Ann Oper Res (2018) 270:287–312
https://doi.org/10.1007/s10479-016-2393-z

Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review

M. M. Malik[1] · **S. Abdallah**[2] · **M. Ala'raj**[2]

Published online: 24 December 2016

Abstract With the widespread use of healthcare information systems commonly known as electronic health records, there is significant scope for improving the way healthcare is delivered by resorting to the power of big data. This has made data mining and predictive analytics an important tool for healthcare decision making. The literature has reported attempts for knowledge discovery from the big data to improve the delivery of healthcare services, however, there appears no attempt for assessing and synthesizing the available information on how the big data phenomenon has contributed to better outcomes for the delivery of healthcare services. This paper aims to achieve this by systematically reviewing the existing body of knowledge to categorize and evaluate the reported studies on healthcare operations and data mining frameworks. The outcome of this study is useful as a reference for the practitioners and as a research platform for the academia.

Keywords Healthcare operations management · Predictive analytics · Data mining · Systematic literature review · Big data

1 Introduction

The continuous and voluminous growth of data commonly known as big data, generated from various business software applications and devices has put the onus on machine aided

✉ M. M. Malik
mohsin.malik@unimelb.edu.au

S. Abdallah
salam.abdallah@adu.ac.ae

M. Ala'raj
maher.alaraj@adu.ac.ae

[1] Department of Management and Marketing, The University of Melbourne, Parkville, Australia

[2] College of Business Administration (COBA), Abu Dhabi University, P.O. Box 1790, Abu Dhabi, UAE

analysis for comprehending the business intricacies. Now, the key issue is that how big data can be most effectively used to generate new insights to improve processes and to aid managerial decision-making (Dubey et al. 2016; Agarwal and Dhar 2014) This process of knowledge discovery from the big data is commonly known as data mining (Fayyad et al. 1996). Data mining and predictive analytics aims to reveal patterns and rules by applying advanced data analysis techniques on a large set of data for descriptive and predictive purposes (Delen and Demirkan 2013). Glowacka et al. (2009) defined data mining as the 'nontrivial extraction of implicit, previously unknown and potentially useful information from data' providing a strong basis for informed decision making. Recently there has been a paradigm shift towards exploring possibilities of utilizing data mining tools and techniques for designing and managing operations and supply chains (Wamba et al. 2015; Waller and Fawcett 2013). Amazon's success has mainly been credited to the early adoption of innovative business ideas and state of the art technologies. They also seem to have realized the potential of big data by filing a predictive order fulfillment and logistics patent that practically has no delivery lead times as data mining helps Amazon determine what potential customers may end up buying and shipping products in their general direction even before they have finalized their purchases (Spiegel et al. 2013). Similarly, Wang et al. (2016) noted the usefulness of big data for determining market trends and customer buying patterns which enables a closer match of supply with demand and therefore, lowering supply chain costs. Studies such as Dubey et al. (2016) have put the potential benefits of using big data in analyzing operations management and supply chain activities to 15–20% increase in return on investment (ROI), productivity and competiveness. This underscores the importance of data mining and predictive analytics for more informed decision making in the realm of operations and supply chain management (Papadopoulos et al. 2016). There seems a near consensus that data mining and predictive analytics have provided an innovative way of improving the supply chain processes and the businesses need to proactively embrace big data to stay competitive (Waller and Fawcett 2013; Hazen et al. 2014).

Like other industries, healthcare sector has also now access to large volumes of data from the medical information systems providing opportunities for more informed clinical and administrative decision making. In addition, Yoo et al. (2012) saw the large healthcare data particularly useful for generating scientific hypotheses for medical research. However, not surprisingly, healthcare data mining seems to have been most widely used for diagnosis, prognosis or treatment planning; selective papers from a plethora of literature include disease management for oncology, liver pathology, neuropsychology or gynecology (Koh and Tan 2011), for telemedicine (Gheorghe and Petre 2014), to predict heart attacks (Srinivas et al. 2010), identify and classify at-risk people (Anderson and Chang 2015) and determining patient acuity levels (Kontio et al. 2014). Other reported data mining and predictive analytics applications in healthcare include customer relationship management (Koh and Tan 2011), detection of fraud (Menon et al. 2014; Yoo et al. 2012) and evaluation of treatment effectiveness (Koh and Tan 2011). With rising healthcare expenditure and shrinking budget allocations, there also seems to be considerable interest in using big data for cutting healthcare costs. Bates et al. (2014) suggested analyzing high-cost patients, readmissions, triage, compensations, adverse events, and treatment optimization plans with big data for potential improvements. Amarasingham et al. (2014) maintained that apart from healthcare cost optimization, the use of predictive modeling for real-time clinical decision making could also be used to improve health outcomes and enhancing patients' experiences. The same ideas seem to have been resonated by Haux et al. (2002), Siau (2003), Harper (2005), and Bonacina et al. (2005); all studies foreseeing data driven

organizational attempts to reduce costs, be more competitive, and provide a better and more personalized patient care. To summarize, there is a significant scope for improving the way healthcare is delivered by resorting to the power of big data. The literature has reported big data applications for the delivery of healthcare services, however, there appears no attempt for evaluating and synthesizing the information on healthcare operations and supply chain management data mining application which may be used as reference for the practitioners and as a platform for future research projects. A systematic review of literature, which has been employed in this study, aims to achieve this by rigorously and systematically searching, describing, classifying and evaluating the available knowledge in an auditable way. The systematic review is concluded by providing a reflective interpretation of the existing gaps in the literature followed by the development of new research propositions. In the next section, a framework for healthcare operations management is adapted which will be used as an analytic category for classifying the reported data mining applications for healthcare service delivery. Section 3 describes the detailed methodology employed for this paper followed by discussions in Sect. 4 and conclusions and implications in Sect. 5.

2 Healthcare operations and supply chain management

Health care policy makers are under intense pressure in the wake of shrinking healthcare budgets to effectively and efficiently meet an ever growing demand. Healthcare expenditure is big; the 2011 average total health expenditure for the Organization for Economic Co-operation and Development (OECD) countries was recorded at 9.38% of their GDP, with the United States' spending being the highest at 17.7%. Given the money involved, the range of healthcare services and its impact on human lives, it is not surprising that operational excellence is now considered central to any hospital setting (Malik et al. 2015). There is a growing realization that the 'process view' of hospitals, where care activities are viewed as a collection of processes that transforms input resources into outputs, is a pre-requisite for healthcare improvement. Vissers and Beech (2005) formally defined Healthcare Operations Management (HOM) as the analysis, design, planning and control of all of the steps necessary to provide a service for a patient. The care chain or healthcare supply chain is a related concept that extends the traditional input-output view of an operation to include all services for patients provided by various medical specialties and functions, within and across departments and also across organizations to include the suppliers. Porter and Teisberg (2006) sees the rebranding of all healthcare activities as a care value delivery chain a necessary transformation for better outcomes and more customer value. To improve the care chain value delivery performance, this process thinking warrants use of operations research concepts and tools for capacity planning, layout designs, facilities location, workforce scheduling, planning staffing levels and the appointment scheduling. Langabeer and Helton (2015) also noted the extensive use of quantitative methods in HOM such as the use of analytical and optimization tools as well as the process and quality techniques to drive the improvements. This healthcare operations and supply or care chain (HOSCM) perspective is particularly important given the criticality of quality and safety in the healthcare industry. Noting the growing interest, Dobrzykowski et al. (2014) identified information technology/new technology in healthcare, operations strategy and objectives, design of the care delivery system, quality issues and capacity planning, scheduling, and control as the five most researched topics within HOSCM. Taking a cue from these most researched themes, we follow Seuring and Gold (2012)'s deductive approach of

Table 1 Key functions and issues in health care operations and supply chain management (HOSCM)

Function	Sub dimensions
Capacity design and planning	Bottleneck and throughput analyses
	Patients/workforce/resource allocation/costing and scheduling
	Use of technology to improve labor productivity
Workflow process	End-to-end mapping of care delivery processes (clinical pathways)
	Reduction of cycle time, steps, and choke points for key processes
	Compliance to clinical pathways
Physical layout	Facility design with the consideration of traffic flow, and operational efficiency
	Floor layouts design to eliminate redundancy (e.g., safety stock)
Physical network optimizations	Positioning appropriate par locations, pharmacy, satellites, warehouses, and suppliers to minimize resources and costs
	Designing optimal locations for clinics or resources to ensure lowest total costs
People, jobs and organizations	Staffing levels
	Job design and analyses
	Managing intra/Inter departmental boundaries
	Measuring performance indicators
Supply chain and management	Single and multi-sourcing
	Vendors and their facilities utilization
	Logistics management
Productivity and process improvement	Measuring patient's experiences
	Just in time operations
	Quality of care
	Productivity management
	Asset utilization
	Inventory turns

analytic category building along with the adaptation of Langabeer and Helton (2015) key functions and issues in healthcare operations management to map the extant literature reported on applications of data mining and predictive analytics in HOSCM (Table 1). Further details of Seuring and Gold (2012)'s deductive approach are mentioned in Sect. 3.

3 Methodology

This paper uses content analyses for a systematic literature review to explore the use of data mining and predictive analytics in healthcare operations and supply chain management. Shapiro and Markoff (1997) defined content analysis as any methodological measurement applied to text for social science purposes. Seuring and Gold (2012) noted the added advantage of content analyses that it can be applied to both in a quantitative and a qualitative way. The authors gave a four step guidelines to conduct content analyses which has been used as the preferred methodology for this paper. Details are as follows:

3.1 Material collection

3.1.1 Defining the research objective, unit of analyses and identifying the relevant keywords

From the above discussion and to bridge the knowledge gap, the objective of this systematic review is to *find out how data mining and predictive analytics have contributed to the delivery of healthcare with an emphasis on healthcare operations management*. The unit of analysis was identified as a single paper. For a systematic literature review, selection of appropriate key words and search criteria ensures that the content selected and assessed would be within the boundaries of the defined research objective. We identified search keywords of "predictive", "analytic(s)", "data mining", "big data", "operations", "process", "supply chain", "process mining", "machine learning" and "optimization" to support our research question. The filter input was "healthcare" or "health" or "health care".

3.1.2 Access to the literature and the search criteria

To ensure that we select quality journals, two major databases (Scopus and Web of Science) were accessed in December 2015. The study focused on reviewing journal articles only, which were further filtered based on research areas that varied in both databases. Therefore, the search mechanism for the two databases is separately mentioned below:

Literature search: Scopus
The following search criteria was used for the Scopus databases: TITLE-ABS-KEY (("big data" OR "data mining" OR "predictive analytics" OR "machine learning" OR "Process Mining")) AND TITLE-ABS-KEY (("supply Chain" OR "Optimization" OR "Process" OR "Operations")) AND TITLE-ABS-KEY (("Health" OR "Healthcare" OR "Health care")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA, "MEDI") OR LIMIT-TO (SUBJAREA, "ENGI") OR LIMIT-TO (SUBJAREA, "HEAL") OR LIMIT-TO (SUBJAREA, "DECI") OR LIMIT-TO (SUBJAREA, "BUSI") OR LIMIT-TO (SUBJAREA, "PHAR") OR LIMIT-TO (SUBJAREA, "NURS") OR LIMIT-TO (SUBJAREA, "IMMU") OR LIMIT-TO (SUBJAREA, "DENT")) AND (LIMIT-TO (SRCTYPE, "j")) AND (LIMIT-TO (EXACTKEYWORD, "Health care") OR LIMIT-TO (EXACTKEYWORD, "Health")). The initial search extracted 1830 papers from various sources but since, our unit of analyses was a journal article, therefore, we kept 901 journal articles for further screening. Following subject areas were carefully selected to match the research question: "computer science", "medicine", "engineering", "decision sciences", "business management and accounting", "dentistry", "Pharmacology", "immunology", and "health professions". This reduced the available documents to 804 journal articles. A further filter was applied to limit the search to health care industry (filters selected: "healthcare" and "health") which limited to the papers of interest to 141 articles from Scopus (Fig. 1).

Literature search: Web of Science
The following search criteria was used for the Web of Science: "big data" OR "data mining" OR "predictive analytics" OR "machine learning" OR "Process Mining") AND TOPIC: ("supply Chain" OR "optimization" OR "process" OR "operations") AND TOPIC: ("Health" OR "Healthcare" OR "Health care"), refined by: LANGUAGES: (ENGLISH) AND DOCUMENT TYPES: (ARTICLE) and research areas: (computer science or research experimental medicine or engineering or medical informatics or health care sciences services or surgery

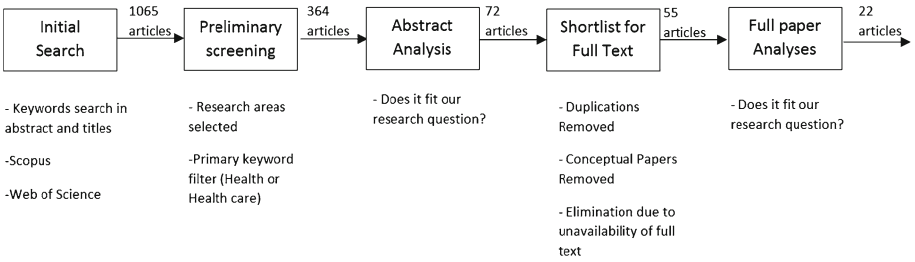

Fig. 1 The screening methodology

or operations research management science or public environmental occupational health or nursing or mathematical computational biology or business economics or information science library science or psychology or pharmacology pharmacy or science technology other topics or immunology). A total of 530 documents were retrieved from this Boolean searching. Out of this 261 were journal articles. Research areas provided by the database were used to further refine the articles and a total of 223 journal articles were selected for further reviewing (Fig. 1).

Abstract analyses and further screening
The 141 articles from Scopus and 223 papers from Web of Science gave us a total of 364 papers which were selected for abstract analyses to check their relevance with our research question. Supporting our research question of how data mining and predictive analytics have contributed to health care operations and supply chain management, we were interested in actual implementation of data mining, therefore, we also rejected conceptual papers. It turned out that only 65 papers were deemed matching the research agenda during the abstract analyses (Fig. 1). The screening criterion was that only those papers were ruled out if the researchers were convinced and in agreement that the rejected paper could not be fit in Table 1 categories. At this stage, we retrieved the complete articles which reduced the shortlisted papers number to 55 because of 7 duplications and full text was not available for 3 papers. Full paper analyses were the final screening step which helped to establish the body of knowledge to include in this study (22 journal articles) by removing irrelevant papers and any paper that gave a bird eye account i.e that it did not give sufficient details regarding the data mining implementation such as the modelling approaches, performance evaluation and the deployment (Fig. 1).

3.2 Descriptive analysis

The time distribution of 22 articles selected for content analyses show that data mining in the healthcare operations management is an emerging area of research with nearly 60% in the last 5 years and almost all publications coming in the last 10 years (Fig. 2). The distribution of the articles across journals is widely spread with only 4 journals publishing more than one paper with Journal of Biomedical Informatics being the highest with 4 publications, followed by 3 in Expert Systems with Application and two papers from International Journal of Medical Informatics and Journal of Operations Research Society. There are 11 journals that have only published once relating to our research questions indicating that the subject has attracted the interest of several fields of knowledge (Fig. 3).

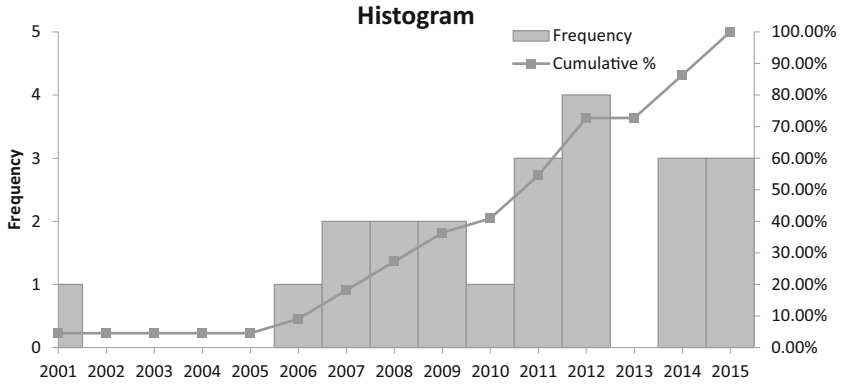

Fig. 2 Time distribution of the 22 papers

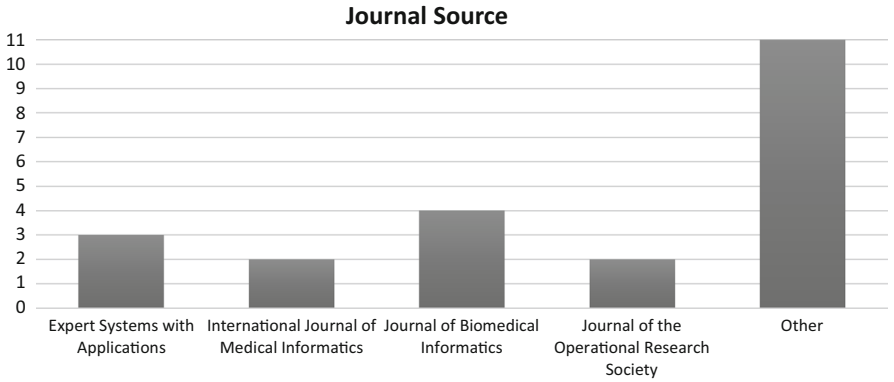

Fig. 3 Journal source distribution

3.3 Analytic category selection

Using Seuring and Gold (2012)'s two step approach for analytic category selection, we use following two established frameworks as overarching categories to deductively fit in the papers reviewed and then, for further sub-classification, we use iterative cycle for inductive category refinement.

- Cross-Industry Process for Data Mining (CRISP-DM) framework
- Healthcare Operations and Supply Chain Management Framework (Table 1).

CRISP-DM is widely considered to be comprising the best practices for data mining projects and was proposed to serve as a nonproprietary standard methodology for data mining (Koh and Tan 2011; Wirth and Hipp 2000). Figure 4 is an illustration of the CRISP-DM six steps namely; Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. This methodology has been adapted for this study to be used as an analytic framework to deductively categorize the reviewed body of knowledge using 5 analytic categories: (1) Business Understanding, (2) Data Collection, (3) Modeling, (4) Evaluation, and (5) Deployment.

Business Understanding phase focuses on understanding the project aims and requirements from a business perspective. Since the research domain for this paper is healthcare operations

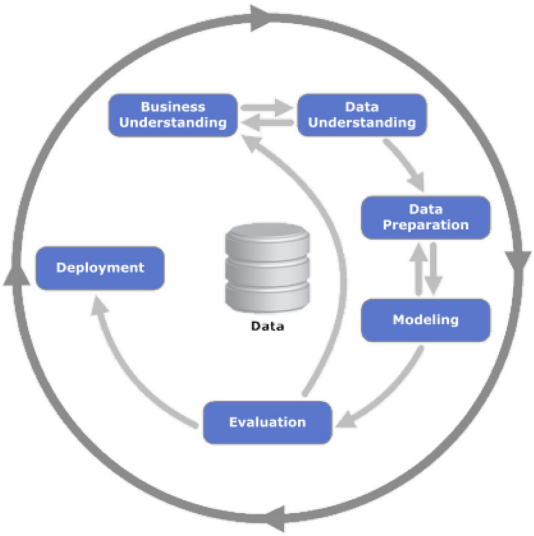

Fig. 4 Cross-industry process for data mining (CRISP–DM) (Wirth and Hipp 2000)

and supply chain management (HOSCM), we use Table 1 as the analytic category for Business Understanding. The data understanding phase starts with an initial data collection followed by attempts to get familiarized with data. Data quality problems are identified and efforts are made to discover data insights. The data preparation phase covers all activities to construct the final dataset from the initial raw data. The tasks comprising this stage may include table, record, and attribute selection, data cleansing, construction of new attributes, and transformation of data for modeling tools. Since in this study, we are not actually working with big data but only reviewing the data understanding and preparation practices, we will combine CRISP-DM data understanding and data preparation under one analytics category and rename it as data collection phase. Modeling phase describes various modeling techniques that may be selected and applied and also, it details the optimal parameters setting. Typically, there are several techniques for the same data mining problem. Evaluation phase measures the accuracy and validity of the modeling approach. Deployment phase focuses on presenting the knowledge gained that can facilitate its use by the end user. The deployment phase may comprise a spectrum with some projects may actually integrate modeling with their information systems for repeatable data mining exercise whereas some studies would only require report generation to aid managerial decision making.

3.4 Category mapping and material evaluation

In this section, we first allocate the reviewed body of knowledge to the five basic categories identified in the preceding section and then further classify it under sub dimensions using an inductive method combining the strengths of established frameworks and the exploratory research method. Details are mentioned in the following five subsections corresponding to the selected basic analytic framework:

3.4.1 Business understanding: healthcare operations and supply chain management

From Table 2, it appears that only three broad functional areas of healthcare operations and supply chain management (HOSCM) have seen the data mining and predictive applications

Table 2 Classification of data mining applications for healthcare functional areas

References	HOSCM—functions from Table 1	HOSCM sub-dimension(s)
Caron et al. (2014)	Workflow Analyses	Clinical/Care Pathways
Ceglowski et al. (2007)	Capacity Design and Planning	Queuing Analyses for Workforce Planning
Chi et al. (2008)	Productivity and Process Improvement	Quality of Care (Effective)
Cornalba et al. (2008)	Productivity and Process Improvement	Quality of Care (Effectiveness)
Demir (2014)	Productivity and Process Improvement	Quality of Care (Effective)
Garg et al. (2009)	Workflow Analyses	Clinical/Care Pathways
Glowacka et al. (2009)	Capacity Design and Planning	Outpatient Appointment Scheduling
Kudyba and Gregorio (2010)	Work Flow Process	Clinical/Care Pathways
Kuo (2011)	Workflow Analyses	Clinical/Care Pathways
Lavrač et al. (2007)	Capacity Design and Planning	Strategic Health Resources Planning
Lee et al. (2011)	Productivity and Process Improvement	Quality of Care (Patient Safety)
Lin et al. (2001)	Workflow Analyses	Clinical/Care Pathways
Malin et al. (2011)	People, jobs and Organizations	Intra/Inter Departmental Relationships
Ng et al. (2006)	Capacity Design and Planning	Hospitals resources planning
Rebuge and Ferreira (2012)	Workflow Analyses	Clinical/Care Pathways
Rovani et al. (2015)	Workflow Analyses	Clinical/Care Pathways
Rubrichi and Quaglini (2012)	Productivity and Process Improvement	Quality of Care (Effectiveness)
Samorani and Laganga (2015)	Capacity Design and Planning	Outpatient Appointment Scheduling
Spruit et al. (2014)	Productivity and Process Improvement People, Jobs and Organizations	Quality of Care (Acceptable/Safety), Staffing and Labour Productivity
Testik et al. (2012)	Capacity Design and Planning	Queuing Analyses for Workforce Planning
Zheng et al. (2015)	Productivity and Process Improvement	Quality of Care (Effectiveness)
Zhong et al. (2012)	Capacity Design and Planning	Strategic/Aggregate Healthcare Capacity Planning

with 7 papers each exclusively dealing with 'Capacity Design and Planning' and the 'Work Flow Analyses' and 6 studies dealt with the 'Productivity and Process Improvement' category. The functional area 'People, Jobs and Organizations' had one paper whereas another study; Spruit et al. (2014) had a wide healthcare data mining agenda comprising this category and also, the 'Productivity and Process Improvement'.

Within the sub-dimensions, the 'the Productivity and Process Improvement' data mining only seem to have focused on improving the quality of care. We have used Bengoa et al.

(2006)'s classification of quality of care dimensions to specify the relevant process improvement application. There were seven papers addressing the 'effectiveness', two studies focused on the 'safety' and one attempt was made for acceptable/patient-centred dimension of healthcare quality. Effectiveness of quality requires the health care delivery to adhere to an evidence base and it should result in improved health outcomes (Bengoa et al. 2006). The studies contributed to effective care delivery by predicting the readmission of emergency patients with a pulmonary disease (Demir 2014), with acute myocardial infarction (AMI) (Zheng et al. 2015) and readmission/mortality prediction during hemodialysis (Cornalba et al. 2008). Chi et al. (2008) devised an expert system based on predictive analytics for referring emergency heart patients. Rubrichi and Quaglini (2012) used text mining to extract information from summary of products ensuring correct drugs administration contributing to improved health outcomes. Safety dimension of quality aims to deliver health care by minimizing risks and harm to service users whereas 'Acceptable Quality' takes into account the preferences and aspirations of individual service users (Bengoa et al. 2006). Spruit et al. (2014) and Lee et al. (2011) predicted the safety incidents based on patient demography and medical history, therefore, their work was classified under the safety sub dimension of quality of care. Spruit et al. (2014) had a wider data mining agenda wherein the study also utilized the customer experience and feedback for knowledge discovery that would support better management of Dutch long-term care institutions.

Predicting clinical pathways have also attracted a great deal of attention for discovering the path care delivery takes on ground. Clinical or care pathways can be loosely described as the instruction set or detailed guidelines for the delivery of specific healthcare services. Kinsman et al. (2010) formally defined clinical pathway as an intervention with a structured multidisciplinary plan of care with three out of the four following criteria:

(a) To translate guidelines or evidence into local structures,
(b) To detail the steps in a course of treatment or care in a plan, pathway, algorithm, guideline, protocol or other inventory of actions
(c) The intervention had timeframes or criteria-based progression and
(d) The intervention aimed to standardize care for a specific clinical problem, procedure or episode of healthcare in a specific population.

Data mining for workflow analyses was used to determine the clinical pathways (Caron et al. 2014; Garg et al. 2009; Kudyba and Gregorio 2010; Kuo 2011; Lin et al. 2001; Rebuge and Ferreira 2012) and in one instance, the deviation of actual clinical pathways from the recommended pathways (Rovani et al. 2015), therefore, also contributing to the improved health outcomes. The relationship between determination of clinical pathways and quality gains was also noted by James and Savitz (2011) while reporting significant process improvement at Intermountain Healthcare because the clinical pathways helped the hospital understand their care chain and the associated variations. Apart from discovering and analyzing recurring patterns from the care delivery logs and characterizing process variants, other objectives of determining clinical pathway have been reported to be prediction of length of stay for patients undergoing through a particular treatment/diagnostic test for example radiology (Kudyba and Gregorio 2010) and identifying pathways with high cost and time (Garg et al. 2009). In short, data mining seems a natural fit to capture the complex interdependencies of clinical pathways and generate insights that can lead to more informed decisions. This point also seems to have been registered within the healthcare industry as shown by the numerous reported applications shown in Table 3.

Table 2 reveals Capacity Design and Planning as the third major healthcare operations and supply chain management area that has had data mining algorithms applied to aid manage-

Table 3 The data collection phase

References	Data sources	Sample size	Data description	Observation period
Caron et al. (2014)	Secondary (EHR)	150,291 logs	Event and time log of care flow	3 years
Ceglowski et al. (2007)	Secondary (EHR)	56,906 all ED presentations	Demography, presentation problem, key time points, disposition, medical procedures	1 year
Chi et al. (2008)	Secondary (State Inpatient Dataset)	360,000 data points	Patient Demographics, Institutional characteristics, patient risks, traveling distance, and chances of survival and complications	1 year
Cornalba et al. (2008)	Secondary EHR	10,095 sessions	Patients' demographics, and the treatment data	4 years
Demir (2014)	Secondary (EHR)	963 patients	Demography, disease and treatment data, readmission time	2.75 years
Garg et al. (2009)	Secondary (EHR)	12,085 admission	Event and time log of care flow	16 years
Glowacka et al. (2009)	Primary (data collection) + Secondary (EHR)	1809 (clinic records)	Registration time, triage time, Consultation time Patient visit data demography medical and family history	3 months + 9 months
Kudyba and Gregorio (2010)	Secondary (EHR)	43,000 inpatient cases	Actual Length of Stay, Medicare Allowed LOS, demography, diagnostic and treatment descriptive	2 years
Kuo (2011)	Secondary (EHR)	980 samples	Patient Demography, disease and treatment data	No information
Lavrač et al. (2007)	Secondary data—Public Health Data	No info	Patient Demography, patients' visits to general practitioners and specialists, diseases, Institutional data human resources and availability organization	No information
Lee et al. (2011)	Secondary (EHR)	725 (fall incidents recorded)	Patients' demographics, other fall related data (physical factors, patients' pre-24-h medications, nursing interventions)	3.6 years
Lin et al. (2001)	Secondary (EHR)	45,200 logs	Event and time log of care flow	1 year

Table 3 continued

References	Data sources	Sample size	Data description	Observation period
Malin et al. (2011)	Secondary (EHR)	7,575,434 data points	EHR users information and their assignments, patients records, accesses of patient information	1 year
Ng et al. (2006)	Secondary – Extraction from a retrospective cohort study	692 (length of stay patient record)	Patient demographics, and associated co-morbidities information	3 years
Rebuge and Ferreira (2012)	Secondary (EHR)	179,354 (emergency episodes)	Event and time log of care flow	6 months
Rovani et al. (2015)	Secondary (EHR)	No info	Start/completion of process tasks together with related context data (e.g., actors and resources) and timestamps	No information
Rubrichi and Quaglini (2012)	Secondary – Farmadati Italia	1253 sentences	Manually annotated interaction sections of specialty medicines	N/A
Samorani and Laganga (2015)	Secondary (EHR)	50,000 appointments data	Patient demographics, arrival and service pattern, no show behavior	No information
Spruit et al. (2014)	Primary (data collection) + Secondary (EHR)	5692 incidents	Interviews, incidents information that included attributes such as client, department, date and time, type of incident, cause, location,physical damage and mental damage	4 years
Testik et al. (2012)	Secondary (EHR)	84,094 donors record	Donors arrival rates	3 years
Zheng et al. (2015)	Secondary (EHR)	1641 instances	Patient age, length of stay, admission acuity, comorbidity index score, gender, patient readmission risk, insurance	No information
Zhong et al. (2012)	Secondary (National Data set)	1,228,234.00	Patients 'visits to general practitioners and specialists, diseases, human resources and availability	1 year

EHR electronic health records

rial decision making. There have been 7 studies grouped under the main HOSCM function area and these studies can be further subdivided in to strategic and medium to short term healthcare resource planning, outpatient scheduling and Queuing Analyses for workforce scheduling. The strategic healthcare resource planning has been on a macro scale with studies estimating cost of stay funding policies for at national, state and local levels (Zhong et al. 2012) and planning for public health availability and accessibility (Lavrač et al. 2007). Ng et al. (2006) used data mining for early prediction of patients requiring longer hospital care, therefore, contributing to short to medium term hospital resources planning for inpatient care for planning resources for the outpatient category, Glowacka et al. (2009) and Samorani and Laganga (2015) used historic data to predict the patients attendance behaviour for their appointments and expected service times based on the medical conditions. This information led to optimized capacity allocation of outpatient clinics. Ceglowski et al. (2007) and Testik et al. (2012) used the knowledge gained from data mining as an input in their queuing analyses for workforce scheduling. However, the implementation environments were very different. Ceglowski et al. (2007) used data mining principles to identify group of patients and this information was used as input for their discrete event simulation model for scheduling of work force in an emergency ward. Testik et al. (2012) used data mining to identify patterns that indicated significantly different daily and weekly arrival rates for blood donors and took these factors into consideration to plan an adaptive work schedule for the facility.

For 'People, Jobs and Organizations' category, Malin et al. (2011) investigated the fluid departmental boundaries in terms of access control by automatically mining usage patterns from electronic health record (EHR) systems. Their approach could be used as an audit for predefined policies and also to detect of unknown behaviors for the EHR usage. Spruit et al. (2014) used historical information about the Dutch long term care facilities to predict various financial performance measures such as staffing levels for each operation and for each facility, operations per facility and the facility's prognosis. These predictions enabled the management to control the expenditure and plan for revenue generation.

3.4.2 Data collection

This category was used to collect and collate information regarding the data sources, sample size, data description and observation period (Table 3). Most studies used the information stored in the electronic health records (EHRs) of hospitals/medical centres whereas in a few cases such as Zhong et al. (2012) used dataset on United States healthcare cost & utilization to predict clinical charge profiles, Chi et al. (2008) used the Iowa state Inpatient dataset (SID) for their expert referral system and Rubrichi and Quaglini (2012) used Italian national pharmacy database for their text mining project. There also have been two instances where along with the EHRs data, actual data was also collected for the project. Spruit et al. (2014) interviewed senior management to clarify the data mining objectives for their project and Glowacka et al. (2009) used a barcode system to keep track of triage time and the consultation time.

The type of data collected for each project was largely dependent upon the healthcare setting (Table 1: Business Understanding). For example, for determining the clinical pathways, event and time log of care flow were collected that generally comprise of patient identifiers, activity type, the time sequencing, functional area, and information on both the diagnosis and treatment type. For improving quality of care, patient demography, medical history, incident information and risks were the main attributes used for prediction purposes. Similarly, for

two sub categories of capacity planning namely; queuing analyses and outpatient scheduling, patients' arrival and service data was of primary concern (Table 3).

3.4.3 Modeling

Different modelling paradigms and tools for the reviewed literature are mentioned in Table 4. The findings tend to confirm the widely reported practice in the data mining community that various modeling techniques have been typically used for one project with two algorithms for the same data mining problem seems to be the norm (Lee et al. 2011; Lin et al. 2001; Rebuge and Ferreira 2012; Testik et al. 2012; Zheng et al. 2015) but up to 4 different algorithms for one project have also been reported (Demir 2014). Possible explanations for this modeling tendency is that a performance comparison of the employed algorithms can help choose the optimal model. There also seems to be a pattern in the modelling approaches depending upon the business understanding (Table 1). The clinical pathways are mainly concerns with determination of event and time dependent sequences with various process mining approaches being the preferred modelling approaches (Caron et al. 2014; Rovani et al. 2015). Sequence pattern mining has also been used multiple times for determination of care pathways (Garg et al. 2009; Lin et al. 2001; Rebuge and Ferreira 2012). Regarding the software tools, the trend seems to have been equally divided into the development of in house coded software and utilizing the existing platforms such as R, SPSS, Weka and Pro M.

3.4.4 Evaluation

This phase is mainly concerned with determining the performance of the modelling approach. Performance measurements are generally defined as regular measurement of outcomes and results, which generates reliable data on the effectiveness and efficiency of models (Moullin 2007). The performance evaluation of a model is considered the most important stage in a data mining implementation.

Throughout this stage, the developed model is tested over the collected datasets and the performance evaluation metrics are applied. The metrics determine the extent to which the model is well-learned and whether the results are robust and reliable in its prediction. Generally, following three types of indicator measures are used to assess the reliability of a data mining implementation (Lessmanna et al. 2013):

1. Measures that assess the predictive power of the model (e.g., Accuracy, Sensitivity).
2. Measures that assess the discrimination power of the model (e.g. Area Under Curve (AUC)).
3. Measures that assess the accuracy of the predictions' probabilities of the model (e.g. Brier Score).

These measures provide a comprehensive view on the developed model performance. As it can be seen in Table 5, most studies applied various performance measures to evaluate their models, therefore, we categorized the evaluation phase in to four measures based on the methods used in the selected studies namely; abstract measures with and without ranking consideration, field relative measures and model validation. Abstract measures that do not take ranking in to consideration only depend on prediction label that convert or assign the ranking or the probability of a classifier to a certain class or label after applying a threshold which is a predetermined value that helps in assigning an instance to a particular class (e.g. Accuracy, Sensitivity, Specificity etc.). However, this may not be sufficient to establish the quality of prediction, therefore, some studies have opted for abstract measures that take into

Table 4 The modeling phase

References	Data mining algorithm	Tool/software
Caron et al. (2014)	Process Mining Algorithms, Custom Patterns and Process Exploration	Heuristics miner, LTL-checker, Trace Alignment, Performance Sequence Diagram, Social Network Miner
Ceglowski et al. (2007)	Discrete Event Simulation model, SOM Clustering guided by Ward's Hierarchical Agglomeration, Simulation	Viscovery SOMine
Chi et al. (2008)	Support Vector machines (SVMs)	In-house Development
Cornalba et al. (2008)	Bayesian Network	Hugin Package, In-house Development
Demir (2014)	Logistic Regression, Regression Trees, Generalized Additive Models (GAMs), Multivariate Adaptive Regression Splines (MARS)	R Studio
Garg et al. (2009)	Markov Models for Sequential Pattern Mining	In-house Development
Glowacka et al. (2009)	Association Rule Mining (ARM), Simulation	SPSS Clementine 10, SimProcess 4.2
Kudyba and Gregorio (2010)	Neural Networks	No information provided
Kuo (2011)	Association Apriori Algorthim	WEKA
Lavrač et al. (2007)	Descriptive data mining methods- Clustering (Agglomerative Classification. Principal Component Analysis, the Kolmogorov–Smirnov Test)	WEKA, In-house Development (Medimap)
Lee et al. (2011)	Neural networks. Logistics Regression	SPSS, STATISTICA 8.0 (StatSoft, Tulsa, OK).
Lin et al. (2001)	Algorithm of mining time dependency(Sequential Pattern Mining), Association Analysis	In-house Development
Malin et al. (2011)	Association Rules, Social Network Analysis	In-house Development
Ng et al. (2006)	Neural networks (incremental learning algorithm)	In-house Development
Rebuge and Ferreira (2012)	Sequence Clustering, Social Network Analysis, Petri Net	In-house Development
Rovani et al. (2015)	Declarative Modeling Languages (Process Mining),, LTL	ProM and Declare
Rubrichi and Quaglini (2012)	Conditional Random Fields (CRFs, Structural Support Vector Machines (SVMs)	In-house Development
Samorani and Laganga (2015)	Cost-Sensitive Classification, Bayesian classifier	WEKA
Spruit et al. (2014)	Association Rule Mining, Apriori algorithm,	R Studio
Testik et al. (2012)	Clustering, Classification, and Regression Tree,	Clementine
Zheng et al. (2015)	Neural Networks, Support Vector Machines (SVM), Particle Swarm Optimization (PSO), Random forest,(RF) algorithm, and the hybrid model of swarm intelligence heuristic	In-house Development
Zhong et al. (2012)	Multi-Level Support Vector Machine (SVM)	In-house Development

account the classifier ranking or probability indicating the confidence of the classifier of its outcome (e.g. AUC, Confidence measure and Brier score). Other performance measures apply classifiers predictions into specific fields such as economy, government or healthcare in order to assess the performance of their models or approaches. We call these as 'field-relative measures'.

Model validation (Sargent 2005; Carson 2002)) is a more wide and complex concept than calculating measures. It is concerned with building the correct model regardless of all the modelling steps were carried out successfully or not. It is utilized to determine if a model is an accurate representation of the real system. Validation is usually achieved through the calibration of the model, which is an iterative process of comparing the model to actual system behavior and using the discrepancies between the two, and the insights gained, to improve the model further. This process is repeated until model accuracy is judged to be acceptable. For some models it is sufficient for validation to calculate basic measures like Accuracy or AUC, while other models need objective validation such as statistical test (e.g. Kolmogorov-Smirnov test), model comparison, and subjective measures such as Field-Experts opinion (Sargent 2005)

As shown in Table 5 several studies have applied only one evaluation measure (Caron et al. 2014; Ceglowski et al. 2007; Garg et al. 2009; Kuo 2011; Spruit et al. 2014). On the other hand, some studies used a combination of two or more (e.g. Chi et al. 2009; Cornalba et al. 2008; Lavrač et al. 2007; Zheng et al. 2015). It is worth noting that most of the studies focused on validating its models to ensure its fitness and acceptability. Also we note that the field relative evaluation measures for healthcare sector settings were also widely used. The abstract measures were only adopted in studies which carried out classification tasks.

3.4.5 Deployment

The deployment phase employs the results of the study. The CRISP-DM suggests a wide spectrum of deployment phase from generating a reported from informed decision-making to integrating data mining with the information system and databases to implementing a repeatable data mining process. Therefore, we defined two sub categories namely; Concept Realization and Actual Deployment:

- Concept realization is a process of creating working model and reviewing that all its parts are working reliably ((Yao et al. 2008)).
- Actual Deployment is a live implementation of the model to assist end users in their daily operations. Here the end users are not required to have working knowledge o] the internal structure of the model, however, they will need to understand up front the possible outcomes and actions (Patterson 2009).

As illustrated in Table 6, while all studies achieved concept realization, only 8 out of 22 data mining implementations can be considered as 'actually deployed'. The model is considered deployed if authors created ready-for-use software or the studies claimed that their approach is practically tested in a real working environment such as in hospitals or clinics. For example, Ng et al. (2006) extracted data from a retrospective cohort study to check if their model can be used for an on-line prediction of length of stay, therefore, we classify it as a study with a focus on concept realization whereas Ceglowski et al. (2007) is categorized as actual deployment because their approach was employed at an emergency department (ED). Similarly, Chi et al. (2008) used data that was linked to hospital descriptive data from the American Hospital Association (AHA) by a hospital identification number, therefore, their model is categorized as 'deployed'. They proposed a hospital referral expert system to assist

Table 5 The evaluation modeling phase

Paper	Abstract (without taking ranking into consideration)	Abstract (with taking ranking into consideration)	Field-relative (measures used in hospitals)	Validation
Caron et al. (2014)	–	–	–	Model Comparison
Ceglowski et al. (2007)	–	–	–	Expert, Kolmogorov–Smirnov, Model Comparison
Chi et al. (2008)	–	Mean square error of predicted probability	Desired hospital outcome	-
Cornalba et al. (2008)	–	–	Adherence to treatment, Hospitalization risk	Model Comparison
Demir (2014)	Sensitivity, Specificity	AUC, Brier score, R_N^2	-	-
Glowacka et al. (2009)	–	–	The doctors' idle time, the nurses' idle time, the doctors' overtime, the nurses' overtime, the total patient waiting time, and the total number of patients seen	IF-Then rules (Association rules), Simulation
Garg et al. (2009)	–	–	–	Model Comparison
Kudyba and Gregorio (2010)	–	–	Length of Stay	–
Kuo (2011)	–	–	Waiting time for consultation	–
Lavrač et al. (2007)	–	–	Availability of health services for patients, Rate of accesses to health services,	Expert, Model Comparison

Table 5 continued

Paper	Abstract (without taking ranking into consideration)	Abstract (with taking ranking into consideration)	Field-relative (measures used in hospitals)	Validation
Lee et al. (2011)	PPV, NPV	ROC	–	Model Comparison
Lin et al. (2001)	Accuracy	–	–	–
Malin et al. (2011)	–	Confidence measure	–	–
Ng et al. (2006)	–	Mean Absolute Difference (MAD)	–	–
Rebuge and Ferreira (2012)	–	–	–	Model Comparison
Rovani et al. (2015)	–	–	–	Model Comparison
Rubrichi and Quaglini (2012)	Accuracy, F1-measure, Recall, Precision	–	–	Model Comparison
Samorani and Laganga (2015)	–	–	Exp. Hospital profit, Exp. overtime, Exp. waiting time, Overall show rate	–
Spruit et al. (2014)	–	–	–	Expert
Testik et al. (2012)	–	–	Average donors waiting time	IF-Then rules (rule inductions)
Zheng et al. (2015)	Accuracy, Sensitivity, Specificity	–	–	Model Comparison
Zhong et al. (2012)	Accuracy, MCC	AUC	–	Model Comparison

PPV positive predicted values, *NPV* negative predicted values, *ROC* receiver operating characteristic, *AUC* area under curve, *MCC* matthews correlation coefficient

Table 6 The Deployment Phase

References	Concept realization	Actual deployment	Healthcare setting	Geographical location
Caron et al. (2014)	Yes	No model deployment information	Inpatients—Gynecological Oncology	The Netherlands
Ceglowski et al. (2007)	Yes	Model deployed	Emergency—Queuing Analyses	Australia
Chi et al. (2008)	Yes	Model deployed	Emergency-Referral system for Acute Myocardial Infarction (AMI)	USA
Cornalba et al. (2008)	Yes	No model deployment information	In Patient—Hemodialysis (Urology)	Italy
Demir (2014)	Yes	No model deployment information	Emergency-Readmission with Chronic Obstructive Pulmonary Disease	UK
Garg et al. (2009)	Yes	No model deployment information	Inpatients Geriatrics Department	UK
Glowacka et al. (2009)	Yes	Model deployed	Outpatient Overbooking with No Shows	USA
Kudyba and Gregorio (2010)	Yes	No model deployment information	In-Patients with Radiology Pathways	USA
Kuo (2011)	Yes	Model deployed	Inpatients-Geriatrics Taiwan	Taiwan
Lavrač et al. (2007)	Yes	No model deployment information	Public Health Resource Allocation	Slovenia
Lee et al. (2011)	Yes	No model deployment information	Inpatients	Taiwan
Lin et al. (2001)	Yes	Model deployed	Inpatient-Brain Stroke Clinical Pathways	Taiwan
Malin et al. (2011)	Yes	Model deployed	Electronic Health Records Logs Usage	USA
Ng et al. (2006)	Yes	No model deployment information	Inpatients-Length of Stay Prediction for Pediatrics Gastroenteritis	Australia

Table 6 continued

References	Concept realization	Actual deployment	Healthcare setting	Geographical location
Rebuge and Ferreira (2012)	Yes	Model deployed	Emergency-Clinical Pathways	Portugal
Rovani et al. (2015)	Yes	Model deployed	Inpatient—Urology Department	The Netherlands
Rubrichi and Quaglini (2012)	Yes	No model deployment information	Pharmacology -Extracting Content	Italy
Samorani and Laganga (2015)	Yes	No model deployment information	Outpatient Overbooking with No Shows	USA
Spruit et al. (2014)	Yes	No model deployment information	Inpatients—Long Term Care Safety	The Netherlands
Testik et al. (2012)	Yes	No model deployment information	Blood Collection	Turkey
Zheng et al. (2015)	Yes	No model deployment information	Inpatients/Emergency-Acute Myocardial Infarction (AMI)	China
Zhong et al. (2012)	Yes	No model deployment information	Strategic/Aggregate Healthcare Capacity Planning	USA

in assigning the best hospital for the patients based on their physical conditions and travel distance required. Glowacka et al. (2009) investigated the effects of several scheduling policies on the clinic's profit by using different rule sets for no-shows. In addition, they investigated their approach in real environment by simulating the problem using optimization software, which is an indication that their concept was fully deployed. Similarly, Rebuge and Ferreira (2012) developed a tool based on business process mining for healthcare environment. The tool collected data from the hospital information system and provide set of process mining techniques for the analysis of selected healthcare processes. The proposed method was empirically applied in a hospital in Portugal. Furthermore, Rovani et al. (2015) proposed a methodology to investigate the difference between expected and actual process behaviors of clinical processes. To implement the methodology, declarative software and graphical interfaces were used to develop a ready use software for practitioners. The model was deployed by applying it in the urology department in a hospital in Netherlands. Kuo (2011) developed a system that integrated information technology and medical-related technologies to develop a comprehensive geriatric assessment healthcare information system for geriatric consultation services. Lin et al. (2001) developed a model that extracted time dependency patterns to discover new clinical pathways for new brain stroke patients. Their model used visualization interface to display the mining results, therefore, their system can be considered ready for deployment. Malin et al. (2011) took another approach, they have empirically evaluated their methods with a special software developed and managed by the Vanderbilt University Medical Center.

Table 6 also reveals the healthcare settings with most applications dealing with in-patient and also their geographical locations. Not surprisingly, USA, Australia and Europe seems to be leading the way in data mining application for HOSCM with 18 papers. Taiwan also seem to be employing the state of the art technologies in delivery of healthcare with 3 papers and 1 paper described data mining for Turkish blood donation supply chain.

4 Discussion

We conducted a research synthesis on the applications of data mining and predictive analytics for the delivery of healthcare services. Rousseau et al. (2008) identified a systematic accumulation, analysis and a reflective interpretation of the full body of relevant empirical evidence related to the research question as the essential components in a good scientific practice of systematic review. In Sect. 3, we employed a systematic search for accumulating data and the use of analytic categories as suggested by Seuring and Gold (2012) for analyzing the collected data. In this section, we follow Saenz and Koufteros (2015) structured approach for the reflective interpretation by highlighting the existing gaps in the reviewed body of knowledge and also, by raising new suggestions for further investigations.

Healthcare operations and supply chain management (HOSCM) is a very broad concept that include all activities that are necessary to provide care to the patients as shown in the Table 1. However, our data analyses shows that most of the data mining applications have focussed on a narrow subset of HOSCM functions such as quality of care, identification of care pathways and patients and workforce scheduling, All these application areas are important for patient safety and also, for matching the scarce healthcare capacity with an ever increasing demand but attempts to use the power of big data for other HOSCM activities such as the physical layout design and the process analyses would also have contributed to improving healthcare productivity. This seems a limitation of the reviewed literature which

may also be termed as unexpected because we have found evidence that there is a growing emphasis on capturing patient pathways and further analyses of the collected data may have led to optimal healthcare facilities layout design and resource planning. The focus of the care flow analyses has been to identify and analyse recurring patterns in patient flows (Kudyba and Gregorio 2010; Lin et al. 2001) and to characterize any deviations from the guidelines (Caron et al. 2014; Rebuge and Ferreira 2012; Rovani et al. 2015). Garg et al. (2009)'s algorithm highlighted the cost structures of multiple pathways but these insights can also be used to restructure the business processes and their layouts for better healthcare outcomes. A plausible explanation for this apparent limitation can be that for our study sample the clinical pathways were collected for partial segments of the care delivery chain, therefore, limiting the use of data driven approaches to the design of entire care chain. Furthermore, the captured data apparently does not seem to be integrated with the healthcare information systems to a degree that could have facilitated advanced process analyses. For example, Kudyba and Gregorio (2010) focussed only on the data mining for radiology exams and Cornalba et al. (2008) for haemodialysis which covers only a narrow part of the care delivery chain. It seems that the model extension to include a complete set of relevant care delivery process was either deemed too complicated or time-consuming for the reviewed studies. Nevertheless, this seems the next logical step and a promising future research area of extending the models to include end to end care delivery processes allowing the captured patient pathways to be used for ongoing healthcare decision making.

In addition to the data mining healthcare applications, this study also examined the data management, modelling and evaluation adopted by the reviewed body of knowledge revealing some apparent areas of improvement. It appears that the modelling trend has been to develop predictive models using simple classifiers like support vector machine (SVM) and neural networks but more complex novel classifiers like Random Forest and multivariate adaptive regression splines (MARS) have been reported to give a superior performance. Furthermore, some of the studies (Ceglowski et al. (2007); Rebuge and Ferreira 2012) did not reveal the model verification information (training/data sets) which restricts the evaluation of the proposed approaches. In Sect. 3.4.4, we differentiated between model validation from model evaluation as an iterative process of comparing the model to actual system behavior or comparing with other modeling approaches for modelling performance improvement. Validation can be internal i.e. comparing between various models developed and external, for example, measuring important real characteristics of the model like cost of patient treatment or length of this treatment. The studies with only internal validation could have strengthened their research findings by employing external validations measures. In a similar vein, the generalizability of the developed models could have been increased if the same approach was successfully implemented in other similar settings. Finally, as pointed earlier, about two third of reviewed studies were not deployed which limits the potential benefits of using big data for healthcare operation management. The factors that can help bridge the gap between concept realization and actual deployment of predictive analytics and data mining applications for healthcare seems to be an area that merits further investigation.

5 Conclusion

The results presented in this study are useful in providing a holistic view of how data mining and predictive analytics have contributed to the delivery of healthcare services. From the examined literature, it has been demonstrated that useful research has been conducted but

data mining in healthcare is still an emerging area of a research with significant potential. While some areas of healthcare operations have attracted more attention than others like characterization of clinical pathways, quality of care and resource allocation but there is scope for using insights generated from the use of big data to improve the facilities layout design and process analyses for operational excellence and to improve patients' satisfaction. Towards this endeavor, this study is unique in its perspective and in advancing research in this field in several ways. Firstly, it examines existing literature and adapts the existing HOSCM frameworks to be more representative of the functions/processes associated with healthcare delivery. Secondly, it provides a multi-layered framework for studying data mining applications for the healthcare operations management. Thirdly, this study synthesizes and classifies the research in the area in the proposed multi-layered framework by using two pronged approach of deductive and inductive fitting. Furthermore, the final framework reveals three major data mining applications for healthcare delivery namely; determination of clinical pathways, healthcare capacity planning and improving the quality of care. The derived knowledge from the identification of clinical pathways has the potential to provide optimal process layout design and analyses contributing further to improved healthcare outcomes. These findings have important managerial ramifications for the healthcare practitioners as they can use the synthesized information in this study as a reference for introducing the data mining measures in their own organizations. Similarly, data scientists and practitioners could benefit from the patterns and associations reported for the data management, modelling, evaluation and deployment stages. Above all, this study can be used as a platform for other academic in the area as a starting point for their own research.

Lastly, we would like to point out that despite following a structured approach supported by the literature, we cannot claim that we have accessed all published material in this area. Our search process was guided by the combination of selected keywords as detailed in Sect. 3 which gives us reasonable confidence that we have been able to review a large body of knowledge on healthcare operations and supply chain data mining.

Acknowledgements This research is funded by a grant from the Centre of Sustainable Processes, Abu Dhabi University, United Arab Emirates.

References

Agarwal, R., & Dhar, V. (2014). Editorial-Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, *25*, 443–448.

Amarasingham, R., Patzer, R. E., Huesch, M., Nguyen, N. Q., & Xie, B. (2014). Implementing electronic health care predictive analytics: Considerations and challenges. *Health Affairs*, *33*, 1148–1154.

Anderson, J. E., & Chang, D. C. (2015). Using electronic health records for surgical quality improvement in the era of big data. *JAMA Surgery*, *150*, 24–29.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, *33*, 1123–1131.

Bengoa, R., Kawar, R., Key, P., Leatherman, S., Massoud, R., & Saturno, P. (2006). *Quality of care: A process for making strategic choices in health systems*. Geneva: World Health Organization. WHO press.

Bonacina, S., Masseroli, M. & Pinciroli, F. (2005). *Foreseeing promising bio-medical findings for effective applications of data mining*. Biological and Medical Data Analysis, Springer.

Caron, F., Vanthienen, J., Vanhaecht, K., van Limbergen, E., de Weerdt, J., & Baesens, B. (2014). Monitoring care processes in the gynecologic oncology department. *Computers in Biology and Medicine*, *44*, 88–96.

Carson, J. S. (2002). Model verification and validation. In *Proceedings of the Winter Simulation Conference* (pp. 52–58), IEEE.

Ceglowski, R., Churilov, L., & Wasserthiel, J. (2007). Combining data mining and discrete event simulation for a value-added view of a hospital emergency department. *Journal of the Operational Research Society*, *58*, 246–254.

Chi, C.-L., Street, W. N., & Ward, M. M. (2008). Building a hospital referral expert system with a prediction and optimization-based decision support system algorithm. *Journal of biomedical informatics*, *41*, 371–386.

Cornalba, C., Bellazzi, R. G., & Bellazzi, R. (2008). Building a normative decision support system for clinical and operational risk management in hemodialysis. *IEEE Transactions on Information Technology in Biomedicine*, *12*, 678–686.

Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems*, *55*, 359–363.

Demir, E. (2014). A decision support tool for predicting patients at risk of readmission: A comparison of classification trees, logistic regression, generalized additive models, and multivariate adaptive regression splines. *Decision Sciences*, *45*, 849–880.

Dobrzykowski, D., Deilami, V. S., Hong, P., & Kim, S.-C. (2014). A structured analysis of operations and supply chain management research in healthcare (1982–2011). *International Journal of Production Economics*, *147*, 514–530.

Dubey, R., Gunasekaran, A., Childe, S. J., Wamba, S. F., & Papadopoulos, T. (2016). The impact of big data on world-class sustainable manufacturing. *The International Journal of Advanced Manufacturing Technology*, *84*, 631–645.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*, 37.

Garg, L., McClean, S., Meenan, B., & Millard, P. (2009). Non-homogeneous Markov models for sequential pattern mining of healthcare data. *IMA Journal of Management Mathematics*, *20*, 327–344.

Gheorghe, M., & Petre, R. (2014). Integrating data mining techniques into telemedicine systems. *Informatica Economica*, *18*, 120–130.

Glowacka, K. J., Henry, R. M., & May, J. H. (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society*, *60*, 1056–1068.

Harper, P. (2005). Combining data mining tools with health care models for improved understanding of health processes and resource utilisation. *Clinical and Investigative Medicine*, *28*, 338.

Haux, R., Ammenwerth, E., Herzog, W., & Knaup, P. (2002). Health care in the information society. A prognosis for the year 2013. *International Journal of Medical Informatics*, *66*, 3–21.

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, *154*, 72–80.

James, B. C., & Savitz, L. A. (2011). How Intermountain trimmed health care costs through robust quality improvement efforts. *Health Affairs*, *30*, 1185–1191.

Kinsman, L., Rotter, T., James, E., Snow, P., & Willis, J. (2010). What is a clinical pathway? Development of a definition to inform the debate. *BMC medicine*, *8*, 1.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management*, *19*, 65.

Kontio, E., Airola, A., Pahikkala, T., Lundgren-Laine, H., Junttila, K., Korvenranta, H., et al. (2014). Predicting patient acuity from electronic patient records. *Journal of Biomedical Informatics*, *51*, 35–40.

Kudyba, S., & Gregorio, T. (2010). Identifying factors that impact patient length of stay metrics for healthcare providers with advanced analytics. *Health Informatics Journal*, *16*, 235–245.

Kuo, N.-W. (2011). Information technology applications for geriatric consultation services in Taiwan. *International Journal of Advancements in Computing Technology*, *3*, 44–52.

Langabeer, J. R, I. I., & Helton, J. (2015). *Health care operations and systesm management. Health care operations management a systems perspective* (2nd ed.). Burlington, MA: Jones and Bartlett Publishers.

Lavrač, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M., & Kobler, A. (2007). Data mining and visualization for decision support and modeling of public health-care resources. *Journal of Biomedical Informatics*, *40*, 438–447.

Lee, T.-T., Liu, C.-Y., Kuo, Y.-H., Mills, M. E., Fong, J.-G., & Hung, C. (2011). Application of data mining to the identification of critical factors in patient falls using a web-based reporting system. *International Journal of Medical Informatics*, *80*, 141–150.

Lessmanna, S., Seowb, H., Baesenscd, B. & Thomasd, L. C. (2013). Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. In *Credit Research Centre, Conference Archive*.

Lin, F.-R., Chou, S.-C., Pan, S.-M., & Chen, Y.-M. (2001). Mining time dependency patterns in clinical pathways. *International Journal of Medical Informatics*, *62*, 11–25.

Malik, M. M., Khan, M., & Abdallah, S. (2015). Aggregate capacity planning for elective surgeries: A bi-objective optimization approach to balance patients waiting with healthcare costs. *Operations Research for Health Care*, *7*, 3–13.
Malin, B., Nyemba, S., & Paulett, J. (2011). Learning relational policies from electronic health record access logs. *Journal of Biomedical Informatics*, *44*, 333–342.
Menon, A. K., Jiang, X., Kim, J., Vaidya, J., & Ohno-Machado, L. (2014). Detecting inappropriate access to electronic health records using collaborative filtering. *Machine Learning*, *95*, 87–101.
Moullin, M. (2007). Performance measurement definitions: Linking performance measurement and organisational excellence. *International Journal of Health Care Quality Assurance*, *20*, 181–183.
Ng, S.-K., McLachlan, G. J., & Lee, A. H. (2006). An incremental EM-based learning approach for on-line prediction of hospital resource utilization. *Artificial Intelligence in Medicine*, *36*, 257–267.
Papadopoulos, T., Gunasekaran, A., Dubey, R., Altay, N., Childe, S. J., & Fosso-Wamba, S. (2016). The role of Big Data in explaining disaster resilience in supply chains for sustainability. *Journal of Cleaner Production*, *142*, 1108–1118.
Patterson, J. R. F. (2009). Handbook of systems engineering and management. In A. P. Sage & W. B. Rouse (Eds.), *System engineering life cycles: Life cycles for research, development, test and evaluation; acquisition; and planning and marketing*. Hoboken: Wiley.
Porter, M. E., & Teisberg, E. O. (2006). *Redefining health care: Creating value-based competition on results*. Brighton: Harvard Business Press.
Rebuge, Á., & Ferreira, D. R. (2012). Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, *37*, 99–116.
Rousseau, D. M., Manning, J., & Denyer, D. (2008). 11 Evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses. *The Academy of Management Annals*, *2*, 475–515.
Rovani, M., Maggi, F. M., de Leoni, M., & van der Aalst, W. M. (2015). Declarative process mining in healthcare. *Expert Systems with Applications*, *42*, 9236–9251.
Rubrichi, S., & Quaglini, S. (2012). Summary of Product Characteristics content extraction for a safe drugs usage. *Journal of Biomedical Informatics*, *45*, 231–239.
Saenz, M. J., & Koufteros, X. (2015). Special issue on literature reviews in supply chain management and logistics. *International Journal of Physical Distribution and Logistics Management*. doi:10.1108/ijpdlm-12-2014-0305.
Samorani, M., & Laganga, L. R. (2015). Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research*, *240*, 245–257.
Sargent, R. G. (2005). Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation, winter simulation conference* (pp. 130–143).
Seuring, S., & Gold, S. (2012). Conducting content-analysis based literature reviews in supply chain management. *Supply Chain Management: An International Journal*, *17*, 544–555.
Shapiro, G., & Markoff, G. (1997). *Methods for drawing statistical inferences from text and transcripts* (pp. 3–31). Lawrence Erlbaum Associates, Mahwah, NJ: Text Analysis for the Social Sciences.
Siau, K. (2003). Health care informatics. *IEEE Transactions on Information Technology in Biomedicine*, *7*, 1–7.
Spiegel, J. R., Mckenna, M. T., Lakshman, G. S., & Nordstrom, P. G. (2013). *Method and system for anticipatory package shipping*, USA patent application.
Spruit, M., Vroon, R., & Batenburg, R. (2014). Towards healthcare business intelligence in long-term care: An explorative case study in the Netherlands. *Computers in Human Behavior*, *30*, 698–707.
Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*, 250–255.
Testik, M. C., Ozkaya, B. Y., Aksu, S., & Ozcebe, O. I. (2012). Discovering blood donor arrival patterns using data mining: A method to investigate service quality at blood centers. *Journal of Medical Systems*, *36*, 579–594.
Vissers, J., & Beech, R. (2005). *Health operations management: Patient flow logistics in health care*. Abingdon-on-Thames: Routledge.
Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*, 77–84.
Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246.

Wang, G., Gunasekaran, A., Ngai, E. W. T., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98–110.

Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Citeseer,* (pp. 29–39).

Yao, Y., Zhong, N., & Zhao, Y. (2008). *A conceptual framework of data mining. Data Mining: Foundations and Practice*. Berlin: Springer.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., et al. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, *36*, 2431–2448.

Zheng, B., Zhang, J., Yoon, S. W., Lam, S. S., Khasawneh, M., & Poranki, S. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications*, *42*, 7110–7120.

Zhong, W., Chow, R., & He, J. (2012). Clinical charge profiles prediction for patients diagnosed with chronic diseases using Multi-level Support Vector Machine. *Expert Systems with Applications*, *39*, 1474–1483.