Computers in Human Behavior 30 (2014) 698–707

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Towards healthcare business intelligence in long-term care An explorative case study in the Netherlands

Marco Spruit *, Robert Vroon, Ronald Batenburg

Utrecht University, Department of Information and Computing Sciences, Princetonplein 5, 3584 CC Utrecht, The Netherlands

ARTICLE INFO

Article history:
Available online 24 August 2013

Keywords:
Healthcare business intelligence
Knowledge discovery
Data mining
Crisp-dm
Long-term care

ABSTRACT

This research contributes to the domain of long-term care by exploring knowledge discovery techniques based on a large dataset and guided by representative information needs to better manage both quality of care and financial spendings, as a next step towards more mature healthcare business intelligence in long-term care. We structure this exploratory research according to the steps of the CRoss Industry Standard Process for Data Mining (CRISP-DM) process. Firstly, we interview 22 experts to determine the information needs in long-term care which we, secondly, translate into 25 data mining goals. Thirdly, we perform a single case study at a Dutch long-term care institution with around 850 clients in five locations. We analyze the institution's database which contains information from April 2008 to April 2012 to identify patterns in incident information, patterns in risk assessment information, the relationship between risk assessments and incident information, patterns in the average duration of stay, and we identify and predict Care Intensity Package (ZZP) combinations. Fourth and finally, we position all data mining goals in a two-by-two matrix to visualize the relative importance of each goal in relation to both quality of care and financial state of care institutions.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction: long-term care in the Netherlands

This research uncovers the relatively unexplored long-term care sector in the Netherlands and the applicability of knowledge discovery techniques as a next step towards the strategic goal of more mature healthcare business intelligence (Mettler & Vimarlund, 2009). During the last fifteen years knowledge discovery has evolved in the healthcare domain from predicting epidemics (Prather et al., 1997) to a broad spectrum of data mining applications (Koh & Tan, 2011). The data that are being captured by healthcare organizations is enormous in size and, therefore, a treasure for data analysts (Lucas, 2004). Koh and Tan (2011) argue that the use of knowledge discovery techniques have become increasingly popular, if not essential for healthcare organizations. However, within the healthcare domain the long-term care sector has been left out in research, so it seems. Up until now.

In 2008 the following general policy for long-term care in the Netherlands was formulated by Mot, Aouragh, Groot, and Mannaerts (2010):

"To ensure that for persons with a long-term or chronic disorder of a physical, intellectual or psychological nature, care of good quality is available and that the cost level of this care is acceptable to society."

Long-term care is care for people with a long-term or chronic disorder, where the chronic disorder can be either of a physical, intellectual or psychological nature. The policy contains two goals for long-term care, which are 'care of good quality' at an 'acceptable cost level'. These goals apply to all care institutions that deliver long-term care. In order to support these goals, care institutions should have insight in the quality and financial state of the internal organization.

Electronic Client Record (ECR) software is used to keep track of the quality and financial state of the internal organization. All the information stored in ECR software is client related, which makes the client the central entity. ECR software contains personal details, medical information, financial information, production information, care plan, incidents, documents, treatment plans, presence and absence. One should note that at least in the Netherlands, ECR software is different than Electronic Patient Record (EPR) software, which is mainly used in hospitals. One could argue that the two are closely related, but one overarching system is unfortunately not yet in place in the Netherlands at this moment. Both ECR and EPR systems are tailor-made for the sector in which they are used.

Long-term care within the Netherlands has become one of the biggest expenses at this moment for the Dutch government, consuming no less than 38% of the total healthcare budget (Schäfer et al., 2010). The expenditures of the Exceptional Medical Expenditures Act (AWBZ) alone have steadily increased from €14 billion in 2000 up to €27 billion (budgeted) for 2012, which is a doubling in just 12 years (Ministerie van Volksgezondheid & Welzijn en Sport,

* Corresponding author. Tel.: +31 30 253 3708.
E-mail address: m.r.spruit@uu.nl (M. Spruit).

0747-5632/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.chb.2013.07.038

2011a, 2011b). According to the population forecast of the Dutch Central Bureau for Statistics (CBS) in 2010, the number of elderly in the Netherlands will increase from 2.4 million to 4.6 million in 2040 (Duin & Garssen, 2010). This implies that a growing number of people will need some form of care, which is also addressed in (Mot et al., 2010). Also, as life expectancy increases, more people will need some form of care for a longer period of time (Duin & Garssen, 2010). From a governmental perspective it is therefore important to increase the efficiency of long-term care in order to avoid that long-term care expenditures will become uncontrollable. The Ministry of Health, Welfare and Sport (2011) also addressed the importance of management information on quality:

> *"In order to supply care of good quality, it is essential that the care institution is managed well and that the institution has permanent information about the quality of the institution."*

From the perspective of long-term care institutions it is important to have a better insight into the internal organization. At this moment ECR software is mainly used to support the core process of care institutions, which is the delivery of care. The ECR systems contain a lot of valuable data that are nowadays used to receive budget from the government and, moreover, to justify their existence. ECR data are also used on an individual basis to support a client in the best possible way. However, at this moment, the data that are collected in ECR software are not fully exploited yet. If only the collected unstructured data could be made explicit to some extent, the information arising therefrom could then be used to improve the efficiency and effectivity of the long-term care processes (Feelders, Daniels, & Holsheimer, 2000).

Therefore, this research aims to firstly discover the information needs of long-term care institutions, and secondly, the extent to which the desired information needs can be made explicit based on a wide range of already proven knowledge discovery techniques, which is captured in the following research question:

> *"How can knowledge discovery techniques support Dutch long-term care institutions to manage their internal organization?"*

Note that we pursue a meta-algorithmic approach in this research. Instead of focusing on developing new algorithms to mine for new insights, we aim to uncover new knowledge by reusing already proven algorithms in new configurations and application domains. In general, a meta-algorithmic knowledge discovery approach provides an informatics perspective onto data analytics research by modelling knowledge discovery *technology* selection to facilitate *process* analysis to improve *people*'s performance.

The remaining part of this paper is structured as follows. Section 2 provides a description of the data material under investigation and the research approach. In Section 3 the Top 5 data mining goals are modelled and their outcomes are elaborated upon. An overall interpretation of these findings is provided in Section 4. Section 5 contains our main conclusions and further discussion of this work.

2. Material and methods: a data-driven case study

Knowledge discovery techniques embed data mining models within an overarching application process to help discover new interesting knowledge from unstructured data. However, Knowledge discovery and Data mining are often used as synonyms by many researchers (Kurgan & Musilek, 2006). In this research data mining is used as one step in an encompassing knowledge discovery process (Cios, Pedrycz, & Swiniarski, 2007). We have structured this exploratory research according to the steps of the CRoss Industry Standard Process for Data Mining (CRISP-DM) process (Chapman et al., 2000) after researching and comparing CRISP-DM with three other knowledge discovery processes: the Knowledge Discovery in Databases (KDD) process by Fayyad, Piatetsky-Shapiro, and Smyth (1996), the Sample Explore Modify Model Assess (SEMMA) method as referred to in (Azevedo & Santos, 2008) and the Three Phases Method (3PM) by Vleugel, Spruit, and van Daal (2010). CRISP-DM is a clearly described process and has been widely used for knowledge discovery processes ever since its inception, making it the 'de facto standard' in the field, for developing data mining and knowledge discovery projects (Giraud-Carrier & Povel, 2001; Harding, Shahbaz, Kusiak, & Srinivas, 2006; Onwubolu, 2009).

The CRISP-DM process consists of the following six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. These phases are also employed in this research.

2.1. Business understanding

Multiple unstructured in-depth interviews have been performed to create an understanding of the long-term sector. Unstructured in-depth interviews are appropriate when the richness of detail through clarification of questions and answers is to be ensured. Yin (2009) states that open interviews are the best way to discover explorative information.

Experts from different perspectives and various organizations have been interviewed in order to create a reliable and complete picture of the information needs in the entire long-term care sector. Our experts represent information needs of Nursing homes, Care homes and Home care (VVT), Mental care (GGZ) and Disability care (GZ). In total 22 experts were interviewed in 18 sessions. Table 1 shows that we interviewed eight experts from the board of directors level, seven experts from the management level and seven experts from stakeholder positions. The concept of Valuation in the left-most column will be explained in Section 3.

Information needs are the result of the interviews, which are consequently translated to data mining goals during this phase. The use of information needs is slightly different than the business goals as prescribed in the CRISP-DM method. In our view information needs are more elaborate than business goals. The data mining goals are input for the next phase: data understanding.

2.2. Data understanding

To allow proper modelling, it is important to understand the gathered data. We performed a single case study, which means that the data that has been gathered by one long-term care institution are used to do the modelling. The care institution has currently around 850 clients, both intramural and extramural, and has been working with the ResidentWeb ECD system since 2008. Table 2 shows the codes, colours and number of beds per location of the long-term care institution in our case study.

In order to use the data, all personal information such as names, BSN numbers, and addresses had to be deleted first. This research was commissioned by the software development company behind ResidentWeb, and the care institution under investigation is closely involved in the on-going development of the ECR software.

Table 1
Overview of interviewed experts in the long-term care sector.

Type of interviewee	Number of experts	Number of sessions	Valuation
Board of directors/Director	8	8	10
Management	7	5	6
Stakeholders	7	5	3
Total	22	18	–

Table 2
The five locations of the long-term care institution in our case study.

Location ID	Visualization colour	Number of beds
Loc1	Blue	142
Loc2	Red	52
Loc3	Yellow	150
Loc4	Black	81
Loc5	Green	34

Using the data has been approved by the board of directors. This care institution has a global agreement with all clients that the collected data may be used for analysis. A confidentiality statement had to be signed before using the data for this research. We are also bound to display the data anonymously. The database that is used for this research contains information from April 2008 until April 17, 2012.

During this step the data are described and explored, and also the data quality is described. Due to security and confidentiality issues the database model will not be described in detail.

2.3. Data preparation

In order to use the gathered data of the care organization, the data must be prepared first. Missing data or wrong data must be revealed and solved in order to do proper modelling and evaluation. For some data mining goals it is necessary to construct, integrate and format the data to enable proper modelling.

2.4. Modelling

Different models will be used during this phase to discover patterns and trends in the data, based on the data mining goals created during the business understanding phase. The models created in this step will be evaluated during the next step.

2.5. Evaluation

In this research phase the models have been evaluated. Experts from the long-term care institution will be used to evaluate the created models. Also the extent to which the models could be used to increase the quality of care, and the effectiveness of the gathered data is evaluated.

2.6. Deployment

The final CRISP-DM stage consists of employing the results of the study. The data mining goals are based on the information needs of the interviewed care institutions and stakeholders, data collected by one care institution are used to discover the availability of data, modelling and evaluation to check whether the models can meaningfully support the management of Dutch long-term care institutions. The conclusions of our current study consist of a clear description of the discovered added value of knowledge discovery processes as well as the shortcomings in the gathered data. Solutions are presented in order to facilitate more effective and efficient use of gathered data.

3. Calculation: exploratory data modelling

The interviews resulted in a large list of information needs, both aimed at the quality of care and financial state of a care institution. These information needs are translated into data mining goals which can be modelled based on the data collected with ECR software. This resulted in 25 data mining goals of which five are elaborated upon in more detail in this paper.

Customer experience is the most mentioned information need of both experts and stakeholders. This information is used to increase the quality of care and improve the quality of life. However, customer experience is mostly measured by an external company of which we did not have access to its data. Staffing with respect to the Care Intensity Package (ZZP) mix is the most important indicator for directors in order to control the expenditure and revenue. Staffing information gives insight in the expenditures, whereas the ZZP-mix provides important information for the directors to control the revenue. Forecasting the future ZZP-mix is also mentioned by many experts. Any consequences of changes in laws and regulations could then be directly discovered by care institutions. Information regarding incidents could be used to increase the quality of care. Especially the causes of incidents could lead to improvements that would increase the quality of care.

The top 10 information needs that have emerged from the interviews are listed in Table 3, wherein **Q** represents the Quality of care information needs and **F** denotes the Financial state information needs. The score for each information need was calculated based on their importance. The following formula was used to create a fair sorting:

$$\text{Score} = \sum_{\text{Expert level}} \frac{\text{Times mentioned}}{\text{Number of interviews}} \times \text{Valuation} \tag{1}$$

Equation 1: Information needs scoring formula.

We value an information need to be more important when mentioned by experts on the director level then when mentioned by experts on stakeholder positions. For example, the score for the number one information need *Customer experience* is calculated using Table 1 and Table 3 as follows:

$$\left(\frac{8}{8}\times 10\right)+\left(\frac{4}{5}\times 6\right)+\left(\frac{8}{8}\times 10\right)+\left(\frac{4}{5}\times 6\right)+\left(\frac{3}{5}\times 3\right)=16.6 \tag{2}$$

Equation 2: Example application of Equation 1.

The complete list of 25 data mining goals can be found in the Appendix. The next sections will investigate the extent to which we can answer the main information needs with our dataset.

Table 3
The Top 10 information needs at the long-term care institution.

#	Type	Information need	Board	Management	Stakeholders	Score
1	Qual.	Customer experience	8	4	3	16.6
2	Fin.	Staffing with respect to ZZP-mix	7	4	2	14.8
3	Fin.	ZZP-mix per business unit	7	4	0	13.6
4	Fin.	ZZP-mix prognosis	7	4	0	13.6
5	Fin.	Staffing with respect to operations	6	4	2	13.5
6	Qual.	Number of incidents occurred	6	4	2	13.5
7	Qual.	Types of incidents occurred	6	4	2	13.5
8	Qual.	Causes of the occurred incidents	6	4	2	13.5
9	Fin.	Operations per ZZP	7	3	1	13.0
10	Fin.	Production information (planned, realized, declared)	7	3	1	13.0

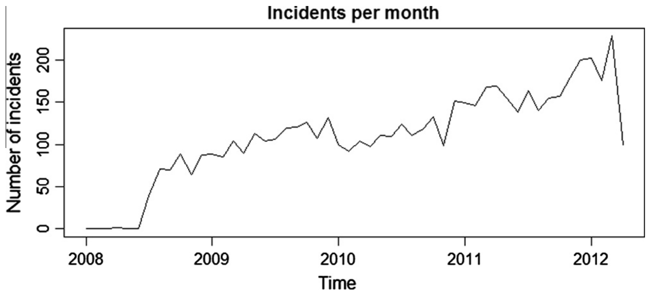

Fig. 1. Number of incidents per month at the long-term care institution.

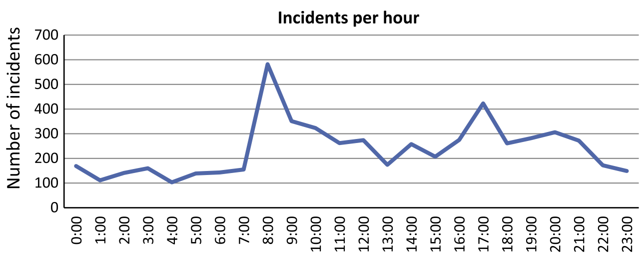

Fig. 2. Number of incidents per hour at the long-term care institution.

3.1. Identify patterns in incident information

Every incident should be registered in the ECR software. The dataset contains 6126 incidents, which includes attributes such as client, department, date and time, type of incident, cause, location, physical damage and mental damage. This collection of data is very valuable, and could be used for various analyses. First of all, all incidents are selected for which a client, department, date and time, type of incident and location are registered, which results in a collection of 5692 incidents. Fig. 1 visualizes the number of incidents per month over the years. The first incident was registered in April 2008, the first month with more than 30 incidents was July 2008. Therefore, this month will be taken as the start date for further analysis. The dataset contains information until April 17, 2012, which is the cause of the low number of incidents in April 2012. Note that although Fig. 1 shows an upward trend, this does not necessarily indicate that increasingly more incidents have happened. We assume that a better registration of incidents is most likely the cause of this trend.

Zooming in on the incidents data, we turn to the number of incidents with respect to the time of the incident. This analysis will create insight in the critical moments during the day at which most incidents occur. For this analysis, all incidents were grouped per hour using a SQL query which counts the number of incidents between, for example, 00:00 and 00:59.

Fig. 2 visualizes the number of incidents at a certain time of day. It turns out that most incidents occur during the day, between 08:00 and 09:00. The peaks between 08:00 and 09:00 and between 17:00 and 18:00 are most likely caused by the transfers of the clients (e.g. getting out of bed and going towards diner).

Also, the location of the incidents is being registered, which could be used to detect geographical problem areas at the care institution. For example, if most incidents take place in the corridor, it could trigger management to research this fact and to increase safety in the corridor. All 5,692 incidents are selected and formatted based on the incident location (in the first column) and the care institution location (in the first row) in Table 4 below.

In Table 4 the Top 3 locations have been marked where incidents occurred, which emphasizes the most common locations of incidents. For all care institution locations it becomes clear that most incidents take place in the living room. The other locations where incidents commonly occur are the bedroom, kitchen and bathroom. For these (problem) areas the percentages are described per location, which makes it possible to compare the locations with each other. Next, we extracted the consequences of these incidents by analysing the physical and mental damages to the clients after each incident, which is elaborated upon in the following tables.

Mental damage after an incident turns out to occur most often after an fall incident, as is shown in Table 5. The three shaded columns mark the incidents without observed mental damage. This means that there is either no observed mental damage, the mental damage is not yet noticeable, or the mental damage is unknown. The latter means that there was no value registered related to the mental damage after an incident, or that this option to register the mental damage was not implemented. Note that for no less than ((3548 + 1010 + 520)/5692=) 89.2% of the incidents no mental damage was registered. Falling incidents cause the highest number of mental damage, but for only ((199 + 202 + 41 + 87)/4292 =) 12.3% of all falling incidents there was mental damage reported. Furthermore, for only ((217 + 233 + 56 + 108)/5692=) 10.8% of all incidents mental damage was reported. Because the mental damage categories are rather loosely defined and therefore multi-interpretable, it is still unclear what the impact of this result is. Anxiety for example could mean that the client does want to leave the bed anymore, but it could also mean that the client is somewhat afraid to walk alone.

The shaded fields in Table 6 visualize the incidents which did not cause any physical damage. The same rules apply as for the

Table 4
Number of incidents per incident location and care institution location.

Incident location	Loc1	Loc2	Loc3	Loc4	Loc5	Total	%
Activities room	11	0	1	10	0	22	0.4
Other department	8	6	1	1	0	16	0.3
Bathroom	126 (6%)	77 (6%)	152 (9%)	43 (10%)	17 (14%)	415	7.3
Outside the building	24	12	52	11	1	100	1.8
Restaurant / Dining room	23	14	15	1	2	56	1.0
Somewhere else in the building	36	24	19	22	2	103	1.8
Corridor	101	72	92	9	3	277	4.9
Shared living room	3	5	1	0	0	9	0.2
Kitchen	225 (10%)	154 (12%)	110 (6%)	46 (11%)	10 (8%)	545	9.6
Bedroom	524 (24%)	278 (22%)	575 (33%)	38 (9%)	31 (26%)	1446	25.4
Toilet	62	45	54	5	8	174	3.1
Stairs	2	1	0	0	0	3	0.0
Living room	895 (41%)	500 (40%)	587 (34%)	231 (55%)	40 (33%)	2253	39.6
Other	120	63	78	5	7	273	4.8
Totals	2160	1251	1738	422	121	5692	100

Table 5
Mental damage per incident type at the long-term care institution.

Mental damage	Anxiety	None	Not yet noticeable	Unknown	Disquiet	Drowsiness	Other	Total
Aggression/Harassment	5	49	7	11	10	0	3	85
Burn/Scorch	0	3	1	0	0	0	0	4
Behaviour	0	0	0	0	0	0	1	1
Medication	1	744	101	149	7	6	5	1013
Unsafe situation	2	19	5	0	2	0	0	28
Prick incident	0	0	0	1	0	0	0	1
Bumps/Pinch/Clash	1	18	3	3	1	0	1	27
Fall	199	2575	855	333	202	41	87	4292
Missing resident	0	32	7	1	4	3	3	50
Other	9	104	30	20	7	6	8	184
Unknown	0	4	1	2	0	0	0	7
Totals	217	3548	1010	520	233	56	108	5692

Table 6
Physical damage per incident type at the long-term care institution.

Physical damage	Bruises	Fracture	Burn	None	Not yet noticeable	Unknown	Pain	Cut	Muscle complaint	Intoxication	Sprain	Other	Total
Aggression/Harassment	3	1	0	56	11	7	3	1	0	0		3	85
Burn/Scorch	0	0	2	1	1	0	0	0	0	0			4
Behaviour	0	0	0	1	0	0	0	0	0	0			1
Medication	0	0	0	784	132	79	3	0	0	0		15	1013
Unsafe situation	0	0	0	22	5	0	1	0	0	0			28
Prick incident	0	0	0	0	0	1	0	0	0	0			1
Bumps/Pinch/Clash	5	0	0	2	1	0	0	16	0	0		3	27
Fall	256	46	0	1791	838	28	559	562	5	2	13	192	4292
Missing resident	0	0	0	44	5	0	1	0	0	0			50
Other	6	1	1	91	45	4	9	11	0	0	1	15	184
Unknown	0	0	0	3	1	0	2	0	0	0	1		7
Totals	270	48	3	2795	1039	119	578	590	5	2	15	228	5692

Table 7
Risk assessment snapshot at the long-term care institution.

Risk description	No increased risk	Increased risk			No risk assessment
		No adequate follow-up	Adequate follow-up	Total	
Falling	117 (23.78%)	12	233	245 (49.80%)	130 (26.42%)
Incontinence	96 (19.51%)	24	241	265 (53.86%)	131 (26.63%)
Depression	101 (20.53%)	184	83	267 (54.27%)	124 (25.20%)
Medication	283 (57.52%)	10	68	78 (15.85%)	131 (26.63%)
Problem behaviour	–	–	–	–	492 (100%)
Weight extramural	–	–	–	–	492 (100%)
Weight intramural	241 (48.98%)	14	61	75 (15.24%)	176 (35.77%)

mental damage overview in Table 5. For both the mental as the physical damage applies that the biggest damage is caused by after fall incidents. For ((2795 + 1039 + 119)/5692=) 69.4% of the incidents there was no physical damage registered. Most physical damage is done after a falling incident; in only ((256 + 46 + 559 + 562 + 5 + 2 + 13 + 192)/4292=) 38.1% of all falling incidents physical damage was reported. The high number of falling incidents has a great impact on the overall percentages. For ((270 + 48 + 3 + 578 + 590 + 5 + 2 + 15 + 228) / 5,692 =) 30.6% of all incidents there was some form of physical damage reported, which is mainly caused by falling incidents.

Because the dataset contains a lot of valuable information about incidents, the number of potentially interesting analyses is also very broad. However, even though the causes of incidents are registered, our investigation has revealed that the high number of differently registered causes prevents an optimal analysis. The 5692 incidents in the dataset are related to no less than 1242 different causes. In order to perform a more useful analysis, the number of causes should be limited first. Another potentially interesting analysis could be to investigate which employees are more confronted with incidents, or the number of incidents for clients in a particular Care Intensity Package (ZZP) category.

3.2. Identify patterns in risk assessment information

Risk assessment is a relatively new functionality in the ECR software of the long-term care institution under investigation, which was introduced in 2011. More specifically, the first risk assessment for a client was performed on August 18, 2011. For the analysis in Table 7, the last risk assessment of the current clients is selected in order to create a reliable snapshot of the risk assessment within the care institution. A query which selects all current clients with a subscription and no date of death was executed to select 492 clients from the dataset. A subscription means that a client has a subscription for receiving some kind of product. Every risk assessment should be performed individually; for 121 clients no single risk

Table 8
Association rules regarding increased risk at the long-term care institution, sorted on lift scores.

Rule	Support (%)	Confidence (%)	Lift
Incontinence, medication → falling	11.76	90.63	1.824
Falling, medication → incontinence	11.76	95.08	1.769
Incontinence, weight intramural → falling	11.56	87.69	1.765
Falling, weight intramural → incontinence	11.56	93.44	1.738
Depression, incontinence → falling	31.64	82.11	1.652
Depression, falling → incontinence	31.64	88.64	1.649
Weight intramural → falling	12.37	81.33	1.637
Falling → incontinence	43.61	87.76	1.633
Incontinence → falling	43.61	81.13	1.633
Weight intramural → incontinence	13.18	86.67	1.612
Medication → incontinence	12.98	82.05	1.526

assessment was performed. These 121 clients are included in the numbers in Table 7 below.

Depression is the most commonly performed risk assessment under current clients. This is also the most interesting risk assessment. For ((101 + 267)/492=) 74.8% of the clients, a depression risk assessment was performed. More than (267/368=) 72.5% of these 368 clients have an increased risk of depression. However, for (184/267=) 68.9% of these increased risk clients, there was no adequate follow-up. Adequate follow-up means that the care institution includes, for example, measurable treatment goals into the care plan of a client in order to decrease the risk of incidents.

Another interesting discovery is the high percentage of clients that do not run an increased medication or weight risk. Medication risk includes the risk of taking the wrong medication, no medication, cognitive impairment or impaired hand function. This means that the client is assessed on various aspects in order to determine if there is an increased risk. Because a large number of medication incidents occurred, it is an interesting finding that only 78 clients run an increased risk. Weight risk includes weight loss, loss of appetite and the need for assistance during eating. Only 75 clients run an increased risk for underweight. Data regarding care related measures such as the client's weight can help explain how many clients are in fact high-risk clients.

In order to find valuable patterns or relationships in the risk assessment data, we have applied association rule mining (Chen, Han, & Yu, 1996; Freitas, 2003). For learning association rules we used the Apriori algorithm (Agrawal & Srikant, 1994), which is well-known, widely used and also available in the software package R Studio (R Studio, 2012). For all analyses a minimal support level of 0.1 and a minimal confidence level of 0.8 were selected. These parameters indicate that at least 80% (confidence) of the clients that for example have an increased falling risk also have an increased medication risk, and at least 10% (support) of the clients have both.

The relationships between the different risks are investigated based on the last risk assessment of the current clients as shown in Table 7. This means that the follow-up of an increased risk outcome was not taken into account for the first analysis and the discovery of association rules.

Table 8 provides insight in the risks to which a client is exposed. In the rightmost column the so-called lift of each association rule is shown. A rule's lift is computed with the following formula:

$$lift(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X) * \text{support}(Y)} \quad (3)$$

Equation 3: The well-known formula to calculate the Lift of an association rule.

In short, a larger lift indicates a stronger association. Both support, confidence and lift are interesting measures to help uncover the most interesting association rules (Hahsler and Chelluboina, 2010).

The most interesting association rule based on the lift score, states that 90.63% of the clients with increased incontinence and medication risks also have an increased risk of falling, almost 12% of the clients run an increased risk on all three. This is the association with the highest lift, and can thus be considered the strongest association.

Note that increased risk of falling or incontinence are part of every association rule. A large number of clients run an increased risk on these two aspects. Also a lot of clients run an increased risk of depression, but depression is only part of two association rules. That indicates that depression is less related to the other risks.

3.3. Identify the relationship between risk assessments and incident information

Our previous analyses in Section 3 showed that falling incidents are the most common incidents with the most damage, both mental and physical. For this reason, this section will focus on analyzing falling incidents to help identify the relationships between incidents and risk assessments. We hypothesize that clients with an increased risk of falling also suffer more falling incidents in their daily lives. In the following analysis we will investigate whether this assumption is true or not.

The possibility to perform a risk assessment is, as already mentioned, a relatively new functionality in the ECR software at the long-term care institution, which has become available on August 18, 2011. In order to find a relationship between risk assessment and the incidents, data are selected from September 2011 onward. The gap of 13 days has been taken into account, because the risk assessment cannot be done for all clients in just one day. We created a SQL query to select all fall incidents since September 2011, together with the outcome of the most recently performed risk assessment before the incident. This selection results in 937 falling incidents and 361 medication incidents, together with the associated risk assessments.

Table 9 shows that the risk assessment does not always create a good insight in the risk group. For falling incidents the risk group is quite well mapped, but that is not the case for medication incidents. This could be due to the fact that most medication incidents happen because care personnel forgets to provide the medication (properly) to the client. The other types of incidents and risks could not be analysed, because there is no further overlap between the two. Problem behaviour risk assessment could be used to gain insight in the target group for aggression incidents, but this type of

Table 9
Number of incidents with respect to risk assessment outcomes.

Risk assessment outcome	Fall incidents		Medication incidents	
No increased risk	71 (7.58%)		124 (34.35%)	
Increased risk – no adequate follow-up	62 (6.62%)		12 (3.32%)	
Increased risk – adequate follow-up	493 (52.61%)	555 (59.23%)	54 (14.96%)	66 (18.28%)
No risk assessment	311 (33.19%)		171 (47.37%)	
Total	937 (100%)		361 (100%)	

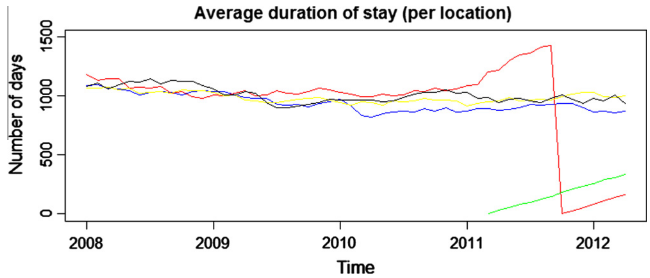

Fig. 3. The average duration of stay per location of the long-term care institution.

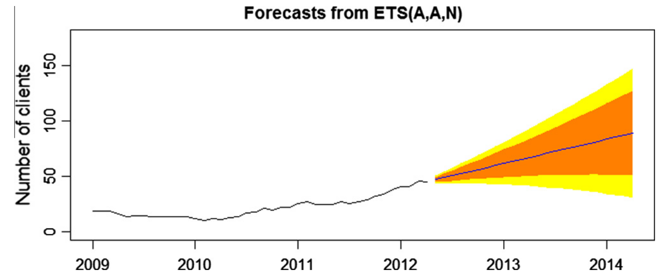

Fig. 5. Number of currently delivered ZZP6 clients with future forecast.

risk assessment is not being performed (yet) at the long-term care institution in our case study. The usefulness of risk assessments can, therefore, not be fully determined.

3.4. Identify patterns in the average duration of stay

A well-known trend in long-term care is that the duration of stay has been steadily decreasing over the years. Therefore, before we describe our analysis on this aspect based on the dataset, we need to explain that the duration of stay is measured periodically. For this analysis the duration of stay is measured every first day of the month, from January 2008 onwards. In other words, the duration of stay is measured from the start of the stay until the first of each month. Fig. 3 visualizes the average duration of stay from January 2008 onwards. The corresponding location colour codes are listed in Table 2.

In 2011 the red location has introduced a new room layout. This new room layout influences the average duration of stay, as can be seen in Fig. 3. One may argue that a move of a client from one room to another is not a reason in itself to reset the duration of stay. However, for this analysis we do reset the duration of stay. Because of that, the dip is prominently visible for the red location. The green location is a new location, which explains the constant increase in the average duration of stay. Another interesting insight is the slightly decreasing lines for the black, yellow and blue locations. Because this is the average duration of stay, it indicates that intramural clients stay shorter in these locations of the long-term care institution.

3.5. Identify and predict the Care Intensity Package (ZZP) mix

Fig. 4 shows the currently delivered Care Intensity Package (ZZP) mix per location. Regarding the Dutch concept of Care Intensity Package (ZZP) mix, in the Netherlands eight ZZP category levels have been defined from 2009 onwards, starting with ZZP-level 1 which represents "Extramural living with some guidance" up until ZZP-level 8 which designates "Intramural living under full surveillance and 24/7 care". The higher the designated ZZP-level of a client, the higher the operational costs to provide proper care to a client (Nederlandse Zorgautoriteit, 2012; Rijksoverheid, 2012).

In order to visualize the current ZZP-mix, the currently delivered ZZPs are selected from the dataset. Every ZZP has a start date and an end date, which makes it possible to select only the current ZZP for each client. The current ZZP mix is, therefore, based on the ZZPs that are actually delivered by the care institution, and not the indicated ZZP. This is important because it is quite possible that a client receives an ZZP5 indication from the care office while being designated at the ZZP4 level by the care institution for various reasons.

In order to predict the future ZZP-mix it is necessary to have historical data which represent the ZZP-mix over a longer period of time. Our dataset contains information about the delivered ZZPs from January 2009. Before January 2009, care institutions did not work with ZZPs yet. Changes in laws and regulations like the newly introduced ZZP system can make it, therefore, hard to provide a solid basis for data mining algorithms to accurately create predictions. Nevertheless, the data that are present in the dataset will be used to predict the future number of ZZP6 clients as an example, as shown in Fig. 5.

The predicted number of future ZZP6 clients was estimated using the forecast() method in the R analytics package (R Project, 2012). This function creates different forecasts and selects the best forecast (Norèn, Hopstadius, Bate, Star, & Edwards, 2010). The best forecast is chosen by selecting the forecast with the lowest error margin (de Gooijer & Hyndman, 2006). For this time series analysis the forecast method selected a model with a trend component and no seasonality. The number of ZZP6 will likely increase, based on the point forecast. The 95% confidence window (in yellow) for April 2014 is between 31 and 147 clients. The 80% confidence window (in orange) for April 2014 is between 51 and 127 clients. At this

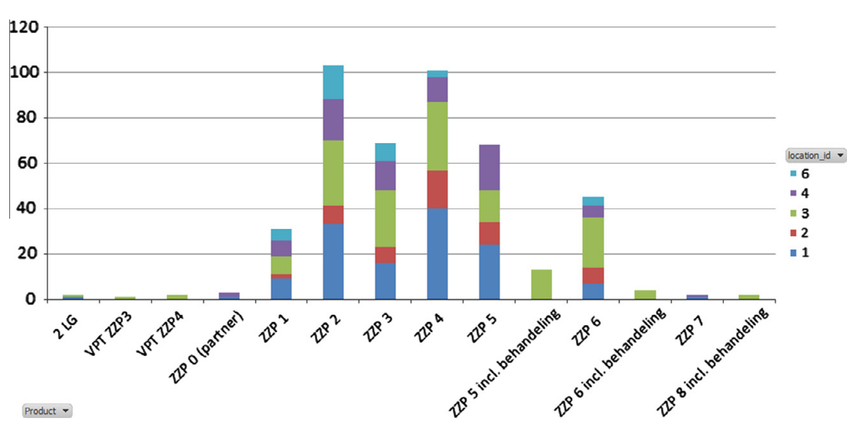

Fig. 4. Currently delivered Care Intensity Package (ZZP) mix at the long-term care institution.

moment the number of ZZP6 clients is, in fact, 45. In other words, we can predict with a likelihood of 80% that the number of ZZP6 clients will increase during the next two years.

4. Results: interpretation of findings

We structure the interpretation and discussion of our findings in two subsections. First, we evaluate the results from the long-term care perspective of the institution in our case study. Second, we discuss our findings from a more technical analytics perspective.

4.1. From a long-term care perspective

Together with the long-term care institution which provided the data for this research, we have evaluated the results of our analyses. Due to the highly explorative nature of this study, the results do not provide deeply profound insights. In order to gain a deeper understanding of the internal business, we could have customized our information needs more specifically for the long-term care institution in our single case study. That would enable to create a list of subjective data mining goals specifically tailored for that one care institution, while accepting the risk of overfitting the list of important information needs.

First of all, the long-term care institution was rather pleased with the results of this study as it provided them with a very interesting view on their internal business. Patterns in the number of incidents gave the care institution insight into the progress of incidents over the course of the week as well as during the day. The incident models gave also insight in the type of incidents per location, which helped pinpoint management of the care institution to implement changes. For example, the bathroom sizes vary per location, and this is now considered to be the main reason for the larger percentage of incidents that occur in the bathroom at some locations. Insight in the progress of incidents during the day gave insight in the peaks between 08:00 and 09:00 and between 17:00 and 18:00. That information is of added value for the care institution in order to increase the quality of care, because the care institution could respond to these facts. Also, the ZZP-mix information created a good insight in the progress of the mix over time. This information is now helping the care institution to create strategic plans and decisions.

Other performed analyses turned out to be less interesting. This is especially the case for the forecasts done in the modelling phase of this study. ZZP-mix, number of clients, occupancy rate could not be forecasted based on the limited historical data. There are too many dependent factors for the forecasts to be practically useful. Also, the analysis regarding care related time was not really useful, because it could not be based on the actual time.

The reasons above on why some analyses could not be performed properly, were confirmed by the care institution. Lack of standardization and data are the two main reasons for not being able to create a proper analysis. In order to be able to create those analyses in the future, ECR software vendors should focus more on usability engineering to help ease registration of these data. Coercive measures and restraints are an exception in this case. Software encoding of database complexity is a main reason which prevents accurate analyses of production information. Business rules, laws and regulations are programmed into the ECR software. The data are not modelled to contain such logic, which makes it hard to properly analyse the data.

For all data mining goals we can state that more elaborate analyses will become possible when data will be gathered in a more structured fashion, and in a quantitative manner. At this moment most data are gathered qualitatively, partly due to the personal process of delivering care. However, by adding a quantitative option, it will become possible to gain more insight from the data. For example, implementing a point scale for so-called General Daily Acts (ADLs) such as providing body care, accompanying toilet visits and helping out with clothing, would create a much more accurate insight into either the decline or progress of a particular client.

4.2. From a knowledge discovery process perspective

Some of the data mining goals could not be achieved due to either the complexity of the data, lack of data or lack of standardization. Production information that is currently being gathered by the care institution was either too complex or unstructured for immediate use in data mining algorithms. Business rules, laws and regulations should first be applied to the data, before proper analyses become feasible. This is also reflected in the data preparation phase of the CRISP-DM method which we applied in this research. Business rules, laws and regulations should be modularly accessible and, thus, unpluggable before performing analyses.

The data mining goals that could not be achieved due to the lack of standardization, make it clear that it is important to start creating standards. For example, a standard set of care goals could create more insight in the feasibility of the goals. Also, standards for ADL registration could create more insight in the decline or progress of a client. Incident information is to a certain level standardized, but further implementation of standards could gain even more insight into the patterns that exist. ECR software vendors should give care institutions the possibility to collect data in a highly standardized manner without actively enforcing a specific level of standardization, because not all care institutions are, in fact, in favour of highly standardized ECR systems.

Lack of data is another important reason for not being able to perform envisioned analyses. Most notably, all information regarding customer experience and complaints are not gathered by the care institution under investigation in this study itself, and are therefore unavailable in their ECR system database, even though these represent the most wanted information needs, as shown in Table 3. Thus, it is impossible to perform any analysis related to these aspects. We advise ECR software vendors to further research the need of care institutions to create such functionality in their software. Customer experience is now measured by external companies, with standardized questionnaires and answers, which results in quantitative data that could very easily be incorporated in data analyses.

Finally, much of the information that is gathered by care institutions, is gathered in a qualitative manner. Free text input is often used to facilitate this. The main reason for this seems to be the core process of care institutions, which is delivering care. This is considered to be a personal process with qualitative data as a result. No two clients are the same, which is also the case for the data that should be gathered for the clients. Coercive measures and restraints are rarely performed, so there is little data available on this aspect. This is also the case for complaints, as only one or two complaints per year are currently being registered, which can obviously not be a basis for any data analysis.

5. Conclusions and further discussion

In this section we answer and reflect on the research question of this study: "*How can knowledge discovery techniques support Dutch long-term care institutions to manage their internal organization?*".

Our research is based on the CRISP-DM model to structure the knowledge discovery process applied on an exploratory single case

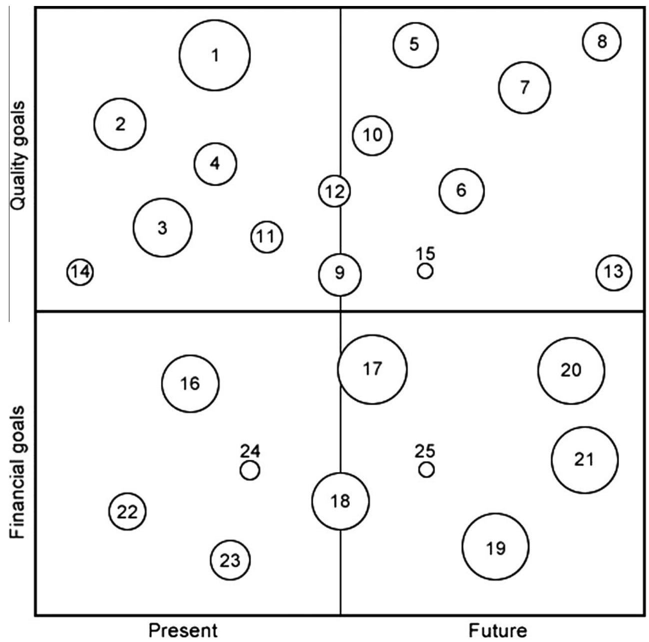

Fig. 6. Two-by-two matrix which positions all data mining goals with respect to quality of care and financial state.

study of a Dutch long-term care institution. This has provided several new insights into both the applicability of the CRISP-DM model, and the way the case study organization at stake can improve their internal management and governance. In this study, first the overarching information needs of long-term care institutions within all subsectors were collected through expert interviews in Nursing homes, Care homes and Home care (Dutch abbreviation: 'VVT'), Mental care (Dutch abbreviation: GGZ), and Disability care (Dutch abbreviation: GZ). Next, these information needs were translated into quantifiable data mining goals. As the data mining goals are based on the overall information needs of institutions throughout all long-term care subsectors, the information needs are defined on a relatively high (abstract, aggregate) level. We have chosen to pursue this wide scope, as knowledge discovery processes in the Dutch long-term care – to the best of our knowledge – have not been scientifically conducted until this study. Hence, this explorative research also contributes as a first example on how to improve efficiency through model-based knowledge discovery on existing data of care institutions.

A practical contribution and deliverable of this research is the positioning of all 25 data mining goals in long-term care (listed in the Appendix) presented in a comprehensive two-by-two matrix, as shown in Fig. 6. The size of each circle represents the relative importance of the represented data mining goal, based on the times it was mentioned in the interviews with the sector experts. In the above two quadrants, all goals related to the quality of care are positioned, whereas in the bottom two quadrants all goals related to the financial state of care institutions are depicted. Based on data analyses performed at the case study organization, the data mining goals are also ordered on a time scale. The data mining goals in the two left quadrants can already be satisfied with the current ECR software, whereas goals in the two right quadrants should be taken into account in future developments.

A large amount of data that is available in care institutions is of a qualitative type, i.e. text based and unstructured, due to the fact that delivering care is only partly protocolled and mostly a personalized and fuzzy process. Extending standardization of processes and registration can hinder health care personnel in their daily work of providing personalized and customized care to often complicated and fragile clients and patients (Zuidgeest, Delnoij, Luijkx, de Boer, & Westert, 2012). Besides the possibility to gain data in a qualitative way, we believe that ECR software vendors should take a role in supporting data recording in a more quantitative way. Numerous data collection and coding techniques exist that could be used to increase the measurement level and usability of the gathered data. E.g., point scales could be added for collecting ADL (daily life activity) information, where one point indicates that a client is very dependent on the help of care personnel and five points means that the client is very independent. Multiple ADL measurements per year could then create insight in the decline or progress of individual or sub-groups of clients, in relation to location, type of mental/medical care or treatment, et cetera. We primarily advocate to develop measurements based on data that are already gathered. Standardization of the data collection and processes to extend/enrich data can increase the effectiveness and usability of the data for further data analysis.

Data collection and provision by health care organizations are highly dependent on various laws and regulations. Long-term care is a publicly financed service in the Netherlands, that should be accessible and affordable to every citizen in The Netherlands (Schäfer et al., 2010). The Dutch government is responsible for the availability, quality and costs of long-term care. Due to political choices and policy considerations, changes in existing, and introductions of new laws and regulations, are frequent. This has direct consequences for the information requirements posed on care institutions and, consequently, the data they need to collect and provide.

Based on the evaluation of the data analysis results, it is clear that predictive models are not yet directly valuable for care institutions. Predictions are too complicated and strongly dependent on exogenous factors and contingencies. For example, the number of clients that a long-term organization will serve in 5–10 years, can radically change due to mergers and acquisitions, the restructuring of care in regions or subsectors, changing laws and regulations, changing population needs, and so on. Still, descriptive modelling of current and future care needs ('volumes') is very interesting for care institutions and – to certain extent – can be done by exploiting the large quantity of valuable information that are already present. Staffing and capacity planning can be improved by analysing the activities of clients' care plans, the number of incidents that occur during the day, and other indicators relevant to debunk structures and patterns in resources required. In the future, predictive modelling could become possible and of increasingly more value if data will be collected and adapted to enable quantitative processing, analysis and reporting.

Ultimately, evolving predictive analytic capabilities may even result in the application of more rigorous analysis & intelligence methodologies, such as the System Dynamics approach of Petruch, Tamm, and Stantchev (2012) to better identify real causalities and hidden dependencies between different datasets, or even a personalized feedback recommendation engine to stimulate learning and more active participation as envisioned by Bodea, Dascalu, and Lytras (2012).

Finally, not only care institutions, but also the relevant stakeholders should have a stronger focus and vision on their information needs. All stakeholders are aware that long-term care, especially the Care homes and Home care (VVT) subsector, is under pressure of decreasing resources and increasing demands, in particular the quality of care. Quality of long-term care however, cannot be measured accurately by only registering incidents and benchmarking care organizations on this indicator, as the validity and reliability of this indicator is too limited. E.g., one can argue whether an incident is a falling incident or not, and how to code an incident when (for example) a client slips out of a wheelchair. Given that registration and administration takes time at the cost of patient-related care, organizations, and their demanding stakeholders, should have a clear vision and support what events and activities should be registered in order to generate data that serve clear objectives.

We claim that large quantities of potentially valuable information being collected by care institutions can be used more efficiently by defining and explicating information needs. Current data are not collected to gain insight in the internal organization yet. They are merely collected to support the primary process of the care institutions. To turn these data into information and knowledge for strategic policy and decision making, care institutions should model their information needs and clearly define their policy purposes to acquire data.

All in all, this research has shown that by exploring knowledge discovery techniques based on a large dataset and guided by representative information needs, we can contribute to the domain of long-term care to better manage both quality of costs of care – as potentially a next step towards healthcare business intelligence in long-term care.

Acknowledgements

We would like to thank all 22 interviewees who participated in this study. They provided us with valuable insights and understanding of the Dutch long-term care sector. We also highly appreciate the care institution which provided their operational data, which enabled the data analysis part of this study. Finally, we would like to thank the CEO of the ECR software vendor for all her efforts in arranging meetings with various experts within the Dutch long-term care sector.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.chb.2013.07.038.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th international conference on very large databases* (vol. 1215, pp. 487–499). Santiago de Chile.

Azevedo, A., & Santos, M. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. In *Proceedings of the IADIS European conference data mining* (pp. 182–185). Amsterdam.

Bodea, C.-N., Dascalu, M.-I., & Lytras, M. (2012). A recommender engine for advanced personalized feedback in e-Learning environments. *International Journal of Engineering Education, 28*(6), 1326–1333.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000). CRISP-DM 1.0 – Step-by-step data mining guide. *SPSS*.

Chen, M., Han, J., & Yu, P. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866–883.

Cios, K., Pedrycz, W., & Swiniarski, R. (2007). *Data mining: A knowledge discovery approach*. Springer Verlag.

de Gooijer, J., & Hyndman, R. (2006). 25 Years of time series forecasting. *International Journal of Forecasting, 22*(3), 443–473.

Duin, C. v., & Garssen, J. (2010). *Bevolkingsprognose 2010–2060: Sterkere vergrijzing, langere levensduur*. Den Haag: Centraal Bureau voor de Statistiek.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37–54.

Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information and Management, 37*, 271–281.

Freitas, A. (2003). A survey of evolutionary algorithms for data mining and knowledge discovery. *Advances in evolutionary, computing* (pp. 819–845).

Giraud-Carrier, C., & Povel, O. (2001). Characterising data mining software. *Intelligent Data Analysis, 5*, 1–12.

Hahsler, M., & Chelluboina, S. (2010). *Visualizing association rules: Introduction to the R-extension package arulesViz*.

Harding, J., Shahbaz, M., Kusiak, A., & Srinivas, S. (2006). Data mining in manufacturing: A review. *Journal of Manufacturing Science and Engineering, 128*, 969–976.

Koh, H., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management, 19*(2), 64–72.

Kurgan, L. A., & Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review, 21*(1), 1–24.

Lucas, P. (2004). Bayesian analysis, pattern analysis and data mining in health care. *Current Opinion in Critical Care, 10*(5), 399–408.

Mettler, T., & Vimarlund, V. (2009). Understanding business intelligence in the context of healthcare. *Health Informatics Journal, 15*(3), 254–264.

Ministerie van Volksgezondheid, Welzijn en Sport. (2011). *Programmabrief langdurige zorg*. Den Haag.

Ministerie van Volksgezondheid, Welzijn en Sport. (2011). *Vaststelling van de begrotingsstaten van het Ministerie van Volksgezondheid, Welzijn en Sport (XVI) voor het jaar 2012*. 's-Gravenhage: Ministerie van Volksgezondheid, Welzijn en Sport.

Mot, E., Aouragh, A., Groot, M. d., & Mannaerts, H. (2010). *The Dutch system of long-term care*. Den Haag: CPB Netherlands Bureau for Economic Policy Analysis.

Nederlandse Zorgautoriteit (2012). *Beleidsregel CA-300-510: Prestatiebeschrijvingen en tarieven zorgzwaartepakketten*. NZa.

Norèn, G., Hopstadius, J., Bate, A., Star, K., & Edwards, I. (2010). Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery, 20*(3), 361–387.

Onwubolu, G. (2009). An Inductive Data Mining System Framework. *IWIM* (pp. 108–113).

Petruch, K., Tamm, G., & Stantchev, V. (2012). Deriving in-depth knowledge from IT-performance data simulations. *International Journal of Knowledge Society Research, 3*(2), 13–29.

Prather, J., Lobach, D., Goodwin, L., Hales, J., Hage, M., & Hammond, W. (1997). Medical data mining: Knowledge discovery in a clinical data warehouse. In *Proceedings of the AMIA annual fall symposium* (pp. 101–105). Nashville: American Medical Informatics Association.

R Project (2012, 05 25). *The R Project for Statistical Computing*. Retrieved from R-Project <http://www.r-project.org/>.

R Studio (2012, 05 25). *RStudio*. Retrieved from RStudio: <http://rstudio.org/>.

Rijksoverheid (2012, 04 15). *Zorgzwaartepakket: Beschrijving van de zorg*. Retrieved from Rijksoverheid <http://www.rijksoverheid.nl/onderwerpen/zorgzwaartepakketten/zorgzwaartepakket-beschrijving-van-de-zorg>.

Schäfer, W., Kroneman, M., Boerma, W., van den Berg, M., Westert, G., Devillé, W., et al. (2010). The Netherlands: Health system review. *Health Systems in Transition, 12*(1), 1–229.

Vleugel, A., Spruit, M., & van Daal, A. (2010). Historical data analysis through data mining from an outsourcing perspective: The three-phases method. *International Journal of Business Intelligence Research, 1*(3), 42–65.

Yin, R. (2009). *Case study research: Design and methods*. Sage Publications, INC.

Zuidgeest, M., Delnoij, D. M. J., Luijkx, K. G., de Boer, D., & Westert, G. P. (2012). Patients' experiences of the quality of long-term care among the elderly: Comparing scores over time. *BMC Health Services Research, 12*, 26.