.aspx
86. Chassang G. The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience* 2017;11:709.
87. National Data Guardian for Health and Care. Review of data security, consent and Opt-Outs. 2017 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/535024/data-security-review.PDF.
88. THIN. The Health Improvement Network (THIN). https://www.visionhealth.co.uk/portfolio-items/the-health-improvement-network-thin/ (accessed Jun 2017).
89. Backman R, Weber P, Turner AM, *et al*. Assessing the extent of drug interactions among patients with multimorbidity in primary and secondary care in the West Midlands (UK): a study protocol for the Mixed Methods Multimorbidity Study (MiMMS). *BMJ Open* 2017;7:e016713.
90. De Weerdt J, Caron F, Vanthienen J, *et al*. Getting a grasp on clinical pathway data: an approach based on process mining. LNCS 7769 LNAI: 22–35, 2013, Emerging Trends in Knowledge Discovery and Data Mining—PAKDD 2012 International Workshops: DMHM, GeoDoc, 3Clust, and DSDM, Revised Selected Papers. 2012.
91. van der Aalst W, Weijters T, Maruster L. Workflow mining: discovering process models from event logs. *IEEE Trans Knowl Data Eng* 2004;16:1128–42.