Title:

Process Mining Routinely Collected Electronic Health Records to Define

Real-Life Clinical Pathways during Chemotherapy

Running title:

Mining patient pathways during chemotherapy

*Karl Baker[a], *Elaine Dunwoodie[b,c], Richard G Jones[d], Alex Newsham[e], Owen Johnson[a,f], Christopher P Price[g], Jane Wolstenholme[h], Jose Leal[h], Patrick McGinley[i], Chris Twelves[b,c] and Geoff Hall[b,c].

*Joint first authors.

[a] X-Lab Ltd, Joseph's Well, Hanover Walk, Leeds, LS3 1AB, UK

- karl.baker@x-labystems.co.uk / software developer

[b] Leeds Institute of Cancer and Pathology, Level 4, Bexley Wing, St James's University Hospital, Beckett Street, Leeds, LS9 7TF.

- e.h.dunwoodie@leeds.ac.uk / clinical research fellow and specialty trainee in medical oncology
- c.j.twelves@leeds.ac.uk / Professor of clinical cancer pharmacology and oncology

- g.hall@leeds.ac.uk / senior lecturer and honorary consultant medical oncologist

[c] Leeds Teaching Hospitals NHS Trust, Level 4, Bexley Wing, St James's University Hospital, Beckett Street, Leeds, LS9 7TF.

[d] Yorkshire Centre for Health Informatics, Leeds Institute of Health Sciences, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT.

- Richard Jones, Professor of chemical pathology and health informatics, is deceased and died prior to submission of the manuscript.

[e] School of Medicine, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT.

- Alex.Newsham@bthft.nhs.uk / research fellow and data scientist

[f] School of Computing, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT.

- o.a.johnson@leeds.ac.uk / senior fellow

[g] Nuffield Department of Primary Care Health Sciences, University of Oxford, New Radcliffe House, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG.

- cpprice1@gmail.com / honorary senior fellow in clinical biochemistry

[h] Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, OX3 7LF.

- jane.wolstenholme@dph.ox.ac.uk / senior health economist
- jose.leal@dph.ox.ac.uk / university research lecturer

[i] Maidstone and Tunbridge Wells NHS Trust, Tonbridge Road, Tunbridge Wells, TN2 4QJ.

- pmcginley@nhs.net / Head of costing & service line reporting

Corresponding author:

Dr Elaine Dunwoodie

Clinical Research Fellow & Speciality Trainee in Medical Oncology, Leeds institute of Cancer and Pathology & Leeds Teaching Hospitals NHS Trust, Level 4, Bexley Wing, St James's University Hospital, Beckett Street, Leeds, LS9 7TF. Email: e.h.dunwoodie@leeds.ac.uk

Key words

Electronic Health Records, Process Mining, Episode of care, Care Pathways, Drug Therapy, Neoplasms.

Abstract

Background

There is growing interest in the use of routinely collected electronic health records to enhance service delivery and facilitate clinical research. It should be possible to detect and measure patterns of care and use the data to monitor improvements but there are methodological and data quality challenges. Driven by the desire to model the impact of a patient self-test blood count monitoring service in patients on chemotherapy, we aimed to (i) establish reproducible methods of process-mining electronic health records, (ii) use the outputs derived to define and quantify patient pathways during chemotherapy, and (iii) to gather robust data which is structured to be able to inform a cost-effectiveness decision model of home monitoring of neutropenic status during chemotherapy.

Methods

Electronic Health Records at a UK oncology centre were included if they had (i) a diagnosis of metastatic breast cancer and received adjuvant epirubicin and cyclosphosphamide chemotherapy or (ii) colorectal cancer and received palliative oxaliplatin and infusional 5-fluorouracil chemotherapy, and (iii) were first diagnosed with cancer between January 2004 and February 2013. Software and a Markov model were developed, producing a schematic of patient pathways during chemotherapy.

Results

Significant variance from the assumed care pathway was evident from the data. Of the 535 patients with breast cancer and 420 with colorectal cancer there were 474 and 329 pathway variants respectively. Only 27 (5%) and 26 (6%) completed the planned six cycles of chemotherapy without having unplanned hospital contact. Over the six cycles, 169 (31.6%)

patients with breast cancer and 190 (45.2%) patients with colorectal cancer were admitted to hospital.

Conclusion

The pathways of patients on chemotherapy are complex. An iterative approach to addressing semantic and data quality issues enabled the effective use of routinely collected patient records to produce accurate models of the real-life experiences of chemotherapy patients and generate clinically useful information. Very few patients experience the idealised patient pathway that is used to plan their care. A better understanding of real-life clinical pathways through process mining can contribute to care and data quality assurance, identifying unmet needs, facilitating quantification of innovation impact, communicating with stakeholders, and ultimately improving patient care and outcomes.

Introduction

Over the last 30 years, Western medicine has become increasingly systematised (1, 2). Drivers bringing this about include a call for improvements in quality, the desire to standardise care, a need to disseminate best practice, demands for cost containment coupled with rising demand and not least, patient safety (3). An important tool in this regard has been clinical pathways that establish the expected standard of care and the processes and procedures that should be followed (4). These pathways are becoming more evidence based (5, 6), but such evidence is generally derived from carefully structured clinical trials rather than learning directly from routine clinical practice which can be very different. Moreover, the majority of such clinical pathways are linear, relate to single conditions, and do not reflect real-life variations in individual patient genotype, phenotype, co-morbidities, environment and response to treatment. There are, therefore, good reasons to question whether perceived patterns of existing care are accurate and meaningful when used to plan and evaluate patient management.

With the trend towards integrated care and electronic health record (EHR) systems, a wealth of data on what actually happens to patients during their episodes of care are potentially available. In oncology, there is increasing attention on the use of “big data” (7), with much of the focus being on applications of next generation sequencing genomics (8). Process mining is one emerging “big data” approach for discovering and analysing process models based on the very large event logs contained within information systems (9) and there is a growing body of literature on process mining in healthcare (10). Our review of the process mining literature identified 37 peer reviewed papers using electronic health record

(EHR) data to map pathways in oncology (11). Challenges included the difficulty of capturing outpatient events (12), the use of data collected for non-clinical purposes, missing data and clinically inaccurate time-stamps (13). Our chemotherapy pathway mapping addressed these issues and is comprehensive and patient-centred, reporting most contacts the patient has with the hospital including in- and outpatient events, telephone contacts and pathology results. We use a novel iterative approach to addressing data quality through using clinical review to refine the model. We argue that mining true real-life clinical pathways of patients with cancer will facilitate the visualisation, quantification and improvement of such pathways.

The United Kingdom (UK) is in a strong position to address the issue of acquiring and applying lessons learned to improve complex patient pathways and outcomes due to the comprehensive nature of the National Health Service (NHS) model which extends across primary and secondary care with equity of access for all. The challenges were identified by the National Confidential Enquiry into Patient Outcome and Death (NCEPOD) in 2008, which reported that 42% of patients who died within thirty days of systemic anticancer therapy were admitted with a treatment related complication to a general medical ward, rather than to a ward with oncology-specific expertise (14). This identified problems at a high level with limited amount of detail. The NCEPOD and the National Chemotherapy Advisory Group highlighted the incompleteness of patient records, many of which were paper-based at the time, along with shortcomings of conventional real-life data collection (15). In the UK, EHRs are now commonly used in routine practice to collect comprehensive clinical data in all patients, at all points of care. We used a large and comprehensive EHR to study clinical pathways in patients with cancer receiving chemotherapy and their risk of developing neutropenic infections in order to get a more detailed understanding of neutropenic complications during chemotherapy.

We describe a methodology to extract detailed information on individual patient care pathways from Patient Pathway Manager (PPM), a mature, large and comprehensive EHR which has been in place for patients with cancer for over a decade in Leeds Teaching Hospitals NHS Trust (LTHT) (16, 17). We studied patients with two common malignancies, breast and colorectal cancer, receiving two frequently used chemotherapy regimens; adjuvant epirubicin and cyclophosphamide (EC90) and palliative oxaliplatin and modified de Gramont 5-fluorouracil (OxMdG). The key objectives of this study were (i) to establish reproducible methods of mining EHRs, (ii) to define accurate clinical pathways of patients with cancer on chemotherapy, and (iii) to gather robust data that is structured to inform a cost-effectiveness decision model of home monitoring of neutropenic status during chemotherapy.

Materials & Methods

Patient Pathway Manager (PPM)

PPM is a mature EHR that is used to capture comprehensive clinical data on all patients undergoing treatment for cancer at the Leeds Cancer Centre within LTHT. Initially developed in 2003 to support collection of the National Cancer Dataset, the system has been extended to collect comprehensive coded data to support the collection and reporting of the Cancer Outcomes Services Dataset (18). The system integrates electronic data held within multiple disparate systems within the Trust into a single EHR database including data on patient admissions and out-patient events from the Patient Administration System (PAS), chemotherapy data from Chemocare (19), radiotherapy data from Mosaiq (20), blood, pathology and microbiology results from laboratory systems and data entered directly into PPM including surgery, cancer waiting times and multi-disciplinary team meetings (Figure 1). The system has recently been extended to collate data on all patients whose care has been delivered at LTHT since 2000 and contains the records of more than 2.39 million patients. It is therefore one of the largest hospital EHRs in the UK. The PPM system now underpins the move away from paper records at LTHT and also provides the platform for the Leeds Care Record, a common integrated digital care record used across the Leeds City Region for primary, secondary/tertiary care, social care, community care and mental health services. Within PPM, each event is recorded with a patient identification and metadata, including date time stamp, enabling a historical pathway to be extracted for every patient.

Patient Selection Criteria

Patient records were included if the patient had (i) a diagnosis of breast cancer (ICD-10 C50) and received epirubicin and cyclophosphamide (EC90) chemotherapy as adjuvant treatment or (ii) a diagnosis of colorectal cancer (ICD-10 C18-C20) and received oxaliplatin and infusional 5-fluorouracil chemotherapy (OxMdG) in the palliative setting. Breast and colo-rectal cancer were selected as both are common cancers in which chemotherapy plays an important role. The breast cancer setting was adjuvant, whereas the patients with colo-rectal cancer were being treated with palliative intent; we sought, therefore, to select populations with potentially different challenges. The chemotherapy regimens chosen represented those most commonly used in Leeds Cancer Centre in the specified time period in each setting. Patients first diagnosed with the cancer between 1st January 2004 and 1st February 2013 were included; 535 with breast cancer and 420 with colorectal cancer.

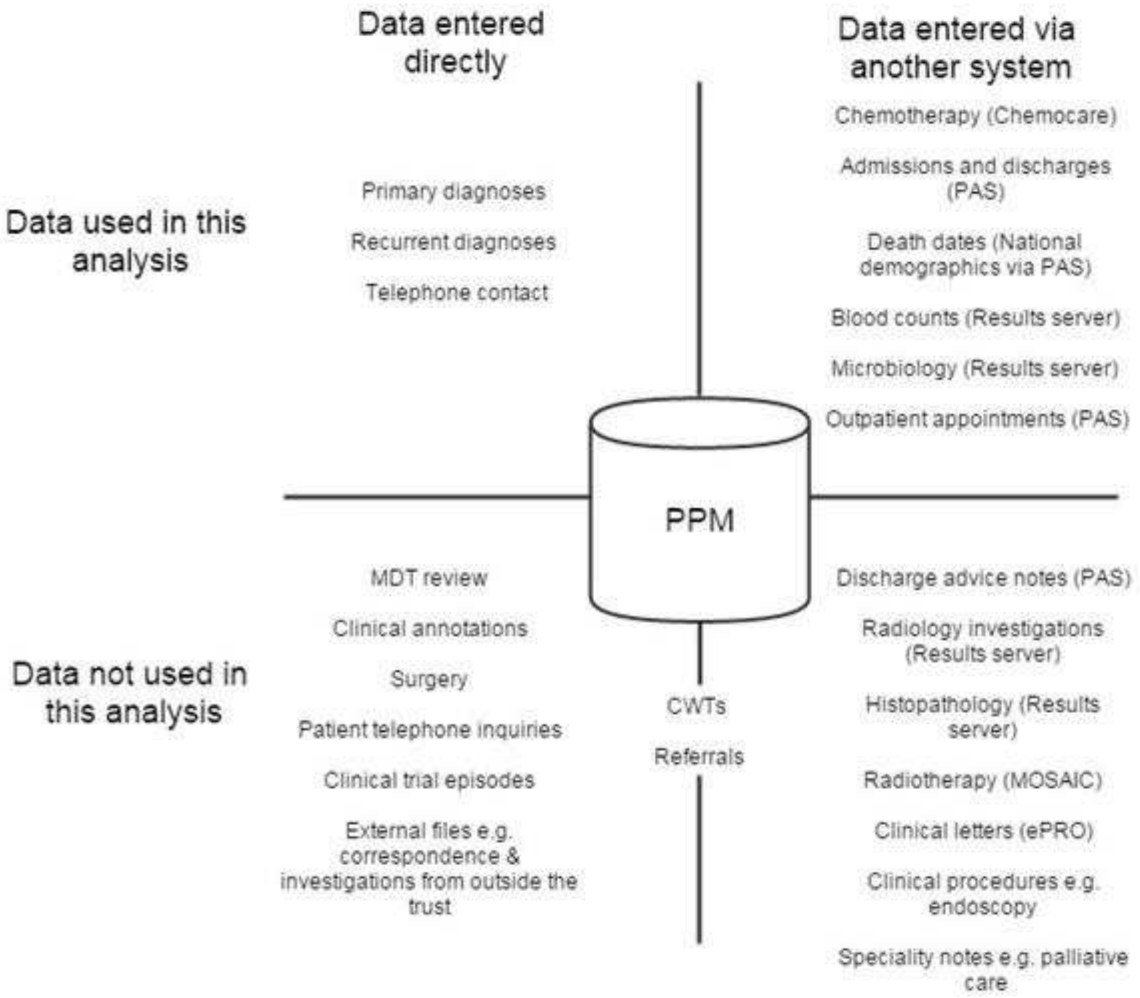
Data entered directly
Data entered via another system
Data used in this analysis
Primary diagnoses
Recurrent diagnoses
Telephone contact
Chemotherapy (Chemocare)
Admissions and discharges (PAS)
Death dates (National demographics via PAS)
Blood counts (Results server)
Microbiology (Results server)
Outpatient appointments (PAS)
PPM
Data not used in this analysis
MDT review
Clinical annotations
Surgery
Patient telephone inquiries
Clinical trial episodes
External files e.g. correspondence & investigations from outside the trust
CWTs
Referrals
Discharge advice notes (PAS)
Radiology investigations (Results server)
Histopathology (Results server)
Radiotherapy (MOSAIC)
Clinical letters (ePRO)
Clinical procedures e.g. endoscopy
Speciality notes e.g. palliative care

Figure 1: Overview of data entered into PPM.

MDT - multi-disciplinary team, Chemocare - chemotherapy electronic prescribing system, PAS - patient administration system, Results server - LTHT software importing all patient investigation results from multiple disciplines, CWT - cancer waiting times, MOSAIC - radiation oncology software, ePRO - clinical information management system working with Winscribe's digital dictation system, telephone contact – hospital staff telephone patient, patient telephone inquiries – patient rings hospital with a clinical inquiry.

Computing Resources

An information technology data analyst developed the data extraction and transformation software and developed the pathway model as described in the next section (Data mining software & Markov model development). Data in PPM is stored in a Microsoft SQL database and records for the two patient groups were extracted using C# with embedded SQL queries to create a secure Microsoft SQL database containing all relevant events. The software followed the extract, transform and load (ETL) design pattern (21). Transformations to the data were coded to address completeness, quality, exclusion and aggregation issues and followed a change control process with clinical oversight. The data was used to generate Comma Separated Value (CSV) exports of aggregate figures and animated visualisations of a Markov model. The key elements of Markov models are described in Appendix A. The software was run on a secure LTHT machine with an Intel Core I7 processor with 16GB of memory.

Data Mining Software & Markov Model Development

Data modelling followed a disciplined agile development process with eight iterations in total (22). The first iteration started with a simplified model schematic of the patient pathway based on expert opinion (Figure 2); this was followed by the development of the initial data mining software that extracted only the Markov events in the patient pathway defined by the initial model. Rules used to transform clinical events into model states and a more detailed description of the whole process are described in Appendix B.

Software was developed to extract patient records from the main PPM dataset using database queries and segment these by regimen type and cycle number. A randomly generated integer was used to identify patients for pseudonymisation. Each patient was allocated to particular states in the model by applying a data mining process in two main stages. Firstly, events were extracted for each patient from PPM and transformed into a single table, ordered by date and time. Creation of the pathway of events for each patient is described in detail in Appendix B. Secondly, this pathway of events was matched to a pathway of states in the model described below (Figure 2). A model state reflects a cluster of patient events happening within a small time span. In Figure 2, chemotherapy delivery is the point at which the pathway starts, repeats and represents state S1 in the model.

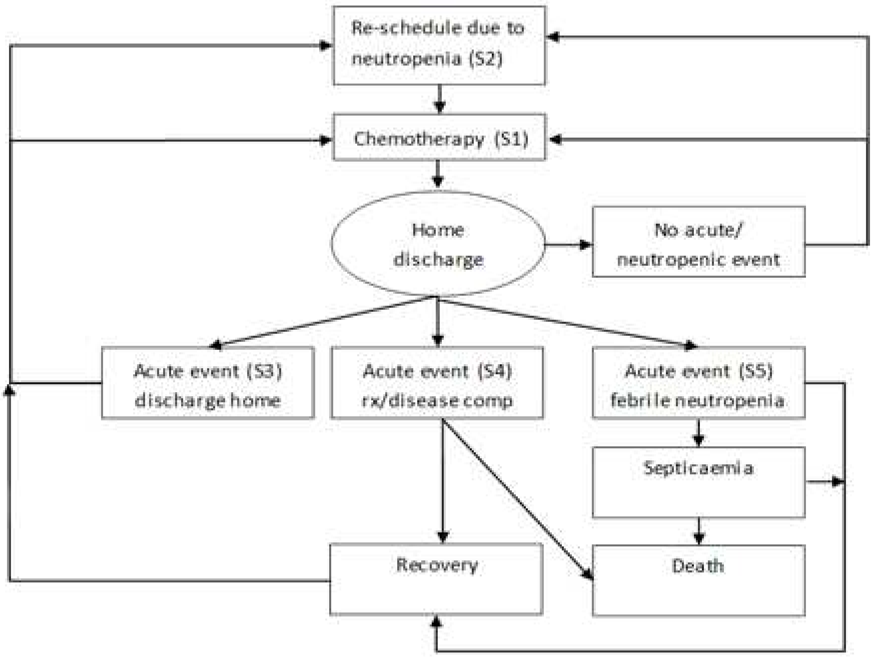
Re-schedule due to neutropenia (S2)
Chemotherapy (S1)
Home discharge
No acute/ neutropenic event
Acute event (S3) discharge home
Acute event (S4) rx/disease comp
Acute event (S5) febrile neutropenia
Septicaemia
Recovery
Death

Figure 2: The patient pathway model at the beginning of the first iteration.

S1, patient presents and the chemotherapy goes ahead as normal. S2, patient attends hospital for chemotherapy but it is re-scheduled/delayed due to neutropenia. S3, patient is admitted acutely to hospital with problems unrelated to their cancer or treatment. S4, patient is admitted acutely to hospital with problems related to their cancer or treatment (not neutropenic). S5, Patient is admitted acutely to hospital with problems related to their cancer or treatment (neutropenic).

There were two main steps in each successive data model iteration (Figure 3). First the model was redefined based on the results of data mining from the previous iteration, in order to account for complexity revealed by the data that was not previously foreseen; for example, a number of patients had blood test results present outside episodes of hospital care due to attendances at their GP. Secondly, there was further refinement of the data mining software based on the outcome of the model redefinition. This consisted of the addition of validation checks, to ensure that state transitions found in the data did not occur between states in the model that did not have a defined path between them. For example, it was initially indicated by the data that patients were still receiving chemotherapy treatment after they had died, which created state transitions in the model that were clearly invalid. This was caused by the pre-emptive booking of many treatments into the system while the patient was still alive which were not subsequently deleted.

The final model schematic of the patient pathway for a cycle of chemotherapy is shown in Figure 4. The model includes a number of health states, each with a code. Across both regimen and diagnosis combinations, 40,047 health states were analysed (18,529 EC90, 21,518 OxMdG) which included 69,326 raw events (34,037 EC90, 35,289 OxMdG). S1 is the point at which the patient receives chemotherapy; states clustered near the bottom right corner represent admissions, whereas those clustered in the bottom left represent events when the patient is not admitted, and states clustered in the upper half of Figure 4 represent events that occur when the patient is due their next cycle of chemotherapy. When patients experienced more than one state within a cycle of chemotherapy, they were prioritised by the most severe event (Table 1), but counts were kept of the frequency of all health states. An uncomplicated case would be represented by a patient traversing through chemotherapy delivery (S1), followed by no contact (D1), then by patient review (R1) and back to chemotherapy delivery (S1). By contrast, an emergency admission to hospital and a blood test result within 24 hours indicating a neutrophil count of <1.5 x10^9/L would represent state S5 in the model. Attendance for chemotherapy that was not delivered on the expected date and a blood test result within 24 hours indicating a neutrophil count of <1.5 x10^9/L would represent state S2 in the model.

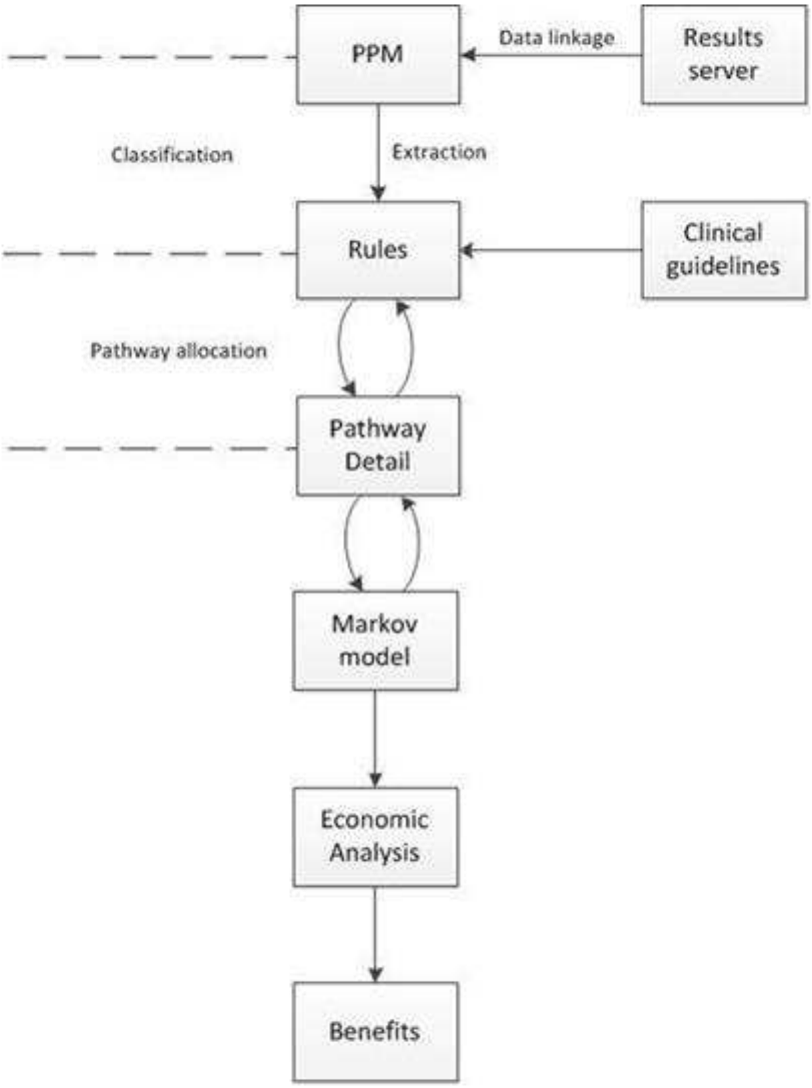

Figure 3: The process of creating and iteratively improving the state model.

Ethical Considerations

This work was sanctioned according to local LTHT Research and Development policy. Data extraction was carried out under strict information governance procedures, including anonymisation of patient-level data.

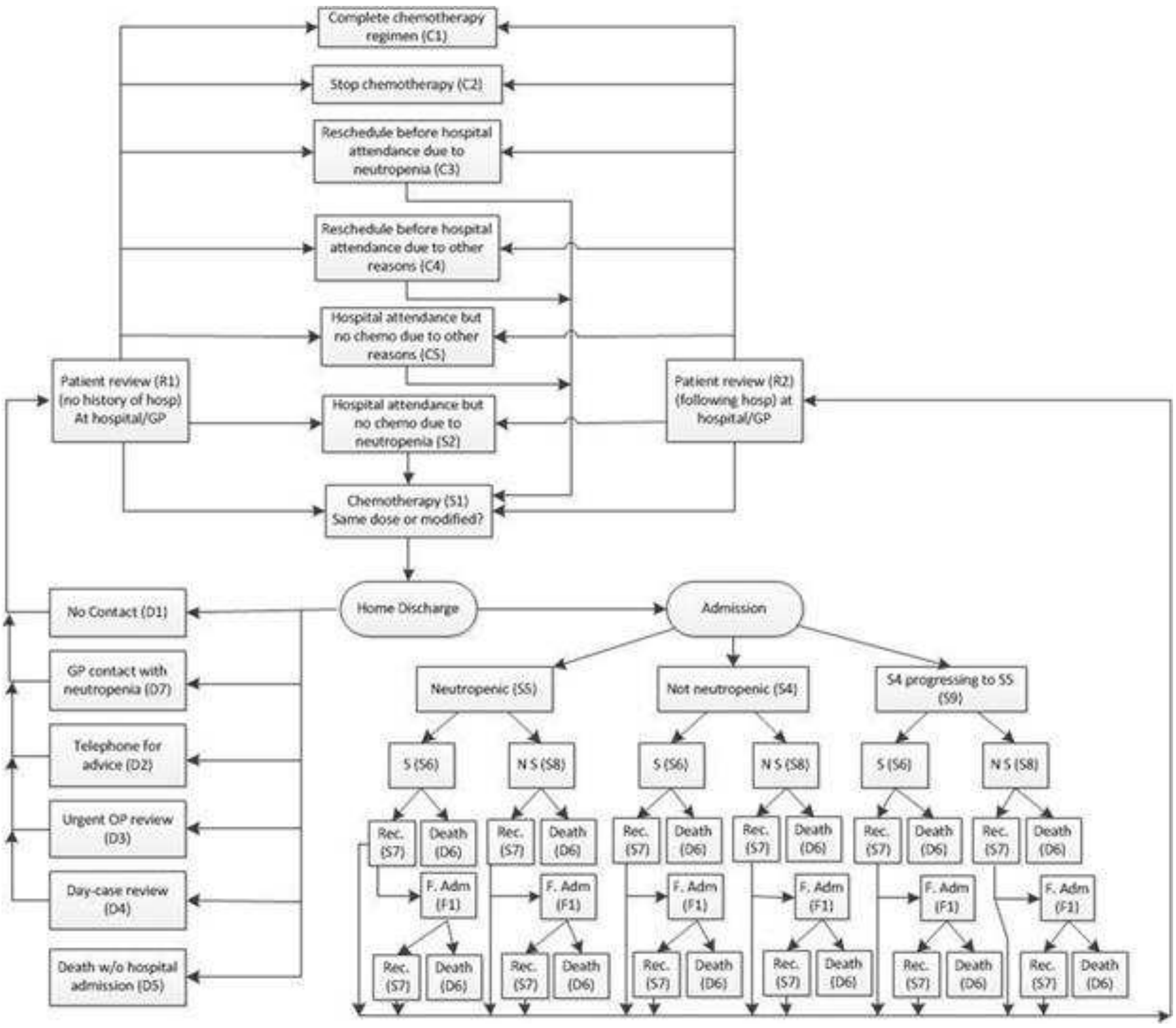
Complete chemotherapy regimen (C1)
Stop chemotherapy (C2)
Reschedule before hospital attendance due to neutropenia (C3)
Reschedule before hospital attendance due to other reasons (C4)
Hospital attendance but no chemo due to other reasons (C5)
Patient review (R1) (no history of hosp) At hospital/GP
Hospital attendance but no chemo due to neutropenia (S2)
Patient review (R2) (following hosp) at hospital/GP
Chemotherapy (S1) Same dose or modified?
No Contact (D1)
Home Discharge
Admission
GP contact with neutropenia (D7)
Neutropenic (S5)
Not neutropenic (S4)
S4 progressing to S5 (S9)
Telephone for advice (D2)
S (S6)
N S (S8)
Urgent OP review (D3)
Rec. (S7)
Death (D6)
F. Adm (F1)
Day-case review (D4)
Death w/o hospital admission (D5)

Figure 4: The state model at the end of iteration eight with the description of states.

S1, patient presents and the chemotherapy goes ahead as normal. D0, home discharge following chemotherapy. D1, patient makes no contact with the hospital. D2, telephone contact between hospital and patient. D3, patient has an urgent outpatient review. D4, patient has an urgent outpatient appointment. D5, death without hospital admission. D7, GP attendance with neutropenia. S2, patient attends hospital for chemotherapy but it is re-scheduled/delayed due to neutropenia. S4, patient is admitted acutely to hospital with no evidence of neutropenia. S5, patient is admitted acutely to hospital with evidence of neutropenia. S6, patient develops bacteraemia. S8, patient does not develop bacteraemia. S9, patient is admitted without neutropenia but develops it during admission. D6, patient dies in hospital. S7, patient fully recovers from acute event. F1, patient undergoes further hospital admissions. R1, review before the day of planned chemotherapy for patients without hospital admission (in that cycle). R2, review before the day of planned chemotherapy for patients with hospital admission (in that cycle). C1, complete all planned chemotherapy. C2, stop planned chemotherapy prematurely. C3, re-schedule before hospital attendance to undergo chemotherapy due to neutropenia. C4, re-schedule before hospital attendance to undergo chemotherapy due to reasons other than neutropenia. C5, patient attends hospital for chemotherapy but it is re-scheduled due to reasons other than neutropenia. Neutropenia is defined as a neutrophil count of < 1.5 x 10^9/L. Bacteraemia is defined as any positive blood culture.

Results

Clinical Pathways

During the first cycle of chemotherapy, 66.4% and 77.4% of patients receiving EC90 and OxMdG chemotherapy, respectively, traversed the planned pathway i.e. without encountering deviation caused either by treatment toxicity, other health complications or personal events unrelated to health such as holidays. Patients receiving the first cycle of EC90 chemotherapy experienced a higher percentage of emergency admission to hospital with neutropenia, urgent outpatient review, GP contact with positive test for neutropenia and telephone contact between hospital and patient (Table 1). Conversely, patients receiving first cycle of OxMdG chemotherapy experienced a higher percentage of death without hospital contact, emergency admissions without neutropenia and day case reviews.

With both regimens, only a small minority of patients received six cycles of chemotherapy on the planned pathway (Table 2). Of the 535 patients with early breast cancer receiving the adjuvant EC90, only 27 (5%) completed all six planned cycles of chemotherapy on the planned pathway (Table 2). In all, there were 474 pathway variants indicating that almost all of the remaining 508 patients followed a unique trajectory through the care system. Likewise, of the 420 patients with advanced colorectal cancer receiving palliative OxMdG, only 26 (6%) completed six cycles of chemotherapy on the planned pathway and there were 329 pathway variants. Additionally, the likelihood of an adverse event decreased with each successive cycle of EC90 chemotherapy, whereas that likelihood fluctuated with successive OxMdG chemotherapy cycles (Table 2). Over the 6 cycles of chemotherapy, 169 (31.6%) of patients receiving EC90 chemotherapy and (190) 45.2% of patients receiving OxMdG chemotherapy were admitted to hospital.

	EC90			OxMdG		
Event	**Count**	**Percentage**	**Cumulative**	**Count**	**Percentage**	**Cumulative**
Death without hospital contact (D5)	0	0.0%	0	14	3.3%	3.3%
Emergency admission to hospital with neutropenia (S5) *	21	3.9%	3.9%	8	1.9%	5.2%
Emergency admission without neutropenia, progressing to neutropenia whilst in hospital (S9) *	2	0.4%	4.3%	2	0.5%	5.7%
Emergency admission to hospital without neutropenia (S4) *	40	7.5%	11.8%	52	12.4%	18.1%
Day case review (D4)	3	0.6%	12.3%	9	2.1%	20.2%
Urgent outpatient review (D3)	19	3.6%	15.9%	0	0.0%	20.2%
Contacted GP and tested positive for neutropenia (D7)	24	4.5%	20.4%	6	1.4%	21.7%
Telephone contact from hospital to patient (D2)	71	13.3%	33.6%	4	1.0%	22.6%
No contact (D1) (planned pathway)	355	66.4%	100.0%	325	77.4%	100.0%
Total	**535**			**420**		

Table 1: Counts of the main pathways traversed for the first cycle of chemotherapy. Ordered from most severe event type to least severe (EC90 in breast cancer patients and OxMdG in colorectal cancer patients). *Emergency admissions include those admitted to hospital and those who attended the oncology acute assessment unit. Letter/number in brackets refers to states in Figure 4.

	EC90		OxMdG	
Cycle	No Contact (D1)	Admission (S5/S4/S9)	No Contact (D1)	Admission (S5/S4/S9)
1	66.6%	11.8%	77.4%	14.8%
2	80.0%	6.5%	80.0%	14.2%
3	72.9%	6.4%	83.0%	8.9%
4	74.2%	6.9%	81.0%	11.7%
5	82.7%	5.4%	80.1%	12.1%
6	86.5%	4.6%	90.0%	5.8%
All	**5.0%**	**31.6%**	**6.2%**	**45.2%**

Table 2: Adverse event rate by cycle of chemotherapy (EC90 for breast cancer and OxMdG for colorectal cancer). Letter/number in brackets refers to states in Figure 4. No contact represents the "planned pathway".

Discussion

We have demonstrated how data collected as part of routine clinical practice can be used to define real-life pathways of care. The key outcomes of this work are the description of the reproducible steps followed to successfully construct the pathways, and the clear demonstration that clinical pathways of real-life patients are far more complex than those defined by standard guidelines, expert opinion, consensus discussions, and clinical trials.

The percentage of patients admitted to hospital over all the six cycles was higher than compared to phase III trials (23, 24). This may reflect the real-life, heterogeneous population served by the Trust delivering the care. The differences between the pathways of patients receiving EC90 and OxMdG chemotherapy may reflect in part the dose intensity and myelotoxicity of regimens used in the adjuvant and palliative settings, patient characteristics and co-morbidities associated with metastatic cancer.

Key strengths of this work are that the data have been collected for routine clinical care from a large base population over a period of almost ten years and that the EHR is sufficiently mature and detailed (drawing from multiple disparate databases) to enable construction of accurate reports of real-life patient pathways. There are, however, limitations to using EHR data. The on-going debate in the UK around weekend admissions and increased mortality highlights how secondary uses of coded data can be valuable, but caution should be applied to avoid misinterpretation by apportioning un-evidenced cause to the findings. The detail and accuracy of pathways is limited by the accuracy and detail of data recorded. For example, we did not have electronic records of temperature measurements so our transformation rules used the taking of blood cultures as a surrogate indicator of fever. There was also inconsistency over time in the chemotherapy regimen nomenclature used due to prescribing and recording systems changing, meaning that the two regimens studied were only referenced in the system by identical names in a subset of the total records. The remaining records were excluded due to uncertainty about whether the treatments were identical in drugs used, dosage and method of delivery. In light of the recent publication of the Caldicott report (25), which this work pre-dates, we acknowledge there was no opt out process for patients from whom data was included in this analysis. This was a service evaluation process, with strict information governance procedures, where data was extracted, anonymised and analysed by aggregated pathways, not at an individual patient level.

Importantly, this pathway mapping provides additional data not apparent from national cancer datasets in the UK such as the National Cancer Waiting Times Monitoring Dataset

(NCWTMD), Cancer Outcomes and Services Dataset, Systemic Anticancer Therapy Dataset and cancer registries (26). Not all care elements are well represented in the national datasets such as suspected cancer diagnoses not referred by a general practitioner, care prior to a 2-week wait referral and care post referral to community palliative care teams. In addition to establishment of and checking adherence to guidelines, mapping real-life pathways gives a clear indication of current clinical practice, making it possible to identify unmet needs and process delays contributing to delayed targets, such as those mandated by the NCWTMD.

Specifically in the context of the research question that prompted this process mining initiative, the data has provided a more detailed picture of the issues around the risk of developing neutropenic sepsis, including how it is managed in routine practice. We selected the populations phenotypically using diagnosis and chemotherapy regimen combinations, but this can be narrowed further by selecting populations using parameters such as genotype and co-morbidity. We have demonstrated a method of organising clinical data into meaningful, quantified pathway models, which practically can be used to identify unmet need, model process improvements and inform health economic analyses. Focusing on neutropenic sepsis, the models of EC90 and OxMdG, can be used to determine where in the pathways home neutrophil monitoring will provide the greatest clinical and economic benefits for both patients and care providers.

As EHRs expand across care-provider boundaries, such as the Leeds Care Record, the ability to map such integrated pathways provides opportunity to measure health and healthcare in ways that have not been possible previously on large numbers of patients, including costs, clinical outcomes and societal impacts. In the UK NHS payments are based on Healthcare Resource Group (HRG) codes, which assume an “average” payment for finished consultant episodes. The size and frequency of the variance in pathways may therefore adversely affect financial performance and the sustainability of an organisation. Basing commissioning contracts on idealised planned pathways is, therefore, prone to serious under estimation of true costs since many of the variant pathways have very high associated costs, such as admissions to intensive care (27). The described pathway in the format of a Markov model, is suitable to inform parameter inputs into health economic analysis of the pathways, without the need to collect additional data. Also, pathway mapping enables improved quality assurance through electronic criteria adherence reporting, computation of quality indicators and visualising pathways thus facilitating identification and checking of anomalies.

We are applying process mining for real-life pathway modelling to all cancer diagnoses and chemotherapy combinations in our large Cancer Centre, in order to model the impact of innovative new pathways aiming to reduce the frequency and severity of neutropenic complications. Further detailed knowledge is required to appropriately inform clinical practice. We are addressing this through process mapping at the individual patient level, and service line costing to quantify the resource utilisation associated with the departures from the planned pathway. All of this is required to support the investment in research applied to addressing the unmet needs, and later to facilitate adoption of new pathways if trials show potential interventions to be clinically effective.

In conclusion, real life healthcare pathways are highly variable. This paper represents "proof-of-principle" that, even in the complex field of oncology, comprehensive data derived from EHRs can generate clinically important information. Understanding and visualising real-life patient-centred pathways should have an important role in identifying unmet needs and facilitating quantification of potential and actual impact of proposed innovations or changes in practice outside of clinical trials. There are implications for planning and supporting routine care, clinical and operational research, performance management and quality improvement.

Acknowledgements

Research Funding:

This research was supported through Innovate UK (Technology Strategy Board) Small Business Research Initiative (SBRI) Grant awarded to Philips Home Clinical Monitoring, who awarded an unrestricted educational grant to the University of Leeds. This forms part of the work funded in the SBRI-1 grant, project number 1203_SBR_SB2_DANR_SBRIApplication.doc 20504-149147. The work was hosted by Leeds Teaching Hospitals NHS Trust.

This article represents independent research supported by the National Institute for Health Research (NIHR) Diagnostic Evidence Co-operatives Leeds at the Leeds Teaching Hospitals NHS Trust and Oxford (GH and ED for Leeds, CP and JW for Oxford). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Professor Rick Jones:

We would like to acknowledge the contribution of Professor Rick Jones who died during the process of this work. He contributed significantly to the conception, planning and performing of the study, much of which would not have occurred without his forward thinking, pro-active approach.

References

1. Gray, J.M. Four box health care: development in a time of zero growth. *The Lancet.* 1983, **322**(8360), pp.1185-1186.
2. Christensen, C.M., Grossman, J.H. and Hwang, J. The innovator's prescription. *A disruptive solution for health care. New York: McGraw-Hill.* 2009, p.441.
3. Baker, G.R. *High performing healthcare systems: delivering quality by design.* Longwoods Publishing, 2008.
4. Kinsman, L., Rotter, T., James, E., Snow, P. and Willis, J. What is a clinical pathway? Development of a definition to inform the debate. *BMC medicine.* 2010, **8**(1), p.1.
5. Rotter, T., Kinsman, L., James, E., Machotta, A., Willis, J., Snow, P. and Kugler, J. The effects of clinical pathways on professional practice, patient outcomes, length of stay, and hospital costs: Cochrane systematic review and meta-analysis. *Evaluation & the health professions.* 2011, p.0163278711407313.
6. Rotter, T., Kinsman, L., James, E., Machotta, A. and Steyerberg, E.W. The quality of the evidence base for clinical pathway effectiveness: Room for improvement in the design of evaluation trials. *BMC medical research methodology.* 2012, **12**(1), p.1.
7. Yu, P., Artz, D. and Warner, J. Electronic health records (EHRs): supporting ASCO's vision of cancer care. *Am Soc Clin Oncol Educ Book.* 2014, pp.225-231.
8. Agarwal, P. and Owzar, K. Next generation distributed computing for cancer research. *Cancer Inform.* 2014, **13**(Suppl 7), pp.97-109.
9. W. Van Der Aalst, Process mining: discovery, conformance and enhancement of business processes, 1st ed. Springer-Verlag Berlin Heidelberg, 2011.10. E. Rojas and J. Munoz-Gama, "Process mining in healthcare: A literature review," J. Biomed. Inform., vol. 61, pp. 224–236, 2016.
11. Kurniati, A.P., Johnson, O., Hogg, D. and Hall, G., 2016. Process mining in oncology: A literature review. In Information Communication and Management (ICICM), 2016 6th International Conference on (pp. 291-297). IEEE. DOI: 10.1109/INFOCOMAN.2016.7784260
12. Bettencourt-Silva, J.H., Clark, J., Cooper, C.S., Mills, R., Rayward-Smith, V.J. and de la Iglesia, B. Building Data-Driven Pathways From Routinely Collected Hospital Data: A Case Study on Prostate Cancer. *JMIR Med Inform.* 2015, **3**(3), p.e26.
13. Caron, F., Vanthienen, J., Vanhaecht, K., Van Limbergen, E., De Weerdt, J. and Baesens, B. Monitoring care processes in the gynecologic oncology department. *Comput Biol Med.* 2014, **44**, pp.88-96.
14. National Confidential Enquiry into Patient Outcome & Death (NCEPOD). *For better, for worse? A review of the care of patients who died within 30 days of receiving systemic anti-cancer therapy.*, 2008.
15. National Chemotherapy Advisory Group (NCAG). *Chemotherapy Services in England: Ensuring quality and safety.* 2009.
16. Newsham, A.C., Johnston, C., Hall, G., Leahy, M.G., Smith, A.B., Vikram, A., Donnelly, A.M., Velikova, G., Selby, P.J. and Fisher, S.E. Development of an advanced database for clinical trials integrated with an electronic patient record system. *Comput Biol Med.* 2011, **41**(8), pp.575-586.
17. Johnson, O. and Abiodun, S.E. Understanding what success in health information systems looks like: the Patient Pathway Management (PPM) system at Leeds. In: *UK Acad Inf Syst Conf Proc*, 2011.
18. National Cancer Intelligence Network. *Cancer Outcomes and Services dataset (COSD).* [Online]. 2015. [Accessed 14th July 2016]. Available from: http://www.ncin.org.uk/collecting_and_using_data/data_collection/cosd
19. *Chemocare.* Belfast: Clinisys, 2013.
20. *MOSAIQ Radiation Oncology software.* Stockholm, 2016.
21. Raghupathi, W. and Raghupathi, V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems.* 2014, **2**(1), p.1.

22. Ambler, S.W. and Lines, M. Disciplined agile delivery. IBM Press, 2012.

23. Henderson, I.C., Berry, D.A., Demetri, G.D., Cirrincione, C.T., Goldstein, L.J., Martino, S., Ingle, J.N., Cooper, M.R., Hayes, D.F. and Tkaczuk, K.H. Improved outcomes from adding sequential paclitaxel but not from escalating doxorubicin dose in an adjuvant chemotherapy regimen for patients with node-positive primary breast cancer. *Journal of Clinical Oncology.* 2003, **21**(6), pp.976-983.

24. Adams, R.A., Meade, A.M., Seymour, M.T., Wilson, R.H., Madi, A., Fisher, D., Kenny, S.L., Kay, E., Hodgkinson, E. and Pope, M. Intermittent versus continuous oxaliplatin and fluoropyrimidine combination chemotherapy for first-line treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *The lancet oncology.* 2011, **12**(7), pp.642-653.

25. Caldicott, F. *Review of data security, consent and opt-outs.*, 2016.

26. Dale, D.C., McCarter, G.C., Crawford, J. and Lyman, G.H. Myelotoxicity and dose intensity of chemotherapy: reporting practices from randomized clinical trials. *J Natl Compr Canc Netw.* 2003, **1**(3), pp.440-454.

27. Dinan, M.A., Hirsch, B.R. and Lyman, G.H. Management of chemotherapy-induced neutropenia: measuring quality, cost, and value. *Journal of the National Comprehensive Cancer Network.* 2015, **13**(1), pp.e1-e7.

Appendix A

Explanation of Markov Models

Markov models can be used to aid health care decision making. They are suited to decisions where the timing of events is important and when events may happen more than once, appropriate when the strategies being evaluated are of sequential or repetitive nature. They can be used to model the cost-effectiveness of interventions by incorporating long terms costs and health outcomes. Here we briefly describe the principles behind Markov models for the uninitiated reader, as there are existing comprehensive introductions to Markov modelling (1, 2).

A cluster of clinical events are simplified by defining them as a clinically important health states or Markov states. For example, in our model, a patient in the health state of febrile neutropenia, had an entry in the hospital admissions table, was neutropenic, had blood cultures taken which did not grow a pathogenic organism. Markov models assume that the health states are mutually exclusive as a patient cannot be in more than one state at any one time. The transition of a patient from one health state to another is assigned a probability, known as the transition probability. In our work, this probability was defined by the patient data from the electronic health records. Markov models represent repetitive processes over time whereby the patient passes through the same health state on more than one occasion, represented in our model by the delivery of chemotherapy. Patients can therefore, only exit the repetitive model via defined exit states, which in our model exist as one of only three states; complete chemotherapy, stop chemotherapy prematurely, death. There is also an assumption that transition to future states depend only on the current state, and not on any events that occurred before the current state. This is known as the Markov assumption. For example, in our model, a patient defined as being treated for febrile neutropenia (A3,S5,M1,S8), is dependent only on the current health state (blood cultures being taken, which is A3, S5, M1), which is in no way influenced by the occurrence of febrile neutropenia in previous chemotherapy cycles.

Markov models enable estimates of cost and outcomes associated with disease and intervention. This is done by assigning estimates of resource usage and health outcomes to health states and transitions between states, then running the model over a large number of the repeated Markov cycles. In this way, the cost and effectiveness of a proposed intervention can be modelled using assumptions of the effect on transition probabilities. Drawing the above together, the steps commonly followed to construct a Markov model are; deciding on the health/Markov states, defining acceptable transitions, identifying transition probabilities, identifying health outcomes, then running the model over repeated Markov cycles.

Appendix A references

1. Sonnenberg, F.A. and Beck, J.R. Markov models in medical decision making: a practical guide. *Med Decis Making.* 1993, **13**(4), pp.322-338.
2. Briggs, A. and Sculpher, M. An introduction to Markov modelling for economic evaluation. *Pharmacoeconomics.* 1998, **13**(4), pp.397-409.

Appendix B

Detailed description of the development of the datamining software and pathway models.

In early iterations of the software for this project, the full analysis took around 20 hours to run on a machine with an Intel Core I7 processor with 16GB of memory, but it was realised that patients could be processed independently of each other and the time to run was reduced to around 3 hours by using multithreading.

The warehousing process is done in three stages:

Stage 1

A cryptographically secure random integer for each patient is generated to be used as an anonymous identifier. This is stored in a separate database on a secure server within the Trust, and the patient is only referenced by their anonymous identifier in the warehouse database.

The relevant data for this study is stored in many different tables in the PPM database, for example, admission and discharge event data is stored in the admissions table, and chemotherapy administration event data is stored in the chemotherapy table. Events from all relevant PPM Tables are merged into one raw events table containing the date and time that each event happened, a code to indicate the type of event and code(s) to indicate additional attributes of the event that will be used for the state categorisation rules in stage 2. Merging the events into one table makes it easy for the software to iterate over each event in sequence during stage 2.

A copy of the regimens table from PPM is created, containing data from only the patients selected in the study, and with columns removed that are not relevant for this study. Computed columns are added that do not exist in PPM; the regimen end date is calculated by taking the start date of the last cycle in the regimen and adding the cycle length. The diagnosis that the regimen is indicated for is inferred by taking the patient's diagnosis that is closest to the regimen start date. This table will assist in creating reports with results categorised by regimen and or diagnosis

A copy of the Cycles table from PPM is created, with only the necessary data, and with additional computed column end date calculated by taking the start date of the next cycle if there is one, or the start date of the current cycle plus the cycle length. This table will assist with categorising results by cycle number.

Stage 2

The events in the raw events table form a single concurrent pathway for each patient. This stage of transformation attempts to fit this pathway to the model (Figure 4 in manuscript) defined by the research team. This stage uses the following process:

- For each patient the events from the raw events table are loaded into memory. All events before their first chemotherapy administration event are skipped as this study is not concerned with events before this.
- Each event is iterated over and the rules in Table B1 are used to decide the appropriate model state.
- Whilst these states are being computed, transition validity is also checked. A transition from one model state to another is valid if there is a line between them on the model diagram, for example S1 to D4 would be valid, but S4 to D4 would be invalid. Although some transitions are valid that are not represented in Figure 4, as further simplification of certain event sequences takes place later on (described in stage 3). If an invalid transition is detected then this is added to an error log. All invalid transitions must be investigated by examining data records in PPM manually. The most common causes of these will be invalid input data, incomplete input data, or an incomplete model.

Stage 3

The sequence of states output by stage 2 can contain sequences of events that are too complicated for the model in Figure 4 to easily represent them. For example a patient can have many admissions, telephone for advice many times, or have any combination of many events. This would make the resultant model extremely complex to visualise if the pathway were to accommodate all outcomes. So this stage attempts to simplify the output.

This is achieved by splitting the long complex sequence of events into small sets. A small event set is a sequence of one or more events that is created by splitting the true sequence of events at each point where the pathway deviates from the model in Figure 4.

This process is shown visually in Figure B1. The red pathway highlights the true pathway taken by the patient through the following states: S1 > D2 > D4 > S4 > S8 > S7. This pathway is split at the blue X's into three small event sets:

1. S1 > D2
2. D4
3. S4 > S8 > S7

The sets are then weighted by severity. Each state in the model has a severity weighting (shown in Table B2 and chosen based on impact to quality-adjusted life year (QALY) for the patient and cost for the hospital). To calculate the severity weighting for a set, the weighting of all its constituent states are summed, giving the following weightings for the above example:

1. 0 + 2 = 2
2. 4
3. 5 + 1 + 0 = 6

Raw event	Rules	Model state
Chemotherapy administered	Chemotherapy was delivered on time or up to 2 days late.	S1
	Chemotherapy was >= 3 days delayed and blood test result indicating neutropenia exists in the period of -7 to +3 days from this event.	C3
	Chemotherapy was >=3 days delayed and no result indicating neutropenia.	C4
Emergency admission	No result indicating neutropenia during admission (between admission event and discharge event).	S4
	Neutropenic result up to 24 hours after admission.	S5
	No neutropenic result during the 24 hours following the admission but neutropenic result before discharge.	S9
	Bacteraemic blood culture result during admission	S6
	No Bacteraemic blood culture result during admission	S8
	Death during admission	D6
Discharge after emergency admission	None	S7
Blood count result	If result indicates neutropenia and date was outside of admission	D7
Telephone contact	None	D2
Urgent outpatient review	None	D3
Death	If date was outside of admission	D5
	If during admission	D6
Non-emergency admission	If chemotherapy was scheduled to be delivered during this admission*, but wasn't, and neutropenic blood count result exists <= 2 days from admission	S2
	If chemotherapy was scheduled to be delivered during this admission*, but wasn't, and no evidence of neutropenia	C5
	No chemotherapy scheduled* during this admission and length of stay <= 1 day	D4

* Calculated by taking previous cycle date plus cycle length

Table B1: The rules used in stage 2 to transform raw events into model states.

In this study the researchers were primarily interested in (for each patient) (i) the most severe small event set, and (ii) the frequency of occurrence of each model state, regardless of inclusion in small event sets. To serve these two interests, each state in the model was given two numbers or counts to be output in the aggregate report data. The first count was the severity count, which indicated how many patients experienced the state as part of their most severe small event set. The second count was the frequency count, which indicated how many times each state had occurred in total. The severity count then has the property that if a single pathway splits into many then the severity count before the split will equal the sum of the severity count on each of its branches, which is useful for checking data validity and for making inferences based on distinct sets of patients. The frequency count is useful for estimating total cost by simply multiplying the frequency count of a state by the cost per occurrence of a state. Reports were then generated showing the severity count and frequency count for every state in the model, and could be grouped by any combination of regimen, indication and cycle number.

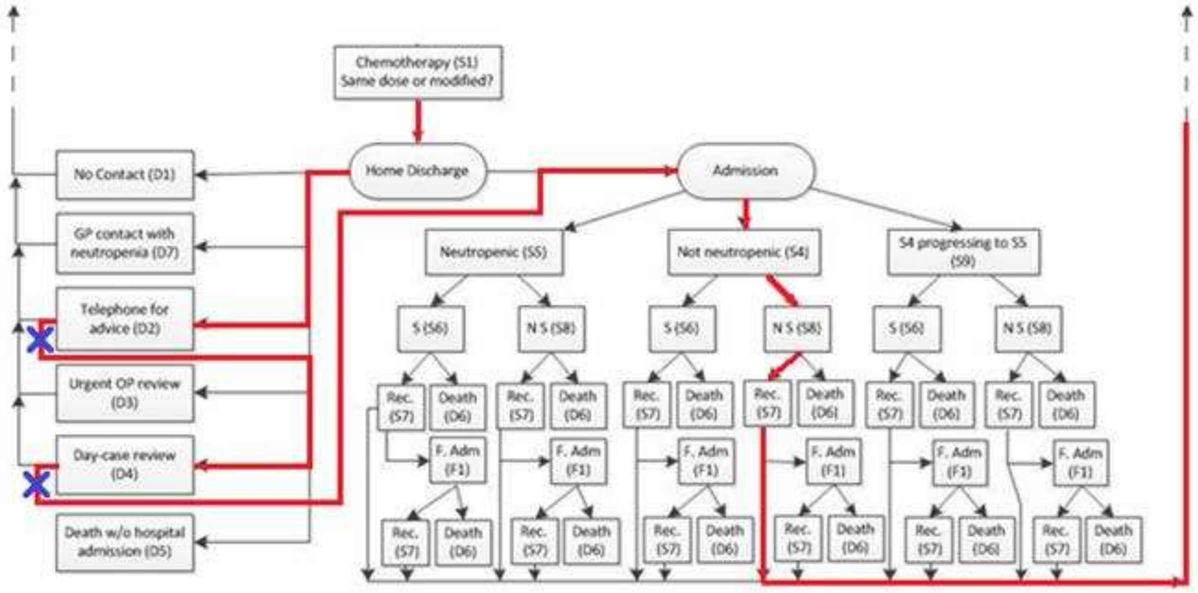

Figure B1: An excerpt of the model in Figure 4 used to demonstrate the process of splitting a long series of events into small event sets.

State	Weighting	Description
S1	0	Chemotherapy received
D0	0	Home discharge
D1	0	No contact
D7	1	GP contact with neutropenia
D2	2	Telephone for advice
D3	3	Urgent OP review
D4	4	Day case review
S4	5	Non-neutropenic emergency admission
S9	6	Non-neutropenic emergency admission progressing to neutropenia during stay
S5	7	Neutropenic emergency admission
S6	0	No bacteraemia during admission
S8	1	Bacteraemia during admission
S7	0	Discharged from hospital
D6	4	Death during hospital stay
D5	9	Death outside of hospital stay
C4	1	Reschedule before hospital attendance due to other reasons
C3	2	Reschedule before hospital attendance due to neutropenia
C5	3	Hospital attendance but no chemotherapy due to other reasons
S2	4	Hospital attendance but no chemo due to neutropenia
C2	5	Stop chemotherapy before completing the regimen
C1	5	Complete all planned chemotherapy

Table B2: The severity weighting of all major states in the model.