Battling the labour shortages in the mentally handicapped-care: a data-driven approach

I.W. Smeerdijk

trajectum

UNIVERSITY OF TWENTE.

Colophon

Thesis BSc Industrial Engineering & Management, University of Twente.

“Battling the labour shortages in the mentally handicapped-care: a data-driven approach”

Number of pages
53

Date of publication
27-08-2024

Author
Ilan Wilton Smeerdijk
BSc Industrial Engineering & Management
Tel. +31 6 42000393
i.w.smeerdijk@student.utwente.nl

University of Twente
Prof. dr. ir. Erwin W. Hans
Dr. ir. Gréanne Leeftink

Drienerlolaan 5
7522 NB, Enschede
Tel. 053 489 9111
e.w.hans@utwente.nl

Stichting Trajectum
Rody Fokke
Bart Slager

Dr. Stolteweg 17
8025 AV, Zwolle
Tel. 088 929 50 90
info@trajectum.inf

Preface

Dear reader,

Before you lies the work that finalises my bachelor's in Industrial Engineering and Management at the University of Twente. Performing this research at – and under the supervision of – the company Trajectum has offered great new learning experiences. Interacting with the people working within the facilities has been inspiring to say the least, and performing my research in the sector of healthcare has taught me more about the direct impact data-science can have on people than I could have predicted before starting this research.

Before introducing you to the work that represents ten weeks of my blood, sweat, and tears, I would like to take this opportunity to thank all the people that made this project possible. First, I want to express my gratitude to my supervisors: Erwin Hans, Gréanne Leeftink, Rody Fokke, and Bart Slager helped me steer the research in the right direction when needed, and held enough trust in me to give me the liberty to independently shape the rest of the research. On the side of Trajectum, I also thank Erik Sierhuis for the interesting tour of the clinic in Rekken (and the excellent lunch).

Additionally, I thank dr. Wonjun Song (송원준), dr. Seoncheol Park (박선철), and dr. Seunghwa Rho (노승화) of Hanyang University, Seoul. These professors have together provided me with the knowledge needed to conduct this research, regarding the use of R. This knowledge in coding has been vital in both processing the data, as well as ideating approaches to the posed problems.

Finally, I thank my friends and family for supporting me through the process of writing this research, of course including my dear friend and housemate Marius. I genuinely hope that reading this research is as enjoyable and interesting for you as it was for me to write.

Yours sincerely,

Ilan Smeerdijk

Managementsamenvatting

Trajectum is een bedrijf gespecialiseerd in het behandelen van mensen met een (licht) verstandelijke handicap, zowel in de vorm van begeleid wonen als in gesloten opvangen voor cliënten met gedragsproblemen. Het bedrijf is onderdeel van het initiatief G-AAN, waarin meerdere zorginstellingen uit Noord- en Oost-Nederland zich inzetten voor de digitalisering van de verstandelijk gehandicaptenzorg. Zo doet het initiatief onderzoek naar de toepassingen van nieuwe technieken – zoals data-gedreven werken – in de gehandicaptenzorg, zodat de bedrijven in deze sector goede zorg kunnen blijven leveren in de toekomst, wanneer het tekort aan arbeidskrachten mogelijk uitbreidt.

Dit onderzoek heeft twee doelen: (1) Een voorbeeld geven aan de bedrijven in het initiatief G-AAN van de mogelijkheden en limieten van data-gedreven werken binnen de verstandelijk gehandicaptenzorg, en (2) het verminderen van incidenten binnen de klinieken van Trajectum door middel van data-gedreven werken. Zo is de onderzoeksvraag die gehanteerd wordt: *"Hoe kunnen we een data-gedreven manier van werken implementeren in organisaties binnen de (verstandelijk-) gehandicapten zorg om toekomstige incidenten te voorkomen"*. Hoofdstuk 1 behandelt de manier waarop deze doelen worden benaderd.

Ondanks een beperkt aanbod aan academische werken over de verstandelijk gehandicaptenzorg, zijn er duidelijke applicaties en limieten voor deze bedrijven, gebaseerd op literatuur uit andere hoeken van de zorg en gesprekken met zorgverleners binnen de klinieken van Trajectum. Hoofdstuk 2 behandelt deze gesprekken. Door de unieke relatie tussen cliënt en medewerker binnen deze sector ontstaan ethische vraagstukken over de rol die data, IT'ers, en mogelijk AI kunnen, mogen, en zouden moeten hebben binnen de bedrijven in deze sector. Hoofdstuk 3 behandelt de resulterende conclusies.

Uit dit literatuuronderzoek en de gesprekken met medewerkers van Trajectum concluderen we dat de meest gepaste toepassing van data-gedreven werken voor het verminderen van incidenten binnen de context van Trajectum een tool is dat het risico van een incident inschat, op basis van de factoren die volgens de voorgaande onderzoeken mogelijk een relatie hebben met het aantal incidenten. Hoofdstuk 4 behandelt deze factoren en de redenatie achter de mogelijke relatie met de incidenten en geeft vorm aan een model van deze relaties. Hoofdstuk 5 test de aanwezigheid van deze relaties door middel van een aantal lineaire regressies tegen een significantie van 99.9%. De kracht van deze legitieme relaties wordt berekend en meegenomen in een model dat op basis van de waarde van factoren het verwachte aantal incidenten binnen elke dienst berekent.

Een van de conclusies die ontstaat uit deze statistische analyse is de prominente (negatieve) relatie tussen de gemiddelde ervaring van de aanwezige medewerkers en het aantal agressie-incidenten. Ook blijkt uit deze regressies een uitgesproken relatie tussen een geagiteerde sfeer (veroorzaakt door een recent agressie-incident) binnen de kliniek en het aantal toekomstige agressie-incidenten. Beide van deze factoren bleken goede voorspellers van toekomstige incidenten binnen de klinieken van Trajectum en werden geïmplementeerd in de tool.

De manier waarop Trajectum deze tool kan implementeren binnen het bedrijf, om met data beslissingen te ondersteunen, en daarmee data-gedreven te gaan werken, wordt behandeld in hoofdstuk 6. Samen met hoofdstuk 7, waar het onderzoek wordt geconcludeerd en het eindadvies wordt gegeven aan Trajectum, levert dit slot van het onderzoek een ontwerp van hoe een bedrijf in de verstandelijk gehandicaptenzorg data-science technieken kan toepassen om bestuurlijke keuzes te ondersteunen.

Management summary

Trajectum is an organisation that concerns itself with treating people with a (light) mental handicap, through guided living, as well as closed detention centres (Dutch: tbs), meant for patients showing unpredictable behaviour. The organisation is a part of the initiative G-AAN, which strives to encourage the digitalisation of the handicapped-care. The initiative looks for application of new technologies – such as data-driven working techniques – in the mentally handicapped-care, such that organisations can continue giving proper care in the future, when labour shortages are possibly even greater.

The objective of this research is twofold: (1) To give an example to G-AAN of the possible applications and limitations of data-driven working in the mentally handicapped-care, as well as (2) decrease the rate of incidents within the clinics of Trajectum through the implementation of data-driven working. The research question is: *"How can we implement a data-driven way of working in companies in the (mentally) handicapped-care to prevent future incidents?"* The manner in which these objectives and the research questions are approached, is discussed in Chapter 1.

Despite examples of data-driven research performed in the mentally handicapped-care being scarce, prominent applications as well as limitations of data-driven working for these organisations became apparent, through research performed in other areas of healthcare and the interviews conducted with employees of Trajectum, which are discussed in Chapter 2. As a result of the unique relation between the workers and patients within the handicapped-care, interesting ethical issues emerge regarding the role that data, computer scientists, and possibly AI should play in this sector. The conclusions regarding these issues are discussed in Chapter 3.

From the literature research and the interviews we conclude that the most appropriate application of data-driven working, in order to decrease the rate of incidents, is a tool that is able to assess the risk of an aggression incident occurring, based on the variables that could possibly have a relation with the number of incidents inside of the clinics of Trajectum, according to the previous research.

These variables, and the reasonings behind their assumed relation with the frequency of incidents, are discussed in Chapter 4, where a model of these variables and their respective relations is ideated. The presence of these relations is tested through a set of linear regressions against a significance of 99.9% in Chapter 5, where the strength of the confirmed relations is calculated and included in the tool, which calculates the expected number of incidents, based on the value of the variables particular to that shift.

One of the most interesting conclusions resulting from the statistical analysis is the (negative) relation between the average years of experience of the present workers and the number of incidents. Besides, a significant relation is found between a hostile atmosphere inside of the clinic, characterised by a recent aggression incident, and the rate of future aggression incidents. Both variables turned out to be good predictors of future incidents within the clinics of Trajectum and were included in the prediction tool.

The manner in which Trajectum can implement the tool into their business decisions, such that their future decisions-making is supported by data, working data-driven, is discussed in chapter 6. In combination with Chapter 7, where we conclude the research and give final recommendations to Trajectum, this final part of the research gives the blueprint for companies in the mentally handicapped-care to utilise data-science in supporting their decision making.

Table of Contents

1 Research plan

1.1 Problem context

This research is completed at – and under the supervision of – the company Trajectum, which specialises in providing care for people with a (light) mental disability. The company is a part of the initiative G-AAN, which strives to digitalise the mentally handicapped-care in the east of the Netherlands. This research is based around one of the objectives of G-AAN: implementing data-driven working (DDW) into the mentally handicapped-care in the Netherlands. Trajectum currently makes little to no use of their data, due to a "gap in their abilities." This research aims to fill this gap.

The problem posed by Trajectum is one that many (mental) healthcare institutes in the Netherlands are currently facing. There is a striking scarcity of workers in about every area of the Dutch healthcare system. This means that clinics are currently dealing with more patients than their employees can manage, with some patients not being able to get into the clinics entirely, being stuck on a long waiting list. As the situation is only expected to become more dire, health institutes are looking to implement new technologies that ensure that they make optimal use of their (labour) resources. The objective of this research is to take weight off of the shoulders of their employees, and to ensure that they are still able to provide proper care in the future, when labour scarcities are possibly greater.

We will make a start on the implementation of DDW into Trajectum to demonstrate how companies in the mentally handicapped-care, as well as other companies in the field of healthcare, can use data as a valuable resource. This research helps battle the (future) shortages in the healthcare system of the Netherlands by showing what value data-science can add to this field.

1.2 Problem description

1.2.1 Identification of action problem

Trajectum wants to move from a descriptive to a predictive use of data. The underlying motivation for this change is rooted in a desire to reduce the number of incidents inside of clinics, which poses as the action problem that needs to be solved through this research. These incidents range from e.g. a patient receiving the wrong medication, to violence and sexual harassment towards members of staff and other patients. However, there are limitations to this action problem, regarding the evaluation of this research.

Considering the scope and timespan of this research, the most appropriate measure of norm and reality (and the gap between them) is the *utilisation of data* within the company. The *reality* for the company is a *descriptive* use of data, meaning that currently the data can only tell the company past information about incidents happening, such as the number of incidents in a specific timespan or clinic. The company is given no insight into what might cause fluctuations in the number of incidents. The *norm*, which represents the preferred situation of Trajectum, is a *predictive* use of data. Meaning that useful predictions for the future can be made based on the past data, regarding the risk of an incident occurring. Based on this information the management can make better informed decisions, which in turn could reduce the number of incidents.

The utilisation of data is only used as a measurement for the company to evaluate the effectiveness of the research, after it is finalised. The action problem, meaning the problem that we aim to solve through this research is reducing the rate of incidents inside of the clinics of Trajectum.

1.2.2 Problem cluster

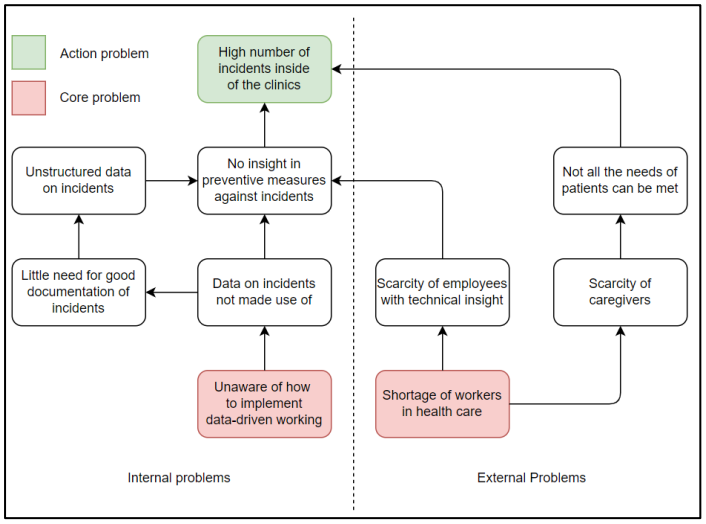

Figure 1.1; Problem cluster

Figure 1.1 poses a problem cluster which shows all relevant problems that Trajectum and other health institutes are facing and the relations between them. Problems that are external to Trajectum, meaning they are shared by other Dutch (mental-) health institutions, are depicted on the right. These are problems that this research does not or barely influence, but that do have an impact on the action problem. The problems that are internal to the company are depicted on the left side of the figure. These are the problems this research aims to address .

1.2.3 Core problems

From this problem cluster arise two candidate core problems: one external and the other internal to the company. The scarcity of workers in healthcare is one of the core problems which causes various other problems for institutions nationwide. For Trajectum specifically, the lack of knowledge regarding the implementation of DDW prevents them from making optimal use of their resources. As the former problem lies mostly outside of our control (Heerkens & Van Winden, 2021), this research will focus on the latter core problem: The lack of insight on the possibilities of data-driven working inside Trajectum.

This core problem unintentionally creates more problems, as it means that the data regarding the incidents are not used and therefore there is little need for structural documentation of these incidents. Consequently, the data on the incidents might be unreliable and less practical (e.g. because of missing data points). Together with the scarcity of technically educated personnel in the field, this leads to the current situation where the company only uses data descriptively, stating how many incidents there were in each clinic each month, unknowing of the underlying causes.

1.3 Research objective

Knowing this, we construct the research objective, which will function as the foundation of the research and the base for the intended deliverables and research questions. This research objective, as well as the respective research questions, are constructed using the ABC-model (Wisse & Roeland, 2022).

The main research objective:

"How can we implement a data-driven way of working in companies in the (mentally) handicapped-care to prevent future incidents?"

This research objective implies a scope that is not limited to the company Trajectum, but rather one that includes other organisations in the mentally handicapped-care, which was considered while ideating the intended deliverables. This ensures that the findings of this research can be of value to the other organisations in the initiative G-AAN.

1.4 Deliverables

To answer this research objective two separate deliverables are constructed. The first deliverable is an assessment of how organisations such as Trajectum can benefit from new technologies, e.g. the implementation of data-driven working, based on a literature research. This research is also meant to explore the limitations of DDW in organisations in the field of healthcare.

The other deliverable is a tool constructed for the IT department of Trajectum, which should be able to give an estimation of the relative risk of an incident happening, based on staff utilisation. This estimation can be made using previously fed data of the incidents and the workers present in the clinic at the time of an incident, considering factors such as the security-level of the clinic. The tool warns the company of an increased chance of an incident occurring, for example because there are too few members of staff present in one of the clinics. Trajectum could make more informed decisions accordingly, e.g. by allocating more (experienced) staff to a clinic with an increased risk of incidents.

The *scope* of the research for the last deliverable is limited to Trajectum, as the tool is specialised for the company. The *scope* of the research for the first deliverable is broader, as the literature research is more generally applicable to other companies in the field of healthcare, who can also use the tool as an example of the possible practicalities of data-driven techniques within their own organisation. Meaning that this research could be beneficial to the field of healthcare as a whole.

Together, these deliverables make up a blueprint of how companies in the mentally handicapped-care can benefit from data-driven working. Besides assisting Trajectum in their making decisions, this research also serves as an introductory example of the practicality of DDW for other organisations within the initiative G-AAN.

1.5 Research questions

The constructed research questions aim to together realise the research objective. Besides that, these questions make up the structure of this research, as one chapter will be dedicated to each of these research questions, which will be answered in their respective order. The first two of these questions together shape the first deliverable, and similarly, the last two question the second deliverable. This section states the research questions together with their relevance. A more extensive overview of the respective types of research, research subjects, research strategies, and data gathering and analysis techniques for each research question can be found in Figure 9.1 in the appendix.

RQ1. What type of incidents have to be prevented at Trajectum? (Chapter 2)

To get a proper view of the problem context of Trajectum, we need to get closer to the processes of the company. In the research conducted to answer this question, we make observations inside of one of the clinics of Trajectum and conduct interviews with the members of staff of various departments. This will give us insight on the types of incidents faced by Trajectum, and their needs from this research.

RQ2. In what ways can data-driven working help the mentally handicapped-care? (Chapter 3)

A systematic literature review aims to find the most prominent applications of DDW-techniques for organisations in the mentally handicapped-care. From this qualitative study of the literature, we find applications that have been used in academic literature inside, as well as outside, the field of healthcare and evaluate how practical these applications are when translated to the problem context of Trajectum.

RQ3. In what ways can data-driven working prevent incidents? (Chapter 4)
This research makes up the foundation of the tool that serves as the second deliverable. We aim to answer this research question through a qualitative study of the literature, in combination with a process of ideation that is meant to explore the possible functions and methods that the tool could incorporate. This knowledge is used in the research for the last research question where a model for the tool is constructed.

RQ4. How can we predict incidents at Trajectum? (Chapter 5)
To answer this last research question, we combine the previously gained knowledge regarding the problems of Trajectum and the ideation based on the literature review to construct a model that will form the basis of the tool. This final deliverable will be a valuable addition to the company, as well as a proper example to all companies in the initiative G-AAN of how DDW can help them make optimal use of resources and remove work from the shoulders of workers. Together with the first deliverable, this concludes the research and demonstrates how companies such as Trajectum can implement and benefit from DDW-techniques to prevent future incidents inside of their clinics, answering the main research objective.

2 Analysis of incidents

This chapter answers the first research question: *"What type of incidents have to be prevented at* Trajectum*?"* This is done by getting closer to the research subjects, which we do by collecting data through observations and interviews with experts, with the data analysis being a summary of the findings. These methods give a clearer view of which incidents can occur inside of the clinics of Trajectum, as well as which type of incidents are most important to reduce.

2.1 Methods

The data collection needed to answer this research question is performed in one of the clinics of Trajectum in the small village of Rekken, in the Dutch Achterhoek. By observing the normal course of events for the staff and patients, we get an overview of the internal structure and workings of the company. Besides that, we conduct short interviews with several members of staff. The questions asked in these interviews can be found in Figure 9.2 in the appendix. These are aimed at better understanding under what circumstances incidents can occur, what types of incidents are most common, and what types of incidents are most important to the care-givers of Trajectum to prevent. To answer the first research question, we determine the type of incidents that need to be prevented based on the failure mode and effect analysis (FMEA), which methodologically prioritises certain types of incidents based on three factors: Severity, occurrence, and detection (Yu et al., 2018).

2.2 Summary

2.2.1 Clinics

Trajectum does not only provide help for people with a (light) mental handicap in the form of assisted living, but also has incarceration facilities for this same demographic. In the Netherlands, Trajectum is the only organisation that people with a mental handicap can be sent to after being prosecuted. To distinguish between these different types of clinics, the departments of Trajectum have a security level ranging from 1 to 3 in increasing order of security. There are also departments without a security level, meaning that there are no security measures. In the continuation of this research, these departments will have an assigned security level of 0. The patients of Trajectum are expected to move through different departments, as they are treated, to eventually end up in a department without security measures.

All departments in the clinic in Rekken have a security level of 2, meaning that most patients cannot leave the premises. The caregivers inside of the clinics work in two day-shifts: the morning-shift from 07:00 to 15:00, and the evening-shift from 15:00 to 23:00. Besides that, there is a night-shift from 23:00 to 07:00, where patients are supposed to sleep. Though incidents can occur during the night-shift, most occur during the day-shifts. Consequently, there are less members of staff present during the night-shift, compared to the day-shifts.

2.2.2 Incidents

From the interviews with the caregivers, some of the limitations of this research became apparent. As e.g. cursing or slamming a door is already seen as an incident, reporting all of these occurrences would be too time-consuming for the staff, considering the frequency of occurrence inside of some of the departments. Consequently, not all incidents are reported by the staff, meaning that the delivered data does not constitute a complete view of the situation inside of the clinics. To ensure meaningful results for the remainder of the research, we look for types of incidents that are essentially always reported.

Aggression incidents are some of the most common types of incidents inside of the clinics with a higher security level, with the most frequent kind being aggression with objects (e.g. throwing objects at other patients or staff members). These incidents are almost always reported and are heavily represented in the data provided by the company. Besides this, the caregivers report that this (aggression towards staff members) is the most important type of incident for them to prevent. This is in part because aggression can often lead to a domino effect, where a troubled atmosphere inside of the clinics indirectly causes other incidents. This makes the study of the aggression incident the most reliable in terms of the validity and reliability of this research.

Another type of incident that is common and almost always reported is patients refusing to take their medication. However, these incidents are barely influenced by the number of caregivers present or the shared experience of these caregivers. Oppositely, caregivers take a big part in reducing the amount of aggression incidents, as they should be able to predict and prevent these types of incidents based on the patient's behaviour. The knowledge of how to identify this hazardous behaviour and how to de-escalate the situation is mostly learned through experience. From this we conclude that aggression incidents are the most prominent type of incidents to perform research on.

2.2.3 Experience

One surprising finding was that the interviewees did not experience that the chance of an incident occurring depended significantly on the number of caregivers present, or the number of patients present. Contrarily, the communication among members of staff is experienced to worsen with a bigger group of caregivers. This could in turn lead to a situation where an incident might occur. In their perspective, the most important indicator of the chance of an incident occurring was the experience of the caregivers present.

The difference in years of experience differs greatly between the caregivers employed by Trajectum. This difference does not only lie in the total experience of working in the mentally handicapped-care, but also in the experience gained in each department, as the ways of working and types of patients vary greatly across clinics and departments.

At the clinic in Rekken, most departments have caregivers who have worked in their respective department for years, meaning that they have enough experience to identify hazardous situations, given the type of patients and their behaviour. Additionally, Trajectum makes use of a flex-pool, which is a group of employees that can be allocated to any given department when there is a case of understaffing, e.g. due to sickness. These caregivers tend to have little experience at the specific department that they are assigned to, but are often familiar with the patients of the department, as patients and caregivers can cross-communicate inside of the clinic.

Lastly, the clinic also makes use of independent workers (Dutch: ZZP'ers), which can operate between different departments as well as different clinics. These members of staff often have little experience at the assigned department and are less familiar with the patients inside of the clinic. Consequently, the presence of these independent workers can cause distrust and hostility from patients. This atmosphere can lead to an increased chance of incidents occurring. The company only hires these independent workers, if there is a clear shortage of workers on a given day.

2.3 Conclusion

From the observations made and the interviews taken in Rekken, we conclude that we should narrow the scope of this research to types of incidents that have a high rate of (1) occurrence, (2) severity, and (3) detection. This includes, but is not limited to, aggression incidents of any sort (towards caregivers, other patients, or themselves in any way). Other incidents that pass these terms include incidents such as sexual assault, auto-mutilation, and suicidal behaviour. Though less common than acts of aggression, these incidents will also be included in the research. From now on in this research, the term *incident* will refer to these types of incidents, unless stated otherwise. Incidents such as the refusal to take medication are disregarded as the (experience of the) staff has little influence on this, skewing the data.

The experience of the caretakers present at the time of an incident seems to be the most appropriate independent variable for this research, as the caretakers believe that this is the best indicator of an incident occurring inside of the clinic. However, during the data-mining we also consider other factors, such as number of caretakers present, as we aim to find as many correlations in the data as possible. In the ideation and construction of the second deliverable, we consider these properties of the staff, including the years of experience they have at Trajectum and their employment status.

3 Literature review

This chapter answers the second research question, which is as follows: "*In what ways can data-driven working help the mentally handicapped-care?*" First, we set up a theoretical framework, which consists of existing theories and definitions from the literature. Building upon this framework, we conduct a literature review to give an overview of different applications of data-driven working that could be valuable to organisations in the mentally handicapped-care. The conclusion evaluates the practicality of these DDW-techniques when translated to the problem context specific to Trajectum. An overview of the databases and articles used in this chapter is given in the Figure 9.3 to Figure 9.9 in the appendix.

3.1 Theoretical framework

3.1.1 Concepts

To construct a theoretical framework – a foundation made up of definitions and relations in existing research – for this research, it is important to understand how other researchers have conceptualised the theories and variables that are relevant to this research, allowing us to make a research design that builds upon the existing research. These following sections discuss the most important concepts for this research, in relation to the topic of DDW.

3.1.2 Data-driven working

The first of these concepts is *data-driven working* itself. As said before, through the initiative G-AAN, organisations in the mentally handicapped-care in the Netherlands have decided to look into the possibilities that data-driven working could offer to the field. Yet, it is not completely clear to all of these organisations what data-driven working entails, and what its applications are inside and outside of the field of healthcare. According to Stahl et al. (2023), data has become an increasingly more valuable asset to companies, as industries digitalise. Consequently, a data-driven business model (DDBM) is an emerging option that aims to put the potential of the company's data to use, by finding trends in the data and making predictions about the future accordingly.

As organisations in the field of healthcare collect increasingly more data e.g. by documenting cases of incidents, or health related factors of patients (e.g. heart rate or blood pressure), the institutions are left with rapidly and continuously growing sets of data. These types of data sets have proven to be a valuable resource to various institutions in the decision-making processes, both inside and outside of the field of healthcare (Walker et al., 2022; Mukhopadhyay, 2023.). To base decisions on previously collected data is what is generally understood as *data-driven working*, and companies such as Trajectum are justified in expecting data-driven approaches to be of value for decision making, as the unused data that is created within companies create new possibilities for insight generation, among other value offerings (Voigt et al., 2021).

Though numerous examples can be found of companies who capitalised on the raw data that they accumulated over time (e.g. Siegal & Ruoff, 2015), examples of these applications in the mentally handicapped-care were scarce. Consequently, the scope of this literature research included – but was not limited to – the applications of DDW applications in healthcare in general. Later, we will assess how these results (mostly from hospitals) can be translated to the institutions that provide care for people with a (light) mental handicap, which in turn answers the second research question.

3.1.3 Trend analysis

One of the aspects of a DDBM is the central use of data and the resulting need for proper use of trend management. Birkel & Hartmann (2021) define the following: "*Trend management and the associated processes, such as trend analysis, do not pursue the goal of predicting trends. Rather, it is about diagnosing trends and the resulting activities.*"

Within the field of healthcare, there is a big push to digitalise the monitoring of patients (Zhang et al., 2010), as is done in Trajectum regarding incidents. This is because the data that results from this monitoring can be used by the organisations to support their decision making on a managerial level. One of the most common uses of large quantities of data, is a trend analysis, which is used to identify possible relations between certain factors, such as patient information, and an outcome variable, such as the number – or even severity – of incidents inside health institutions (Härkänen et al., 2021). Finding these relations is valuable to the management of organisations, as Mahajan et al. (2019) state: "*This knowledge can allow healthcare leaders to make informed decisions on where to start making changes and how to explore the consequences of potential actions*"

This method of data-analysis can be of use for the case specific to Trajectum. Approaches like these have already shown prominent result in hospitals, but have mostly been executed with data regarding the patients' clinical data (Lucero et al., 2018) or experience of the quality of care (Garay et al., 2015), opposed to the data that is available to Trajectum, such as staff utilisation (a notable exception comes from Leary et al., 2016). Nevertheless, a trend analysis on the data of the incidents in the clinics of Trajectum, considering the staff utilisation, could be the base of a useful model to assess the circumstances under which an incident might occur. This would provide valuable information for the management of the company, as it can be used as a way of forecasting future incidents.

3.1.4 Predictive data

Within the context of the case of Trajectum, what is meant with predictive data, is data which makes predictions about future developments (e.g. the chance of an incident occurring), based on data on previous incidents. This concept is closely related to trend analysis, as Chauhan & Jangade (2016) say the following, in their article specifically regarding the use of big data in healthcare organisations: *"..., Predictive data analytics for big data has potential to take advantages of exploration among healthcare data and extract trends which can benefit the future informed decision making".*

Figure 3.1 explains the stages of data utilisations, as conceptualised in the Gartner Analytics Maturity Model (Davenport & Harris, 2007). In this research, we aim to shift Trajectum's use of data from descriptive (simply describing how many incidents occurred) to predictive (evaluating the chance that an incident will occur). One step further would be prescriptive data, where a decision is automatically made, based on past data. We consider this beyond the scope of this research.

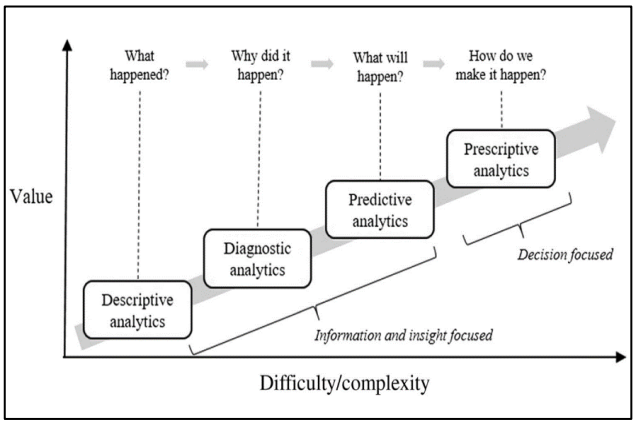

Figure 3.1; types of data utilisation

3.2 Applications

3.2.1 Staff allocation

Staff allocation (also referred to as manpower planning) problems have grown to become a well-studied field of industrial engineering. These types of problems revolve around finding the optimal allocation of staff, based on a stochastic (uncertain) demand (e.g. Campbell, 2011; Zhu & Sherali, 2009). The management of Trajectum faces the same problem within their clinics, where there can be a heavy in-and-out flow of patients (demand), and a scarcity of workers is common.

Numerous institutions have used data-driven approaches to optimise their staff allocation (Walker et al., 2022), but for healthcare institutes, and especially in the (mentally) handicapped-care, these models are not easily adapted. Resource planning in these care institutions is difficult, as patients live in small groups and have a bond with the staff members, meaning that one staff member cannot simply be replaced by another to fulfil the same job in the same place, as would be possible in e.g. a car factory or possibly even a big hospital (various studies were conducted regarding the staff allocation of registered nurses, most notably by Leary et al. (2016) and Zaranko et al. (2023)). In other words, the staff inside of the clinics of Trajectum are not as one-sided of a resource as in other fields and organisations, which makes it hard to simply construct a demand-based working schedule. There are also ethical objections to letting the IT department (and later possibly AI) make decisions on staff allocation, overtaking this responsibility from the management.

Because of these reasons, a suitable application of DDW for Trajectum would not be aimed at making recommendations to the management directly (prescriptive data), regarding where to allocate staff. But rather, we will look into the possibility of making predictions of the chance of incidents occurring (predictive data), based on trends in the data. Providing this information, allows the management of Trajectum to make more informed decisions; working data-driven while keeping the human touch. For institutions that have members of staff whose labour is interchangeable, it might be worth looking into more direct, demand-based models of staff allocation.

3.2.2 Risk

One field that came up in the literature is that of risk prediction-models. These models, which have been applied to hospital settings, aim to assess a prediction of an undesirable event (such as an incident in a clinic or a medical failure), based on a trend analysis within the available data. Prominent examples of this type of model include the efforts of Yu et al. (2018), who aim to assess an indication of the risk of medical failure. One of the models posed in this article was used to assess the relevance of incidents in this research (see Section 2.1)

In another research performed by Lucero et al. (2019), various methods are proposed to assess the risk of falling incidents in hospitals occurring, based on collected clinical data of patients. Beside the relations found between the patients' health-related factors and falling incidents, a significant relation was also found between the incidents and the mix of nurses present, as well as the ratio of registered nurses (similar to the research of Zaranko et al., 2022). Finding a similar relationship in the data of Trajectum would hugely help shape the deliverables. Based on this relation, predictions of higher risks of an incident could be made based on the scheduled staff. If an increased chance of an incident can be identified by the management in this manner, they could make better informed decisions regarding the staff utilisation on any given day.

3.2.3 Neural networks

Though it is outside the scope of this particular research, due to limited data and time constraints, it should be noted that there is a big movement within academics towards the application of artificial intelligence and neural networks to assess risks in healthcare institutions. The field of healthcare produces enormous volumes of data every day, more than other fields such as manufacturing, finance, or media (Balaji & Prasathkumar, 2020). The sheer volume of these so-called big data makes them subject to machine learning techniques and AI-related approaches to find trends in the data. The practicality of these methods is demonstrated by Al Nammari (2020) and Nirmalajyothi (2023), among others. The former looks into the applications of machine learning and AI to improve the safety of patients, while the latter discusses the application of these same methods to construct a risk assessment tool for patients that suffer from chronic diseases.

As the volume of the data of Trajectum is too small to train a proper neural network, and due to a time constraint, we do not expect to be neural networks in the deliverables. The prominence of these techniques, especially in healthcare, should be stated regardless, as a recommendation to Trajectum and other organisations for future data-driven projects.

3.3 Conclusion

From this literature review multiple useful approaches resulted for the problems that Trajectum faces. Through identifying the limitations of the posed models and keeping into account the time constraints and limitations of this research, the most prominent application of data-driven working in the context of the company's current problems, is in the form of a risk-prediction tool. This tool would assess the risk of an incident occurring, given any composition of staff members. This calculated risk is based on the experience of the staff compositions that were present at the incidents. The tool would not propose a different mix of staff members, such as a typical staff allocation model might do, but would rather inform the management of the company of an increased risk of incidents occurring, assisting them in making better informed decisions.

To what extent the construction of this described tool is realistic, as well as the practicality of the tool, depends on the trend analysis, which would stand at the centre of the research. If no significant relations are found between the allocation of staff and the number of incidents, the tool is of minimal value. However, it could – in combination with the recommendations that make up the other deliverable of this research – be a stepping stone for the company to give an impression of the possibilities that data-driven working techniques bring, and later possible applications rooted in machine learning and artificial intelligence.

The available literature on all of these previously mentioned techniques were limited to regular hospitals, rather than clinics in the mentally handicapped-care, such as Trajectum. Having to make a start in the academic literature for this branch is both exciting and limiting. The information that can be taken from previous works is limited and the generalizability of previous works is not fully known, which is at the expense of the validity and reliability of this research. On the other hand, the novelty of the research objectives makes the research more impactful in the academic landscape and can be of great use to healthcare as a whole, if performed right. The next chapter uses parts of this literature review to ideate the models needed to construct the tool that forms the second deliverable.

4 Model ideation

This chapter answers the third research question: *In what ways can data-driven working prevent incidents?* We answer this question by first proposing two models that visualise and conceptualise the possible causes of the incidents that we research at Trajectum. Further, we evaluate the literature to explain how data mining techniques can help us verify meaningful relations in the data, including examples based on the data provided by Trajectum. We conclude with an elaboration on how these supplied data in combination with the models, can be used to prevent incidents.

4.1 Theoretical model

4.1.1 Models

By visualising the variables and the relations between them in a model, the problem becomes more approachable and easier to understand. According to the definitions of Cooper & Schindler (2014), the models shown in Figures 4.1 and 4.2 were constructed. A more detailed explanation of the workings of each type of variable can be found in Figure 9.10 in the appendix. These models visualise how the staff-mix and other variables can influence the number of incidents inside of the clinics of Trajectum, and are based on interviews held with the staff members (Section 2.2.3). The models are slightly different and in the next chapter we statistically analyse both to evaluate which model most closely resembles reality.

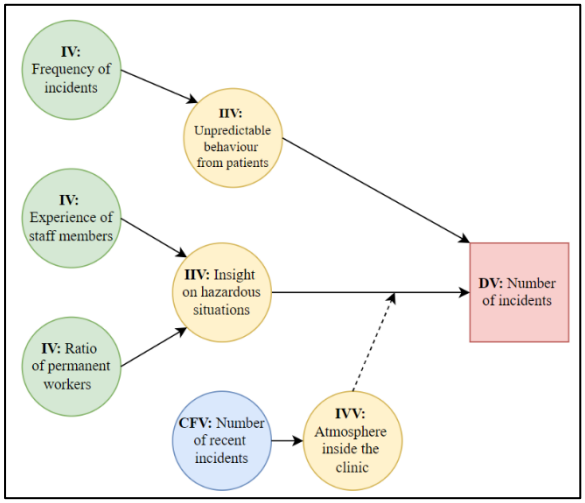

Figure 4.1; First theoretical model

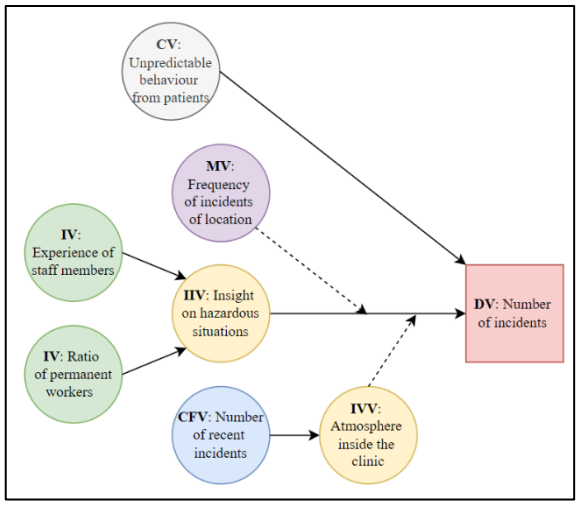

Figure 4.2; Second theoretical model

4.1.2 Similarities

In both models, we believe that the experience of the staff members is an independent variable, which has a significant (negative) relation with the dependent variable (the number of incidents). According to the findings of the first research question, workers with more experience can recognise hazardous situations faster and take preventive measures more effectively than workers with relatively little experience working at Trajectum. Similarly, workers that are permanently hired by Trajectum are regarded as more desirable to plan in than self-employed forces (Dutch: ZZP'ers), as they also tend to be more efficient at recognise hazardous situations, due to their experience at one specific location, whereas self-employed workers often go from clinic to clinic. We assume that the ratio of permanently-employed workers (permanently-employed workers planned at a shift ÷ all workers planned at a shift) is also an independent variable, which has a significant (negative) relation with the dependent variable.

In both models we assume that the atmosphere in a clinic has an effect on the strength of the relation between this independent and the dependent variable. If this atmosphere is hostile, e.g. due to a series of recent incidents, patients tend to get scared and/or more aggressive themselves, meaning that the experience of the workers at preventing and de-escalate hazardous situations is even more important. In this case, the effects of previous incidents inside the clinic are passed through the atmosphere variable to change the relation between the independent variable and the dependent variable. We expect that the (negative) relation between the independent and dependent variable becomes stronger in cases where there were recent incidents.

4.1.3 Differences

In the first model we assume that the level of unpredictable behaviour from the patients can be estimated based on the past frequency of incidents of the location, as a location with a past high frequency of incidents tends to house patients that are more likely to show unpredictable behaviour. We assume this unpredictable behaviour to be an attribute of locations with a high past frequency of incidents, meaning that future unpredictable behaviour will stay more common here than at other locations. However, in the second model we take a different approach by assuming that the level of unpredictable behaviour from the patients is not identifiable by location. This would mean unpredictable behaviour could occur which is virtually impossible to avoid, despite the experience of the workers present. This level of unpredictable behaviour has an impact on the dependent variable, but determining or predicting it is outside of our control, meaning that it is disregarded in the model.

This means that in the first model, the past frequency of incidents at the clinic is considered an independent variable that is expected to have a positive relation with the dependent variable. In the second model, we assume that this frequency is a moderating variable that impacts the *relation* between the insight on hazardous situations and the number of incidents. We expect that in clinics with a higher past frequency of incidents, the relation between the dependent and independent variable is stronger, as the patients show more unpredictable behaviour, meaning that good insight in hazardous situations can be more of use compared to low-security locations. This means that in the second model, the frequency of incidents of the location is expected to have a positive relation with the *strength* of the relation between the independent and dependent variable.

4.2 Data mining

4.2.1 Methods

Especially in healthcare organisations, which often produce continuously and rapidly growing datasets, data mining techniques can be used to find significant relations in these big datasets (Zhang, 2010). In our case, these techniques can be used to verify the relations posed in the models. According to Wissen (2005), data mining entails a set of techniques that aim to find patterns in a set of data, in order to explain the data and make predictions from it. One of these methods is in the form of a linear model, which aims to make a prediction of the output variable (number of incidents) based on the set value of the input values (experience of staff, past frequency of incidents, and ratio of permanent workers).

When all input and output variables can be described in a numeric manner, *multiple linear regression* can be considered to statistically prove the presence of the relations in the data of the incidents. Here the output variable (y) is described as the result of the input variables ($x_1, x_2, ...$) times their respective weight ($\beta_1, \beta_2, \ldots$) and a random error (ε). Such that, $y = \beta_0 + x_1 \beta_1 + x_2 \beta_2 + \cdots + \varepsilon$. Using this technique, we can find the value of the weights ($\beta_1, \beta_2, \ldots$) that represent the strength of the real-world relations between all independent variables and the dependent variable.

If either model resembles reality, the estimate for the variables that measure the experience and the ratio of permanent workers should be negative and statistically significant. If the first model resembles reality, the estimate of the variable that measures the past frequency of incidents should be positive. If this second model resembles reality, the relation between the other independent variables and the dependent variable should be stronger among locations with a higher past frequency of incidents.

Once these posed relations are verified, we can use the results of the regressions to form a predictive model. If the proven relations remain unchanged in the future, we can predict the (average) number of incidents in a shift, based on the variables that posed a statistically significant relation with the dependent variable. Calculating this expected number of incidents in a shift would follow the same formula as the regression: $y = \beta_0 + x_1 \beta_1 + x_2\beta_2 + \cdots$. Here, the dependent variable y, which represents the number of incidents in the shift is unknown and calculated based on the estimates calculated with the regression ($\beta_0,\ \beta_1, \ldots$) in combination with the values that the independent variables ($x_1, x_2, \ldots$) take on for that particular shift.

If this value is higher than usual in that clinic, we can speak of an increased chance of an incident occurring. By identifying shifts with an increased risk of incidents, the management can make better informed decisions on their staff allocation. This data-driven form of risk management can function as a stepping stone for Trajectum and other companies in the healthcare sector alike to explore how their data can assist them in their business processes.

4.2.2 Limitations

A statistical analysis, based on linear regressions, can find meaningful relations between two or more variables in the data provided by Trajectum, as we demonstrate in the next chapter, where we use linear regression to verify the possible relations that were ideated in this chapter. However, there are limitations to this method's ability to find meaningful relations. This is mostly due to the fact that a linear regression is only able to accurately identify linear relations between variables. Using the data provided by Trajectum, we demonstrate this in a simple example.

One of the most straightforward possible indicators of the chance of an incident is the time of day, where workers of Trajectum have experienced that incidents often happen later in the day, as patients' frustration accumulates over the day. This sentiment might very well be true, but due to the cyclical nature of the time of day no signature relation between the time of day and the frequency of incidents can be found using only a linear regression. Figure 9.11 in the appendix shows that there is indeed a pattern in the time of (aggression) incidents that could not be found through a regression, as the relation is not perfectly linear. Meaning that if one of these ideated variables is not shown to have a significant relation with the dependent variable, we should consider that a *non-linear* relation might still be present.

4.3 Conclusion

The research question of this chapter is answered by the ideation of a mathematical model that aims to describe the situation of Trajectum. The most appropriate way for Trajectum to utilise their data is by using it to predict the risk of incidents occurring in their clinics. This is done by applying data mining techniques on the data regarding the past incidents, combined with the work rosters, to find meaningful relations between the number of incidents and variables that might influence this number. Possible relations are ideated and visualised in the proposed models in Figures 4.1 and 4.2. In the next chapter, we will attempt to verify statistically significant relations, which we will utilise in the creation of the prediction model. This model gives the company predictive insight and can assist decision making.

5 Prediction tool

This chapter aims to answer the final research question: "*How can we predict incidents at* Trajectum*?*" This is done by first demonstrating how we manipulated the datasets for this research. Further, data mining techniques are performed, in the form of a regression, following the findings of the third research question. Based on the statistically significant relations found, we propose a model that is able to predict incidents inside of the clinics.

5.1 Data manipulation

Figure 9.12 in Appendix 9.5 shows a relational model of the structure of the database, as it was supplied by Trajectum. By discarding the irrelevant variables from the data tables and including new useful variables that are calculated from the initial data, we are left with a database that is more useful for our research. The relational model of this database can be found in Figure 9.13 in the appendix. The new variables are *Experience* and *Shift*.

The *Shift* variable is added to the data table that holds the information regarding the incidents that have occurred inside of the clinics (Incidents), and the data table that holds all shifts worked at Trajectum (Roster). In the table "Roster", the variable divides the shifts in three different categories by assigning each entry a number between 1 and 3, with the value 1 being the morning-shift (starting between 06:15 and 14:00), value 2 being the evening-shift (starting between 14:00 and 20:00), and value 3 being the night-shift (starting between 20:00 and 06:15). Similarly, in the table "Incidents", each incident is assigned a shift, based on the time of the incident. By specifying this variable in both data tables, it becomes easy to identify which workers were present at the time of an incident, as the variables *Date*, *Location*, and *Shift* together uniquely identify a specific moment and place.

The *Experience* variable is added to the data table that holds the information of every employee of Trajectum (Workers), and reports the years of experience that any given worker has at the company by subtracting the current date from the hiring date. This variable is also referred to in the data table that holds all shifts worked at Trajectum (Roster). For each shift, the experience is linked to the employee that works that specific shift. Through this variable, we can calculate the average years of experience of all workers present during any given shift. The average number of incidents during a shift is expected to be higher if the average experience of the present workers is relatively low.

5.2 Statistical analysis

5.2.1 Data

To perform data mining techniques on the restructured data, we create one table that contains all relevant variables. We did this by (left-)joining the *Roster* data table and the *Incidents* data table, such that for each shift (unique date, location, and shift) the average years of experience of the present staff members and the number of incidents in that shift are specified. Additionally, based on the information in this table, we add the variables that describe the average number of incidents per shift for each specific location, the permanent worker ratio, and the time since the last incident at the location. A more detailed representation of the data table can be found in Figure 9.14 in the appendix. This dataset is used to perform the regressions for both models proposed in the previous chapter. In these regressions, the relations posed in the models are verified by statistically analysing the relation. The model that resembles reality best, based on the findings of the regression, will be used to build a prediction model.

5.2.2 Independent variables

The first regression performed is a multiple linear regression in the form:

$$y_i = \beta_0 + x_{1,i}\,\beta_1 + x_{2,i}\beta_2 + x_{3,i}\,\beta_3 + \varepsilon_i.$$

Where:

y_i represents the number of incidents during shift i,
β_0 represents the intersection of the regression (expected number of incidents if all IVs were to be 0),
$x_{1,i}$ represents the average years of experience of the staff members present at shift i,
β_1 represents the coefficient that determines the strength of the relation between y_i and $x_{1,i}$,
$x_{2,i}$ represents the ratio of permanent workers of shift i,
β_2 represents the coefficient that determines the strength of the relation between y_i and $x_{2,i}$,
$x_{3,i}$ represents the security level of the location of shift i,
β_3 represents the coefficient that determines the strength of the relation between y_i and $x_{3,i}$,
and finally ε_i represents the random error of the regression for each shift i.

By minimising $\sum \varepsilon_i$, we find the values for β_0, β_1, β_2 and β_3. We expect the coefficient β_0 to be a positive number, indicating a positive number of incidents if the other coefficients take on the value 0 (as a negative number of incidents would realistically be impossible). For the coefficient $\beta_1$we expect a negative number, indicating that there is a negative relation between the experience of the staff members and the number of incidents. For the coefficient β_2 we expect a negative number, indicating that there is a negative relation between the share of permanent workers at the shift and the number of incidents. Finally, for the coefficient β_3 we expect a positive number, indicating that there is a positive relation between the security level of the location and the number of incidents. The significance of these relations is determined by the p-value, which should ideally be lower than 0.01.

	Dependent variable:
	`Number of incidents`
Constant	0.001
	p = 0.762
`Average experience`	-0.001***
	p = 0.00001
`Worker Ratio`	0.008**
	p = 0.047
`Average number of incidents`	0.991***
	p = 0.000
Observations	123,765
Residual Std. Error	0.403 (df = 123761)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 5.1; results first regression IVs

Figure 5.1 demonstrates the results of this regression. The estimate for the variable "Worker ratio" is different from what we expected and not statistically significant, meaning we will disregard it in the model. Both the experience of the workers and the frequency of incidents of the location seem to have a statistically significant relation with the dependent variable. The former of these variables can be used in the model to predict the chance of future incidents occurring. However, in this regression, the inclusion of the variable that measures the frequency of incidents of the specific location ("Average number of incidents"), causes some complications.

Though this variable is a good predictor of the chance of an incident occurring *between* locations, its inclusion in the regression complicates the verification of the impact of other independent variables on the dependent variable *within* locations. As the variable is a good predictor of incidents, we will use it in the prediction model. However, the conclusion that future incidents are more likely to occur in clinics with a higher past frequency of incidents does not serve us any new information. Therefore, this variable will be disregarded in future regression (unless mentioned otherwise), such that we can analyse the effect of the individual variables within the clinics.

We perform the regression again, this time only including the relevant independent variable, to evaluate its isolated effects. The results of this regression can be seen in Figure 5.2. In this regression table, we see that the estimate of the variable is negative (as expected) and statistically significant. Meaning that we can assume that the experience of the present staff members has a real-life impact on the number of incidents inside of the clinics. Concludingly, the experience of the workers will be used as an independent variable in the prediction model.

	Dependent variable:
	`Number of incidents`
Constant	0.156***
	p = 0.000
`Average experience`	-0.006***
	p = 0.000
Observations	123,765
Residual Std. Error	0.420 (df = 123763)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 5.2; results second regression IVs

5.2.3 Moderating variable

As we have previously disregarded the frequency of incidents as an independent variable – as it was posed in the first model – due to complications in the regression as the result of the inclusion of this variable, we now research the possible relation proposed in the second model. In this model we propose that the variable has an impact on the *relation* between the other independent variables and dependent variable, making it a moderating variable. As we have discarded the variable that describes the share of permanent workers due to no statistically significant relations being found in the first regression, we aim to verify whether the frequency of incidents has an impact on relation between the average experience of the present workers and the number of incidents.

We do this by evaluating the relation of these variables for five different datasets that include different locations selected based on the frequency of incidents inside of these clinics. We divide the initial dataset into these five smaller datasets such that they all have approximately the same amount of observation. In the regression that follows, if the second model resembles reality, we expect to see that for all datasets the estimate of the variable "Average Experience" is a statistically significant negative number, which decreases (becomes more negative) as the frequency of incidents of each dataset increases. This would represent a larger impact of the experience of workers on the number of incidents in clinics where more incidents occur, the reasoning behind this relation is elaborated on in Section 4.1.3.

	Dependent variable:				
	`Number of incidents`				
	Lower Frequency				Higher Frequency
Constant	0.015***	0.055***	0.100***	0.180***	0.349***
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
`Average experience`	-0.0005***	-0.001***	-0.002***	-0.003***	-0.010***
	p = 0.0001	p = 0.00004	p = 0.006	p = 0.00001	p = 0.000
Observations	25,232	25,844	25,416	22,311	24,282
Residual Std. Error	0.112 (df = 25230)	0.244 (df = 25842)	0.349 (df = 25414)	0.479 (df = 22309)	0.657 (df = 24280)
Note:					*p<0.1; **p<0.05; ***p<0.01

Figure 5.3; Results regression MV

Figure 5.3 shows that the strength of the relation between the experience of the workers (IV) and the number of incidents (DV) is indeed dependent on the frequency of incidents, in the manner that we predicted. The constant logically increases as the frequency of incidents increases, as this constant represents the average number of incidents of each shift. The estimate of the variable "Average experience" is negative and statistically significant for each dataset. Additionally, the estimate decreases as the frequency of incidents of the dataset increases, indicating a stronger relation with the dependent variable, as the frequency of incidents in the clinics increases.

5.2.4 Confounding variable

Finally, we evaluate the effects of the confounding variable of the proposed models, which is the variable which measures the hostility of the atmosphere in the clinic, based on if there was a recent aggression incident. According to both proposed models, a recent incident in a clinic could result in a hostile atmosphere in the same clinic. In this hostile environment, experienced workers would be especially valuable for identifying and deescalating hazardous situations (see Section 4.1.3). Consequently, we expect that the relation between the independent and dependent variable becomes stronger (more negative) as the latest incident occurred more recently.

The presence of this possible relation is verified in a manner similar to the moderating variable of the previous section. Each shift is assigned a number ranging from 0 to 6, indicating how long ago the last aggression incident at that specific location occurred, where a value of 0 indicates that the last incident was longer than a week ago, and the value of 6 indicates that an incident occurred the day prior at that location. The data is split up into different datasets, based on the value of this variable, and a separate regression is performed, to verify the relation between the independent and dependent variable at every level of hostility. The results can be found in Figure 5.4.

	Dependent variable:						
	`Number of incidents`						
	1 day ago	2 days ago	3 days ago	4 days ago	5 days ago	6 days ago	>=7 days ago
Constant	0.302***	0.227***	0.175***	0.155***	0.142***	0.113***	0.059***
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
`Average experience`	-0.009***	-0.007***	-0.004***	-0.004***	-0.004***	-0.003***	-0.002***
	p = 0.000	p = 0.000	p = 0.00001	p = 0.0003	p = 0.001	p = 0.002	p = 0.000
Observations	22,682	14,585	10,454	7,971	6,262	5,015	56,796
Residual Std. Error	0.611 (df = 22680)	0.520 (df = 14583)	0.467 (df = 10452)	0.433 (df = 7969)	0.404 (df = 6260)	0.344 (df = 5013)	0.239 (df = 56794)
Note:							*p<0.1; **p<0.05; ***p<0.01

Figure 5.4; Results first regression CFV

This result supports the previously posed hypothesis, which states that the strength of the relation between the independent and dependent variable is dependent on the variable "Recent incident". As speculated, this strength increases as the last incident becomes more recent. However, we observe that the constant, which represents the expected number of incidents per shift, increases as past incidents become more recent. This was to be expected, as it is more likely that there was a recent incident at a location where incidents are frequent than at a location where virtually no incidents occur. It begs the question whether the change of strength of the relation between the independent and dependent variable is actually caused by a recent incident occurring, or if the effects of the variable that represents the frequency of incidents might mediate through the "Recent incident" variable, causing this result. To study the isolated effect of this variable, we perform another regression, including both the variable "Recent incident" and "Frequency of incident". The results are found in Figure 5.5.

	Dependent variable:						
	`Number of incidents`						
	1 day ago	2 days ago	3 days ago	4 days ago	5 days ago	6 days ago	>=7 days ago
Constant	0.072***	0.034***	0.023**	0.010	0.029***	0.027***	0.010***
	p = 0.000	p = 0.001	p = 0.022	p = 0.323	p = 0.008	p = 0.009	p = 0.00000
`Average experience`	-0.004***	-0.003***	-0.001	-0.001	-0.002	-0.002*	-0.001***
	p = 0.0002	p = 0.002	p = 0.189	p = 0.323	p = 0.104	p = 0.098	p = 0.006
`Average number of incidents`	0.976***	0.925***	0.821***	0.878***	0.771***	0.648***	0.651***
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Observations	22,682	14,585	10,454	7,971	6,262	5,015	56,796
Residual Std. Error	0.600 (df = 22679)	0.509 (df = 14582)	0.459 (df = 10451)	0.424 (df = 7968)	0.397 (df = 6259)	0.340 (df = 5012)	0.236 (df = 56793)
Note:							*p<0.1; **p<0.05; ***p<0.01

Figure 5.5; Results second regression CFV

The results show that the isolated effect of the "Recent incident" variable is limited. We observe a slight downward trend in the value of the estimates for the independent variable, but the results are less striking as well as less statistically significant. It seems that most of the predicting power of the variable "Recent incident" was characterised by its correlation with the variable that describes the frequency of incidents between the clinics. The correlation between all potential independent variables is depicted in the correlation matrix in Figure 9.15 in the appendix. This matrix also confirms that there is a significant correlation between these two variables which could explain the results of these regressions.

One should wonder whether the "Recent incident" variable might be a good predictor of the dependent variable itself. We perform another regression, including this variable among the other independent variables to evaluate whether it is more appropriate to treat the "Recent incident" variable as an independent variable or a confounding variable, in the construction of the prediction tool. Note that we include the variable which measures the frequency of incidents at the location and shift, as otherwise it is possible that the effects of this variable get transmitted through the "Recent incident" variable, giving a skewed view of the effects of the "Recent incident" variable on the dependent variable. Including both variables in the regression ensures that we can observe the isolated effect of the "Recent incident" variable.

	Dependent variable:
	`Number of incidents`
Constant	-0.002
	p = 0.300
`Average experience`	-0.001***
	p = 0.00001
`Recent incident`	0.011***
	p = 0.000
`Average number of incidents`	0.850***
	p = 0.000
Observations	123,765
Residual Std. Error	0.405 (df = 123761)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 5.6; Results first regression CFV as IV

The results of this regression, displayed in Figure 5.6, show that the estimate of the variable is positive (as expected) and statistically significant, meaning that there is a meaningful relation between this variable and the dependent variable. Consequently, in the construction of the predicting tool, we now treat the variable "Recent incident" as an independent variable, meaning manipulating it has a direct impact on the expected number of incidents. As a final step, we inspect whether the relation between this variable and the independent variable might be influenced by the moderating variable "frequency of incidents".

The presence of this relation is verified by dividing the dataset into multiple datasets, in the same manner as for the independent variable "Average experience" in the previous section. These smaller datasets hold the shifts of different locations based on the level of frequency of incidents. We observe the estimates of the variable "Recent incident" and look for trends between the different levels of frequency of incidents. The results of this final regression are given in Figure 5.7.

	Dependent variable:				
	`Number of incidents`				
	Lower frequency				Higher frequency
Constant	0.013***	0.047***	0.076***	0.119***	0.218***
	p = 0.000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
`Average experience`	-0.0004***	-0.001***	-0.002**	-0.003***	-0.009***
	p = 0.0004	p = 0.0001	p = 0.016	p = 0.00000	p = 0.000
`Recent incident`	0.003***	0.006***	0.010***	0.018***	0.029***
	p = 0.00000	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Observations	25,232	25,844	25,416	22,311	24,282
Residual Std. Error	0.112 (df = 25229)	0.243 (df = 25841)	0.349 (df = 25413)	0.477 (df = 22308)	0.655 (df = 24279)
Note:					*p<0.1; **p<0.05; ***p<0.01

Figure 5.7; Results second regression CFV as IV

From this final result, we can see that there is a positive relation between the frequency of incidents inside of the location and the strength of the relation between the independent variable "Recent incident" and the number of incidents, which serves as the dependent variable. This means that in clinics where incidents happen more frequently, the occurrence of a past incident is expected to have a higher chance of causing future incidents than in a clinic where incidents are relatively less frequent.

5.2.5 Predicting model

Concluding the statistical analysis, we have analysed every variable discussed in the proposed models and we have evaluated their position in relation to the dependent variable. Neither of the proposed models perfectly resemble reality, according to the regressions. However, the models served as a benchmark which helped the research by giving direction to the regressions. The conclusions made based on these regressions helped us form the model visualised in Figure 5.8.

In this model, both the experience of the present workers and the days since the last incident are expected to have a direct relation with the number of incidents that happen during a shift. The strength of both of these relations is dependent on the past frequency of the location, increasing the strength of the relations as this frequency increases (in the case of the "Average experience" variable, the relation becomes more negative. in the case of the "Recent incident" variable, the relation becomes more positive).

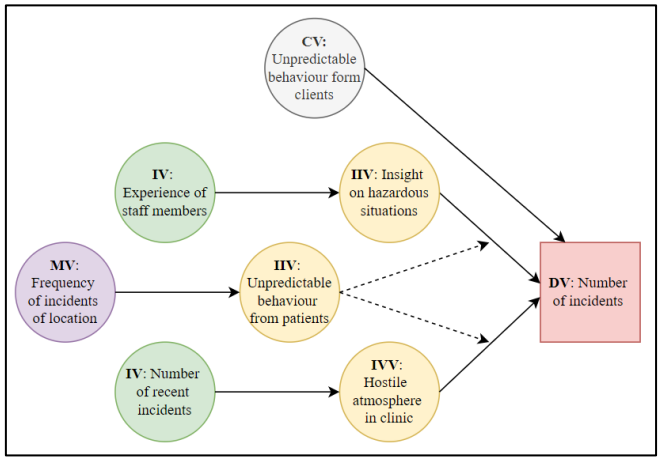

Figure 5.8; Final theoretical model

5.3 Prediction method

5.3.1 Relations

Following the model proposed in Figure 5.8, we can assign a value to the expected number of incidents at any given shift, when we know the exact values of the other variables and the strength of the relations between them. This "fitted" value represents the increased or decreased chance of an incident occurring during the shift, depending on whether the value is above or below the average expected number of incidents at that specific location. In this section, we calculate the strengths of the relations of the proposed model and give the formula that provides the fitted value. We calculate the fitted value on past data and compare it to the actual number of incidents, to evaluate the practicality of the tool. Finally, we demonstrate the workings of the tool on future shifts, as to demonstrate how the company could identify especially hazardous situations.

The strength of the relation between both independent variables and the dependent variable is dependent on the frequency of incidents of the location. This means that to calculate a fitted value, which predicts the relative chance of an incident occurring, we first need to establish the average frequency of incidents at the location for which we are calculating this value. This is done using the data on past incidents. Further, we calculate the strength of the relation between the "Average experience" variable and the dependent variable and the strength of the "Recent incident" variable and the dependent variable. The fitted value can be calculated using the previously calculated coefficients (which represent the strength of the relations), in combination with the values that the variables take on, which are extracted from the future schedule data (in the case of the "Average experience" variable) and the past incident data (in the case of the "Recent incident" variable). This process is visualised in Figure 5.9.

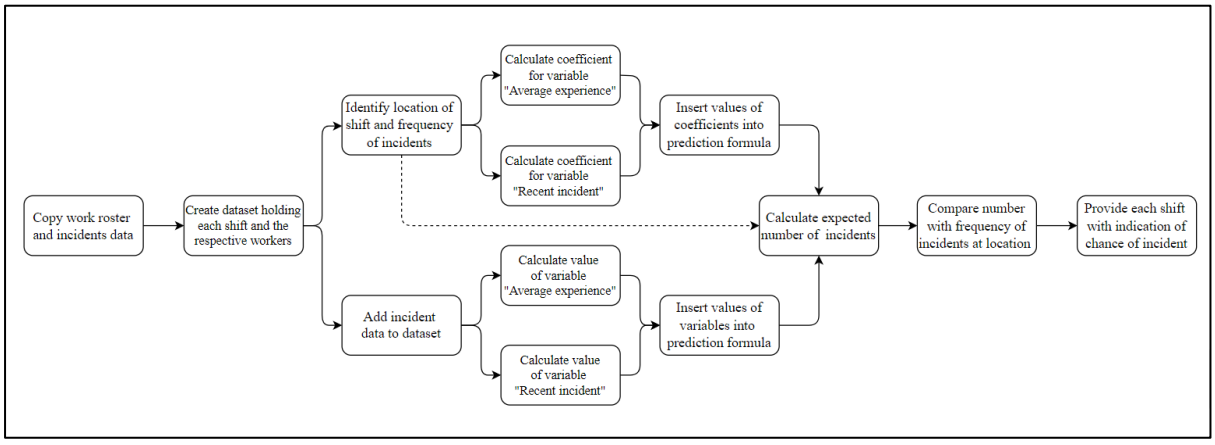

Figure 5.9; Process model prediction tool

5.3.2 Coefficients

First, we performed simple linear regressions for every individual location, measuring the relation between both independent variables against the dependent variable. Using the coefficients that resulted from these regressions and the respective frequency of incidents at each location, we performed another regression to establish how the value of the coefficient changes as the frequency of incidents is manipulated. Figures 5.10 and 5.11 show the change in coefficient for the variables "Average experience" and "Recent incident", respectively. The regression tables in Figure 9.16 and 9.17 in the appendix show that the estimate of the regression constant is not statistically significant, meaning we can assume that this constant is in reality zero.

Additionally, the coefficient for the *strength* of the relation between the "Average experience" and the dependent variable, measuring the dependency of the relation's strength on the frequency of incidents at the location, is -0.045. This means that if a location were to have an average of 0.1 incidents per shift, the value in the prediction model for the estimate of the coefficient of the relation between this independent variable and the dependent variable would be $-0.045 * 0.1 = -0.0045$, meaning that with every (mean) year of experience the expected number of incidents during that shift decreases with 0.0045. The same type of coefficient for the strength between the "Recent incident" variable and the dependent variable is 0.085, meaning that in the same clinic with an average frequency of 0.1 incident per shift, the coefficient measuring the strength between this independent variable and the dependent variable takes on the value of $0.085 * 0.1 = 0.0085$. This means that if an incident occurs in this clinic after a week of no aggression incidents ("Recent incident" variable goes from 0 to 6) the expected number of incidents during a shift the next day increases with $0.0085 * 6 = 0.051$.

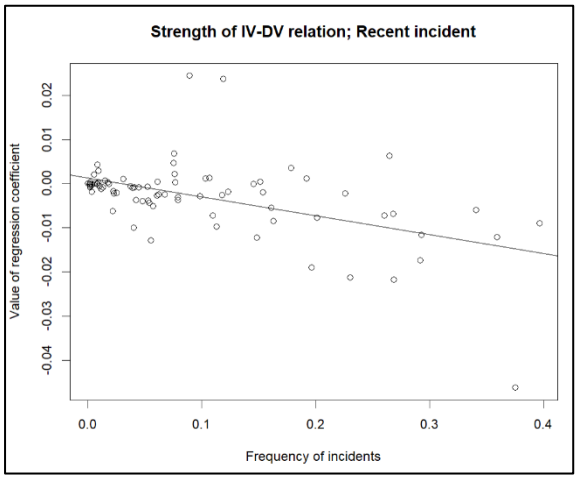

Figure 5.10; coefficients "Average experience"

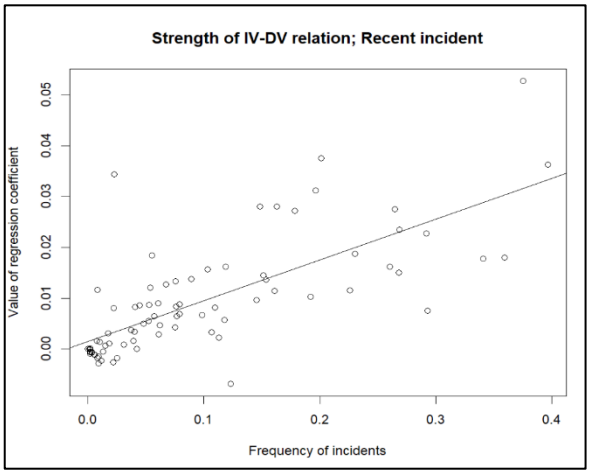

Figure 5.11; coefficients "Recent incident"

5.3.3 Prediction formula

Following these regressions, the prediction formula is constructed and defined as follows:

$$E[\#\ of\ incidents] = FOI\ + \mathrm{x}_1 * (FOI * -0.045) + \mathrm{x}_2 * (FOI * 0.085)$$

Where:
E[# of incidents]: Expected Number of incidents during the shift,
FOI: Frequency of incidents of the location and shift,
x_1: Value of the variable "Average experience" (mean years of staff present during the shift),
x_2: Value of the variable "Recent incident" (ranging from 0-6 in increasingly recent order).

The outcome of this formula is heavily dependent on the past frequency of incidents (FOI) of the locations, meaning that this value is a good prediction of the risk of an incident considering the scope of the whole of Trajectum. However, this value is not a good predictor of the *relative* risk within locations, caused by suboptimal staff allocation. By dividing the outcome of the formula with the FOI of the type of shift at the specific clinic, we obtain the relative chance of an incident occurring during the shift, we call this the relative risk factor, or RRF.

$$RRF = \frac{E[\#\ of\ incidents]}{FOI}$$

Based on the RRF, we assign a categorical value to each shift, which describes the risk of an incident occurring during the shift in nominal terms. The terms are: Very Row Risk, Low Risk, Normal Risk, Increased Risk, and High Risk. An example of an output is given in Figure 5.12.

Date	Location_SKey	Shift	Average experience	Recent incident	Average number of incidents	Expected incidents	Relative hazard	Riskmeter
2023-01-01	106	1	5.0000000	1	0.042094456	0.036201232	0.8600000	Normal Risk
2023-01-01	107	1	7.1333333	0	0.061601643	0.041827515	0.6790000	Low Risk
2023-01-01	108	1	12.7500000	0	0.013347023	0.005689168	0.4262500	Very Low Risk
2023-01-01	109	1	4.4000000	0	0.047227926	0.037876797	0.8020000	Normal Risk
2023-01-01	110	1	3.4000000	0	0.027720739	0.023479466	0.8470000	Normal Risk
2023-01-01	111	1	10.4666667	0	0.047227926	0.024983573	0.5290000	Low Risk
2023-01-01	112	1	8.8000000	6	0.015400411	0.017156057	1.1140000	Normal Risk
2023-01-01	113	1	10.4000000	0	0.012320329	0.006554415	0.5320000	Low Risk
2023-01-01	137	1	1.6200000	6	0.240246407	0.345258111	1.4371000	High Risk
2023-01-01	138	1	4.9333333	6	0.030800821	0.039671458	1.2880000	Increased Risk
2023-01-01	144	1	3.3333333	0	0.224845996	0.191119097	0.8500000	Normal Risk
2023-01-01	145	1	0.8000000	5	0.096707819	0.134327160	1.3890000	High Risk
2023-01-01	146	1	8.4333333	5	0.132443532	0.138469713	1.0455000	Normal Risk
2023-01-01	147	1	11.5666667	5	0.183778234	0.166227413	0.9045000	Normal Risk

Figure 5.12; Output tool

5.3.4 Evaluation

The evaluation of this tool is best done using data regarding incidents and the work roster that has not been used to train the model. By verifying the relation between the expected number of incidents (output first formula) and the realised number of incidents, using a regression, a strong relation should be found. The same should be true for a simple linear regression between the RRF (output second formula) and the realised number of incidents. To test the model, we use data from 2024-05-29 to 2024-07-11, which was not included in the training set (2022-01-01 to 2024-05-28).

The results of the regressions between the expected number of incidents and the realised number of variables is given in Figure 5.13, and similarly for the risk indicator RRF in Figure 5.14. Both estimates are of remarkable statistical significance with the estimate of the regression for the expected number of incidents and the estimate of the regression for the RRF posing t-values of 28.6 and 15.0 respectively. This result indicates that both outputs of the tool would be good predictors of risk of an incident.

	Dependent variable:
	`Number of incidents`
Constant	-0.010
	t = -1.276
	p = 0.202
`Expected incidents`	1.023***
	t = 28.648
	p = 0.000
Observations	5,596
Residual Std. Error	0.426 (df = 5594)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 5.13; Evaluation E[# of incidents]

	Dependent variable:
	`Number of incidents`
Constant	-0.139***
	t = -7.375
	p = 0.000
`Relative hazard`	0.283***
	t = 15.025
	p = 0.000
Observations	5,596
Residual Std. Error	0.447 (df = 5594)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 5.14; Evaluation RRF

5.4 Conclusion

The results of the research performed in this chapter shows how we can predict incidents inside of the clinics of Trajectum, based on the independent variables ideated in the previous chapter. According to a statistical analysis, the variables that measure the average experience of the workers present during a shift ("Average experience") and the variable that represents the time since the last incident at the location ("Recent incident") both proved to have a significant relation with the number of incidents during the shift (DV), making them independent variables. Additionally, the statistical analysis revealed the relation between the variable that measures the average frequency of incidents inside of the location ("Frequency of incidents") and the *strength* of the relations between the aforementioned independent variables and the dependent variable, making it a moderating variable (see Figure 5.10 and 5.11).

Using this knowledge, we construct a model that calculates the expected number of incidents at any given shift, given the value of these variables. The output of the model is effective at predicting an increased chance of an incident occurring, mostly measuring the differences in risk *between* locations. Additionally, the relative risk factor (RRF) is calculated, which removes the location-based bias and provides the company with an assessment of the risk of an incident occurring relative to the usual frequency of incidents at the location. This output is effective at predicting an increased risk of incidents *within* different locations.

6 Implementation plan

This chapter concludes the research by answering the main research objective: *"How can we implement a data-driven way of working in companies in the (mentally) handicapped-care to prevent future incidents?"* This is done by elaborating on how Trajectum and the other organisations of the initiative G-AAN should implement the deliverables of this research into their organisations. This is first done for the deliverable that resulted from the first two research questions. This deliverable is an assessment of how organisations such in the mentally handicapped-care can benefit from data-driven working. These conclusions regard all organisations of G-AAN. Additionally, we share the implementation plan specific to Trajectum, where we discuss how the company can properly utilise the prediction tool, the deliverable that resulted from the research performed for the last two research questions.

6.1 First deliverable

The assessments in the Chapter 3, which make up the first deliverable, provide companies in the mentally handicapped-care with an introduction to the broad applications of data-driven working, as well as the limitations that are specific to this sector. Making this deliverable of use to both Trajectum as well as to the other companies that make up the initiative G-AAN.

For Trajectum specifically, this deliverable is the foundation of the prediction tool that serves as the second deliverable of this research. Whereas other companies can implement the results of this research into their company by assessing which of the posed applications and limitations of DDW applies to their own organisations. To these companies, this deliverable is most useful as an example of how data-science can help their own organisation through risk prediction models, which uses relatively simple data-mining techniques. Depending on the organisation, different risks might need to be assessed and reduced, where different limitations, independent variables, and different types of incidents might apply. This deliverable might serve as a blueprint for how these models can be ideated. Additionally, for some organisations the limitations specific to Trajectum, regarding the interchangeability of employees across clinics, may be less relevant. They could conclude from the research that it would be valuable for them to look into more direct forms of data-driven staff allocation systems.

It is crucial that the people that make up these companies gain trust in the potential of data-driven working. Companies in the sector of healthcare, and especially the (mentally) handicapped-care, tend to be hesitant at implementing new technologies into their decision making. A sentiment that limits the speed of the implementation of new technologies. This research demonstrates that data can support business processes, while retaining the human touch needed in a sector where these processes are as personal as at Trajectum. As a result, this research could be a gateway for Trajectum, and other companies alike, to experience the functionality of data-driven working and its further applications.

6.2 Second deliverable

To fully benefit from this research, some steps have to be taken at Trajectum to make optimal use of the tool. First, the roster maker of Trajectum has to be informed about the workings of the tool, such that when an increased risk of an aggression incident is detected, the management can be alerted and can take a look at the factors that result in the increased risk (low experience of staff and/or hostile atmosphere in clinic). The responsibility of acting on a possible increased risk lies completely with the management of Trajectum, as the tool is meant to merely assist the management to make decisions based on the data. Consequently, the impact that data-driven decision making can have on the company is hugely dependent on the management's trust in data-driven working.

The tool needs to be regularly updated, as the value of the variable “Recent incident” for a shift can hugely differ depending on the day the shift is evaluated. If an aggression incident occurs in a clinic where no such incident had occurred in the previous week, the variable “Recent incident” on the day of the incident will be of the value 0, while the next day this value changes to 6. This change hugely influences the expected number of incidents and the predicted relative risk factor of the shift.

For shifts that are at least a week in the future, the “Recent incident” variable always takes on the value 0, with the estimates becoming more accurate as shifts move closer to the present. Consequently, the estimations of the risk of incidents for shifts in the future is structurally underestimated, as these estimates do not accurately reflect the effects of the “Recent incident” variable, sometimes underrepresenting the eventual value of the variable, but never overrepresenting it. The resulting underestimation is visualised in Figure 6.1, where the data of incidents stops at 2024-05-28. Note that, similarly to Figure 5.12 and Figure 6.2, the variable “Location_SKey” refers to the clinics of Trajectum. Due to privacy reasons, the names of these clinics are not shown. Though, in the UI that the company receives the names of these clinics may be shown alongside the “Location_SKey” variable.

Date	Location_SKey	Shift	Average number of incidents	Average experience	Recent incident	Expected incidents	Relative hazard	Riskmeter
2024-05-22	189	2	0.4529877	1.1000000	6	0.6343187	1.4003000	High Risk
2024-05-23	189	2	0.4529877	2.3142857	6	0.5974649	1.3189429	High Risk
2024-05-24	189	2	0.4529877	2.0000000	6	0.6070035	1.3400000	High Risk
2024-05-25	189	2	0.4529877	4.8142857	6	0.5215894	1.1514429	Increased Risk
2024-05-26	189	2	0.4529877	5.0142857	6	0.5155194	1.1380429	Increased Risk
2024-05-27	189	2	0.4529877	2.6000000	6	0.5887934	1.2998000	High Risk
2024-05-28	189	2	0.4529877	3.2714286	6	0.5684154	1.2548143	High Risk
2024-05-29	189	2	0.4529877	1.7333333	5	0.5793109	1.2788667	High Risk
2024-05-30	189	2	0.4529877	2.4571429	4	0.5215571	1.1513714	Increased Risk
2024-05-31	189	2	0.4529877	1.6333333	3	0.5107738	1.1275667	Increased Risk
2024-06-01	189	2	0.4529877	2.5500000	2	0.4471668	0.9871500	Normal Risk
2024-06-02	189	2	0.4529877	3.3857143	1	0.3860167	0.8521571	Normal Risk
2024-06-03	189	2	0.4529877	4.8714286	0	0.3051390	0.6736143	Normal Risk
2024-06-04	189	2	0.4529877	1.5285714	0	0.4065953	0.8975857	Normal Risk
2024-06-05	189	2	0.4529877	2.8142857	0	0.3675736	0.8114429	Normal Risk
2024-06-06	189	2	0.4529877	3.1000000	0	0.3589022	0.7923000	Normal Risk

Figure 6.1; Structural underestimation of RRF as incident data stops at 2024-05-28

It would be a possibility to run the tool every day, as runtime is around 10 minutes when using 2 years of data. However, the easiest way to update the tool regularly is to integrate it into the IT-infrastructure of Trajectum, relating it to the database that holds the data of the roster and incidents. To do this, efforts of the IT-department are necessary, as they understand this infrastructure best. After integrating the tool into the IT-infrastructure of Trajectum, the management of Trajectum can be presented with an understandable user interface to make the decision-making process more approachable.

This UI can be displayed in Excel, where approaching shifts are represented in rows and highlighted when an increased hazard is predicted. Besides the shift information (date, location, etc.) and prediction (RRF, risk-meter), the row should also include the independent variables, such that the user can identify the reasoning behind the increased hazard. A UI of this kind could look like the UI in Figure 6.2.

Date	Location_SKey	Shift	Average experience	Recent incident	Average number of incidents	Expected incidents	Relative hazard	Riskmeter
01/06/2024	172	1	5,22	0	0,02048	0,01981	0,97	Normal Risk
01/06/2024	172	2	8,25	0	0,06834	0,05729	0,84	Normal Risk
01/06/2024	173	1	7	0	0,0262	0,02335	0,89	Normal Risk
01/06/2024	173	2	3,35	0	0,07877	0,08244	1,05	Normal Risk
01/06/2024	174	1	3	3	0,09226	0,12007	1,3	Increased Risk
01/06/2024	174	2	1,47	3	0,19886	0,27177	1,37	High Risk
01/06/2024	176	1	12,13	0	0,05233	0,03524	0,67	Low Risk
01/06/2024	176	2	10,8	0	0,05263	0,03842	0,73	Low Risk
01/06/2024	176	3	11	0	0,01164	0,0084	0,72	Low Risk
01/06/2024	177	1	9,78	0	0,10239	0,0792	0,77	Low Risk
01/06/2024	177	2	2,6	0	0,21502	0,2319	1,08	Normal Risk
01/06/2024	177	3	3,2	0	0,05239	0,05517	1,05	Normal Risk
01/06/2024	178	1	17,9	3	0,17065	0,11404	0,67	Low Risk
01/06/2024	178	2	4,3	3	0,21047	0,26229	1,25	Increased Risk
01/06/2024	178	3	1,5	3	0,07973	0,10885	1,37	High Risk
01/06/2024	179	1	3,03	0	0,14448	0,15316	1,06	Normal Risk
01/06/2024	179	2	17,06	0	0,21843	0,10134	0,46	Very Low Risk
01/06/2024	179	3	12,6	0	0,09101	0,05948	0,65	Low Risk
01/06/2024	180	1	4,35	0	0,16951	0,17021	1	Normal Risk
01/06/2024	180	2	2,8	0	0,15245	0,16312	1,07	Normal Risk

Figure 6.2; User interface in Excel

It should be noted that the highlighted results represent the risk of an incident occurring *relative* to the average frequency of incidents in the same clinic during the same type of shift (morning, evening, or night). This means that the value “High Risk” does not necessarily indicate an abnormally high chance of an incident occurring during that shift, but rather a situation with exceptionally high risk of an incident for that clinic. Management should decide to act when they deem the *combination* of the expected number of incidents (measures risk *between* clinics and shifts) and the RRF (measures risk *among* clinics and shifts) too high.

Trajectum should evaluate the performance of the tool by verifying the relation between the predicted risk factors and the realised number of incidents within the clinics. The formula that is used to predict the expected number of incidents uses values calculated using the last 2 years of data. We do not expect these values to change significantly over time. However, as new incidents are reported, it is possible for the strength of any relation between variables to (slowly) change. To make this tool future proof, these variables should be recalculated from time to time, as different underlying factors might influence the strength of the independent variables on the dependent variable. As these values are based on large sums of data, they do not have to be evaluated constantly, once every quartile will presumably be enough to keep the tool accurate to its environment.

To conclude, what is needed for Trajectum to take full benefit from this research is (1) an understandable UI of the tool, which is integrated into the IT-infrastructure and can be used by the management of Trajectum. (2) Occasional (yet planned) evaluations of the strength of the variables and the performance of the tool itself. And importantly, (3) trust from the management of Trajectum in the potential of data-driven decision making in the mentally handicapped-care.

7 Recommendation

In this final chapter, we give our separate recommendations regarding the use of this research to Trajectum and to the companies that are part of the initiative G-AAN, in the first and second section, respectively. Both of these sections include recommendations on relevant topics of future research that we believe could have fruitful results. These researches could add to or build on this research.

7.1 Recommendations Trajectum

The first recommendation that we have for the company to make optimal use of this research is to follow the implementation plan, integrating the tool into the IT-infrastructure of Trajectum with the help of the company's IT-department. Subsequently acting on the predicted risks can be a first step for Trajectum to consciously make decisions supported by data. If these decisions result in a significant decrease in aggression incidents, this would lead to an increase in the employee's confidence in DDW that is needed for the company to fully utilise this research. This trust is crucial for an environment where (future) data-driven projects are frequent and effective.

This leads to the second recommendation that we have for the company, which is to immediately look into more data-driven projects. It became apparent how much potential there is for data-driven decision-making at Trajectum, considering the large volumes of data that they record constantly. This leads to the recommendations we have for Trajectum regarding future research.

First, future research could build on this research by expanding this existing model to include more (independent) variables, relations, and by e.g. including more types of incidents (where the current model mainly includes aggression incidents). Additionally, new (risk prediction-)models can be constructed based on different processes of Trajectum such as the prediction of omissions or the flow of patients through the different clinics of Trajectum. These models can also utilise the information of the first deliverable, and can be constructed using the same methodology of this research (Chapter 4 and 5). These data-driven projects can also help increase the trust of the company in DDW.

7.2 Recommendations G-AAN

To the companies of G-AAN we also make the recommendation to look into new data-driven projects based on this research. Each company should consider the results of this research and how these techniques apply to their respective business processes. Similarly to Trajectum, they should consider risk prediction models to support their decision making. This way, research similar to this one can be performed in different problem contexts, resulting in the use of different variables and relations.

Additionally, they should consider the limitations of this research and evaluate whether the same limitation applies to their company. For example, it could be possible that an organisation in G-AAN is less limited by the rate of specialisation of their staff, meaning they could look into more direct demand-based staff allocation solutions, which got ruled out in this research as a result of the specific problem context of Trajectum. Regarding the use of AI in future research, the organisations of G-AAN could consider centralising the data, meaning the data (e.g. regarding incidents) collected at each organisation follow the same format and are pooled together. This would result in a more efficient and larger scale dataset to train the neural network on.

8 References

Al Nammari, R. H. (2020, November). Promoting Patient Safety through Machine Learning. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 657-662). IEEE.

Balaji, S., & Prasathkumar, V. (2020, January). Dynamic changes by big data in health care. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-4). IEEE.

Birkel, H., & Hartmann, E. (2021). Development of a trend management process for supply chain management in the context of industry 4.0. Digital Business Models in Industrial Ecosystems: Lessons Learned from Industry 4.0 Across Europe, 23-34.

Campbell, G. M. (2011). A two-stage stochastic program for scheduling and allocating cross-trained workers. Journal of the Operational Research Society, 62(6), 1038-1047.

Chauhan, R., & Jangade, R. (2016). A robust model for big healthcare data analytics. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 221-225). IEEE.

Cooper, D. R., & Schindler, P. (2014). Business research methods. McGraw-Hill.

Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: the new science of Winning. *Harvard business review press, Language*, *15*(217), 24.

Delen, D., & Ram, S. (2018). Research challenges and opportunities in business analytics. Journal of Business Analytics, 1(1), 2-12.

Garay, J., Cartagena, R., Esensoy, A. V., Handa, K., Kane, E., Kaw, N., & Sadat, S. (2015). Strategic analytics: towards fully embedding evidence in healthcare decision-making. Healthcare quarterly (Toronto, Ont.), 17 Spec No, 23–27.

Härkänen, M., Vehviläinen-Julkunen, K., Franklin, B. D., Murrells, T., & Rafferty, A. M. (2021). Factors Related to Medication Administration Incidents in England and Wales Between 2007 and 2016: A Retrospective Trend Analysis. Journal of patient safety, 17(8), e850–e857.

Heerkens, H., & Van Winden, A. (2021). *Solving managerial problems systematically*. Routledge.

Hlavac, M. (2022). Stargazer: Well-Formatted Regression and Summary Statistics Tables; 2018. R package version 5.2. 2.

Houtmeyers, K. C., Jaspers, A., & Figueiredo, P. (2021). Managing the training process in elite sports: From descriptive to prescriptive data analytics. International Journal of Sports Physiology and Performance, 16(11), 1719-1723.

Jong, T. C. de. (2019). *Optimizing mental health care processes through data-driven operations management* [Bachelor thesis, University of Twente]. University of Twente. https://essay.utwente.nl/79494/

Kaiser, I. (2012). Collaborative trend analysis using web 2.0 technologies: a case study. International Journal of Distributed Systems and Technologies (IJDST), 3(4), 14-23.

Leary, A., Cook, R., Jones, S., Smith, J., Gough, M., Maxwell, E., Punshon, G., & Radford, M. (2016). Mining routinely collected acute data to reveal non-linear relationships between nurse staffing levels and outcomes. BMJ open, 6(12), e011177.

Lucero, R. J., Lindberg, D. S., Fehlberg, E. A., Bjarnadottir, R. I., Li, Y., Cimiotti, J. P., ... & Prosperi, M. (2019). A data-driven and practice-based approach to identify risk factors associated with hospital-acquired falls: Applying manual and semi-and fully-automated methods. International journal of medical informatics, 122, 63-69.

Mahajan, A., Madhani, P., Chitikeshi, S., Selvaganesan, P., Russell, A., & Mahajan, P. (2019). Advanced Data Analytics for Improved Decision-Making at a Veterans Affairs Medical Center. Journal of healthcare management / American College of Healthcare Executives, 64(1), 54–62.

Mukhopadhyay, S. (2023). Modelplasticity and abductive decision making. Decisions in Economics and Finance, 46(1), 255-276.

Narisetty, N., Sai, V. H., Siddiqa, S. S., Bedadhala, S. M., & Banala, S. (2023, December). Empowering Health Surveillance: A Machine Learning Based Risk Alert System. In 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) (pp. 1-6). IEEE.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2, 1-10.

Siegal, D., & Ruoff, G. (2015). Data as a catalyst for change: stories from the frontlines. Journal of Healthcare Risk Management, 34(3), 18-25.

Stahl, B., Häckel, B., Leuthe, D., & Ritter, C. (2023). Data or Business First?—Manufacturers' Transformation Toward Data-driven Business Models. Schmalenbach Journal of Business Research, 75(3), 303-343.

Voigt, K. I., Brechtel, F., Schmidt, M. C., & Veile, J. (2021). Industrial data-driven business models: towards a goods-service-data continuum. Digital Business Models in Industrial Ecosystems: Lessons Learned from Industry 4.0 Across Europe, 137-153.

Walker, D., Ruane, M., Bacardit, J., & Coleman, S. (2022). Insight from data analytics in a facilities management company. Quality and Reliability Engineering International, 38(3), 1416-1440.

Wisse, M., & Roeland, J. (2022). Building blocks for developing a research question: The ABC-model. Teaching Theology & Religion, 25(1), 22-34.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2005, June). Practical machine learning tools and techniques. In *Data mining* (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.

Yu, S. H., Su, E. C. Y., & Chen, Y. T. (2018). Data-driven approach to improving the risk assessment process of medical failures. International Journal of Environmental Research and Public Health, 15(10), 2069.

Zaranko, B., Sanford, N. J., Kelly, E., Rafferty, A. M., Bird, J., Mercuri, L., ... & Propper, C. (2023). Nurse staffing and inpatient mortality in the English National Health Service: a retrospective longitudinal study. BMJ quality & safety, 32(5), 254-263.

Zhang, Q., Pang, C., Mcbride, S., Hansen, D., Cheung, C., & Steyn, M. (2010, July). Towards health data stream analytics. In IEEE/ICME International conference on complex medical engineering (pp. 282-287). IEEE.

Zhu, X., & Sherali, H. D. (2009). Two-stage workforce planning under demand fluctuations and uncertainty. Journal of the Operational Research Society, 60(1), 94-103.

9 Appendix

9.1 Introduction

Research question	Type of research	Research subjects	Research strategy	Data gathering	Data analysis
What type of incidents have to be prevented at Trajectum?	Descriptive	management, staff, and patients of Trajectum	Qualitative	Observation, Expert Interviews	Summary
In what ways can data-driven working help the mentally handicapped-care?	Exploratory	Literature	Qualitative	Systematic literature review	Content analysis
In what ways can data-driven working prevent incidents?	Exploratory	Literature	Qualitative	Literature study	Ideation
How can we predict incidents at Trajectum?	Explanatory	Trajectum IT department, Patients of Trajectum.	Quantitative	Primary sources, Statistical analysis	Trend analysis, Linear regression (using R)

Figure 9.1; Research design

9.2 Analysis of incidents

	Question (English)	Question (Dutch)
1	What different types of incidents have you witnessed inside of the clinics?	Wat voor soort incidenten heb je meegemaakt binnen de klinieken?
2	What types of incidents occur most frequently?	Welke types incidenten komen het vaakst voor?
3	What types of incidents are most important to you to prevent?	Welke types incidenten vind je het belangrijkst om te voorkomen?
4	What type of incidents are not always reported?	Zijn er incidenten die genegeerd/niet gerapporteerd worden?
5	Under which circumstances could an incident occur?	Onder welke omstandigheden ontstaan incidenten (is het vaak voorspelbaar)?
6	How often do you feel like there is a shortage of staff?	Hoe vaak heb je het idee dat er te weinig personeel is voor de mix cliënten?
7	How many hours of labour are you generally missing a week?	Hoeveel werkuren komen jullie over het algemeen te kort per week?
8	Do you notice a relation between the number of incidents occurring and the present staff?	Merk je een verband tussen de hoeveelheid incidenten en (tekort aan) aanwezige personeel?
9	Do you notice a difference between full time staff and independent staff?	Merk je dit verschil ook tussen bijvoorbeeld vaste medewerkers en ZZP'ers?
10	Do you think that more present members of staff could prevent these incidents?	Heb je het gevoel dat de aanwezigheid van meer medewerkers incidenten voorkomt?
11	How do the patients move from clinic to clinic within Trajectum?	Hoe stromen cliënten door de verschillende afdelingen van Trajectum?
12	How do clinics communicate the information of each new patients and their hazards?	Hoe wordt de informatie van de cliënten door gecommuniceerd wanneer zij naar een andere kliniek verplaatsen?

Figure 9.2; Interview questions

9.3 Literature review

Key search term	Related term	Narrower term	Broader term
Data-driven	Data driven business model*, DDBM*, trend analysis	Data-mining, AI	Big data, predictive data, trend*
Staff allocation	Manpower planning, job scheduling	Stochastic scheduling model*, Patient safety; Probabilistic risk assessment	Resource allocation, schedul*, risk planning, risk management
Mentally handicapped-care	Sheltered housing, ICF/ID, ICF/MR,	Mild mental handicap, Intermediate Care Facilities for individuals with Intellectual disability	Healthcare, hospital*, intellectual disability

Figure 9.3; Key search terms

Inclusion criteria	Justification
Peer reviewed	Ensures reliable scientific contents. Only databases were used that exclusively hold peer review literature.
Full access through UT library	As access to other databases can be pricey, only databases were used that give full access to the literature through the UT subscription.
English publication	No translated works are included to prevent irrelevancy.
Exclusion criteria	**Justification**
For articles regarding data-engineering: older than 10 years	As data-engineering techniques can change/ develop quickly, records of the last 10 years are included (only applies to articles on e.g. data-mining techniques).
No recorded author	Seems implicit, but occasionally records were found from IEEE conferences that did not formally state any author, as it was part of a collaboration. These records were eventually excluded.
Surveys as centre part of study	As seen from the results from PubMed, many papers used surveys as main means of data-collection. As this research concerns itself with data-science on primary data, these records were excluded.

Figure 9.4; Inclusion and exclusion criteria

Database	Justification
Scopus	Multidisciplinary database with exclusively peer reviewed articles to ensure articles that are relevant but possibly not in the relevant scientific fields are included in the research. Access through UT find.
PubMed	Medical science database to find articles closely related to the healthcare aspects of the research. Contains healthcare specific articles that are not included in Scopus. Access through UT find.
IEEE	Database with technical literature to find articles closely related to the data-mining/engineering aspects of the research. Contains articles related to data-driven working that are not included in Scopus. Access through UT find. IEE also publishes literature themselves, which is not always available in other scientific databases.

Figure 9.5; Choice of database

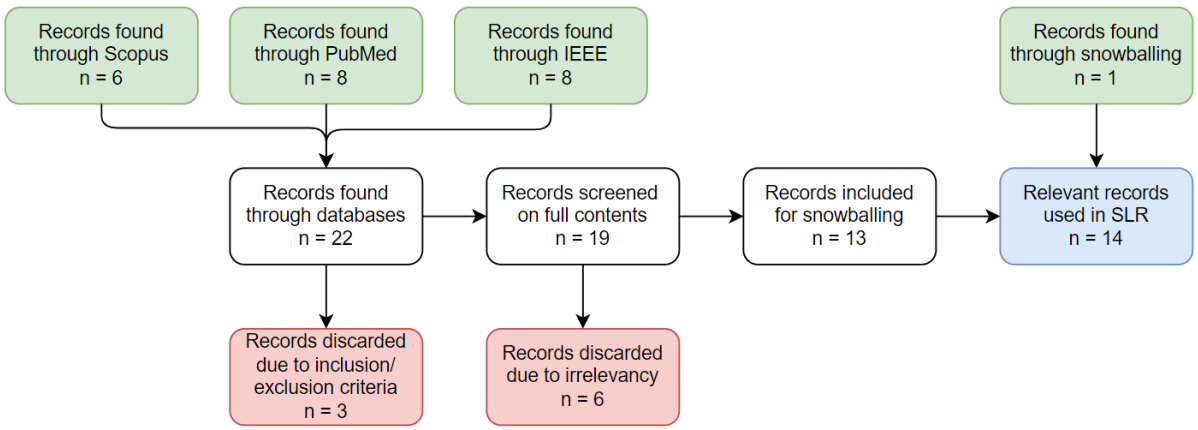

Figure 9.6; Flowchart of search process

Date	Database	Search query	Hits	Comment
17/04/2024	Scopus	TITLE-ABS-KEY (("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning") AND ("Mentally handicapped-care" OR "Sheltered housing" OR icf/id OR icf/mr))	0	This first search was too specific to my own case. From here I will look for relevant papers that include applications of data-driven techniques outside of the mentally handicapped-care, to see if I should broaden this search term.
17/04/2024	Scopus	TITLE-ABS-KEY (("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis" OR "big data" OR "predictive data" OR trend*) AND ("Staff allocation" OR "Manpower planning" OR "resource allocation"))	7,312	From this adjusted query, where broader terms were included for the first two key search terms, we get too many irrelevant articles (mostly to do with resource allocation). In response, the first two terms are narrowed again, and the last term (mentally handicapped-care) broadened
17/04/2024	Scopus	TITLE-ABS-KEY (("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning") AND ("Mentally handicapped-care" OR "Sheltered housing" OR healthcare OR hospital* OR "intellectual disability" OR icf*))	5	This was the first fruitful query, where relevant papers came up. Though, preferably a broader query should be used to ensure more hits.
17/04/2024	Scopus	TITLE-ABS-KEY (("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "risk management") AND ("Mentally handicapped-care" OR "Sheltered housing" OR healthcare OR hospital* OR "intellectual disability" OR icf*))	64	After broadening the key search term *staff allocation*, successfully including more articles in the search, various useful new papers were found. Still, various papers were found that were not performed in the medical sector. For this reason, we tried again in a medical database.

18/04/2024	PubMed	ALL(("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "risk management") AND ("Mentally handicapped-care" OR "Sheltered housing" OR healthcare OR hospital* OR "intellectual disability" OR icf/id OR icf/mr))	34	Though new useful articles were found, most results did not regard staff/resource allocation. Instead, a lot of articles regarding risk management came up that were irrelevant. In the next query, risk management is let out of the search terms, and the first key search term is broadened.
18/04/2024	PubMed	ALL(("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis" OR "Big data" OR "predictive data") AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning") AND ("Mentally handicapped-care" OR "Sheltered housing" OR healthcare OR hospital* OR "intellectual disability" OR icf/id OR icf/mr))	3	No new articles were found with this query. For the next search, I included more terms for the second key search term. The third key search term was redefined according to the commonly used terms in the library of PubMed (using “Intermediate Care Facilities for individuals with Intellectual disability” instead of “mentally handicapped-care”).
18/04/2024	PubMed	ALL(("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "Stochastic scheduling model*" OR "Patient safety" OR "Probabilistic risk assessment") AND ("Mentally handicapped-care" OR "Sheltered housing" OR hospital* OR "intellectual disability" OR "Intermediate Care Facilities for individuals with Intellectual disability"))	132	Many new useful articles came up after redefining and broadening terms. I decided to try another database that was more specific to data science, as many results were related to medicine, which has little relevance to our research.

18/04/2024	IEEE	ALL(("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "Stochastic scheduling model*" OR "Patient safety" OR "Probabilistic risk assessment") AND ("Mentally handicapped-care" OR "Sheltered housing" OR hospital* OR "intellectual disability" OR "Intermediate Care Facilities for individuals with Intellectual disability"))	8	Two new relevant articles were found, both published by IEEE, demonstrating the value of using this database. The number of hits is low, suggesting that more OR operators should be used to include articles in a broader scope. As the articles in the database of IEEE are engineering related, the third search term stays broad, as the relation with healthcare is not implicit, like it was with PubMed.
18/04/2024	IEEE	ALL(("Data-driven" OR "Data driven business model*" OR ddbm* OR “data*mining” OR ai OR “big data” OR “predictive data” OR “trend*”) AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "Stochastic scheduling model*" OR "Patient safety" OR "Probabilistic risk assessment") AND ("Mentally handicapped-care" OR "Sheltered housing" OR hospital* OR "intellectual disability" OR "Intermediate Care Facilities for individuals with Intellectual disability"))	74	This search resulted in various relevant articles. From here, I believed that I had accumulated enough relevant literature to, after snowballing, start thoroughly screening the articles. This was the last search in an academic database.

Figure 9.7; Search log

Method (Database/Query)	Article Title	Author(s)/year	Main points
Scopus / TITLE-ABS-KEY (("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning") AND ("Mentally handicapped-care" OR "Sheltered housing" OR healthcare OR hospital* OR "intellectual disability" OR icf*))	Insight from data analytics in a facilities management company.	Walker, D., Ruane, M., Bacardit, J., & Coleman, S. / 2022.	Explores the applications of data-science. One of them being a tool for job scheduling (outside of the field of healthcare).
Scopus / TITLE-ABS-KEY (("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "risk management") AND ("Mentally handicapped-care" OR "Sheltered housing" OR healthcare OR hospital* OR "intellectual disability" OR icf*)	A data-driven and practice-based approach to identify risk factors associated with hospital-acquired falls: Applying manual and semi-and fully-automated methods.	Lucero, R. J., Lindberg, D. S., Fehlberg, E. A., Bjarnadottir, R. I., Li, Y., Cimiotti, J. P., & Prosperi, M. / 2019.	Risk-management study in hospital. Prediction model to assess risk factors within patient population.
	Data-driven approach to improving the risk assessment process of medical failures.	Yu, S. H., Su, E. C. Y., & Chen, Y. T. / 2018.	Risk assessment for medical failures using data envelopment analysis (DEA).
	Modelplasticity and abductive decision making	Mukhopadhyay, S. / 2023.	Overview of data-driven decision making. (including healthcare)
	Data as a catalyst for change: stories from the frontlines.	Siegal, D., & Ruoff, G. / 2015.	Various successful applications of data-science for the reduction of risk in healthcare.

PubMed / ALL(("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "risk management") AND ("Mentally handicapped-care" OR "Sheltered housing" OR healthcare OR hospital* OR "intellectual disability" OR icf/id OR icf/mr))	Factors Related to Medication Administration Incidents in England and Wales Between 2007 and 2016: A Retrospective Trend Analysis	Härkänen, M., Vehviläinen-Julkunen, K., Franklin, B. D., Murrells, T., & Rafferty, A. M. / 2021.	Applications of trend analysis in medication incidents. Aims to find relations between patients' factors and reported severity of incidents.
	Strategic analytics: towards fully embedding evidence in healthcare decision-making.	Garay, J., Cartagena, R., Esensoy, A. V., Handa, K., Kane, E., Kaw, N., & Sadat, S. / 2015.	Data-analysis that supports decisions in healthcare making through predictive analytics solutions.
PubMed / ALL(("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "Stochastic scheduling model*" OR "Patient safety" OR "Probabilistic risk assessment") AND ("Mentally handicapped-care" OR "Sheltered housing" OR hospital* OR "intellectual disability" OR "Intermediate Care Facilities for individuals with Intellectual disability"))	Advanced Data Analytics for Improved Decision-Making at a Veterans Affairs Medical Center.	Mahajan, A., Madhani, P., Chitikeshi, S., Selvaganesan, P., Russell, A., & Mahajan, P. / 2019.	Data-driven methodology for decision making in healthcare. Background information regarding data-science applications.
	Mining routinely collected acute data to reveal non-linear relationships between nurse staffing levels and outcomes.	Leary, A., Cook, R., Jones, S., Smith, J., Gough, M., Maxwell, E., Punshon, G., & Radford, M. / 2016.	Data-driven technique to find relations between staffing levels and patient outcomes, such as safety factors and physiological data. Explores possible applications of big data.
Forward snowballing from Leary et al. (2016)	Nurse staffing and inpatient mortality in the English National Health Service: a retrospective longitudinal study.	Zaranko, B., Sanford, N. J., Kelly, E., Rafferty, A. M., Bird, J., Mercuri, L., ... & Propper, C. / 2023.	Explores the relation between nurse staffing level and mortality rate in hospitals. Statistical analysis, descriptive results.

IEEE / ALL(("Data-driven" OR "Data driven business model*" OR ddbm* OR "trend analysis") AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "Stochastic scheduling model*" OR "Patient safety" OR "Probabilistic risk assessment") AND ("Mentally handicapped-care" OR "Sheltered housing" OR hospital* OR "intellectual disability" OR "Intermediate Care Facilities for individuals with Intellectual disability"))	Towards health data stream analytics.	Zhang, Q., Pang, C., Mcbride, S., Hansen, D., Cheung, C., & Steyn, M. / 2010.	Explores the need for proper continuously growing data-sets in healthcare. Inner workings of possible trend-analysis tools.
IEEE / ALL(("Data-driven" OR "Data driven business model*" OR ddbm* OR “data*mining” OR ai OR “big data” OR “predictive data” OR “trend*”) AND ("Staff allocation" OR "Manpower planning" OR "job scheduling" OR "Resource allocation schedul*" OR "risk planning" OR "Stochastic scheduling model*" OR "Patient safety" OR "Probabilistic risk assessment") AND ("Mentally handicapped-care" OR "Sheltered housing" OR hospital* OR "intellectual disability" OR "Intermediate Care	Dynamic changes by big data in health care.	Balaji, S., & Prasathkumar, V. / 2020.	Applications of big data in healthcare. Machine learning in healthcare.
	Promoting Patient Safety through Machine Learning.	Al Nammari, R. H. / 2020.	Explores the use of machine learning in healthcare to reduce medical errors.
	Empowering Health Surveillance: A Machine Learning Based Risk Alert System.	Narisetty, N., Sai, V. H., Siddiqa, S. S., Bedadhala, S. M., & Banala, S. / 2023.	Method to use machine learning to make assessments on risk of the development of diseases among patients.

Figure 9.8; Overview used articles SLR

Concept / Article	Specific to healthcare	Trend analysis	Big data / ML	Risk	Staff allocation
Al Nammari (2020)	X		X		
Balaji et al. (2020)	X		X		
Garay et al. (2015)	X	X			
Härkänen et al. (2021)	X	X			
Leary et al. (2016)	X	X	X		X
Lucero et al. (2019)	X			X	
Mahajan et al. (2019)	X	X			
Mukhopadhyay (2023)		X	X		
Narisetty et al. (2023)	X		X	X	
Siegal et al. (2015)	X	X		X	
Walker et al. (2022)		X			X
Yu et al. (2018)	X	X		X	
Zaranko et al. (2023)	X	X			X
Zhang et al. (2010)	X	X	X		

Figure 9.9; Conceptual matrix

9.4 Model ideation

Abbreviation	Variable type	Effect on relationship	Action needed
DV	Dependent	Variable of interest, this research aims to reduce this variable.	Measure
IV	Independent	Believed to have significant effect on the DV.	Manipulate
IVV	Intervening	IVs effect is believed to be trans-mitted through here to affect DV.	Measure
CV	Control	Might have an effect on DV, but outside of our control.	Ignore
MV	Moderating	Possible contributory effect on relation between IV and DV.	Assess effect from IV and MV on DV
CFV	Confounding	Unknown effect on the relation between IV and DV.	Measure/discuss

Figure 9.10; Variables of the model in Figure 4.1

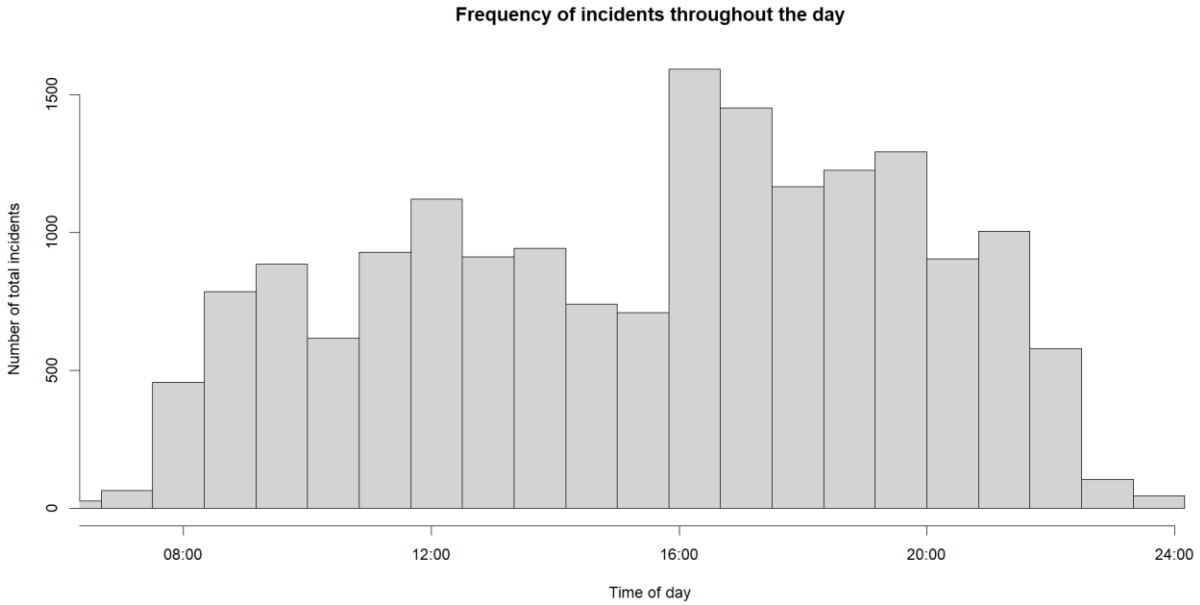

Figure 9.11; Frequency of incidents based on time of day

9.5 Prediction tool

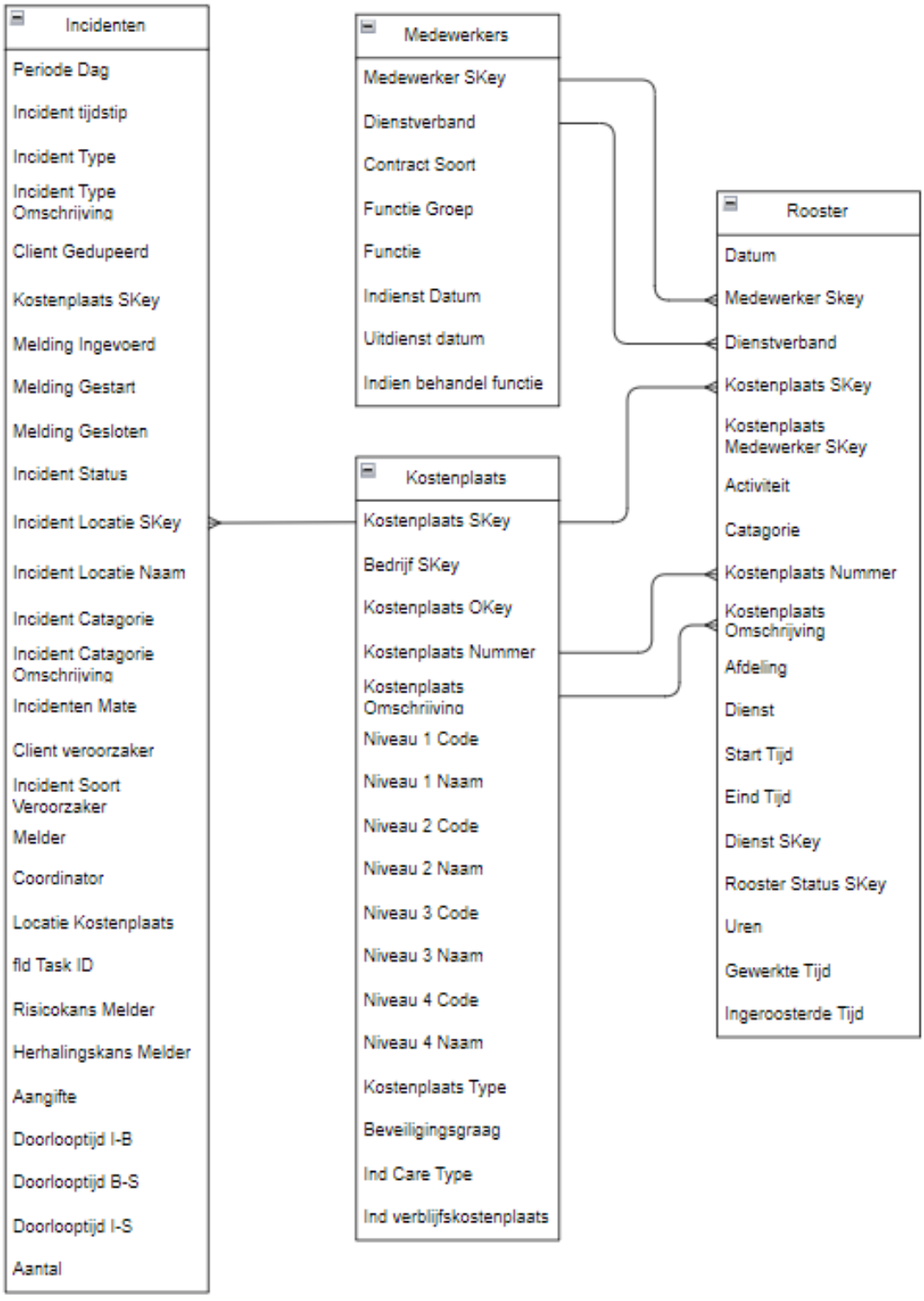

Figure 9.12; Relational model of the data provided by Trajectum

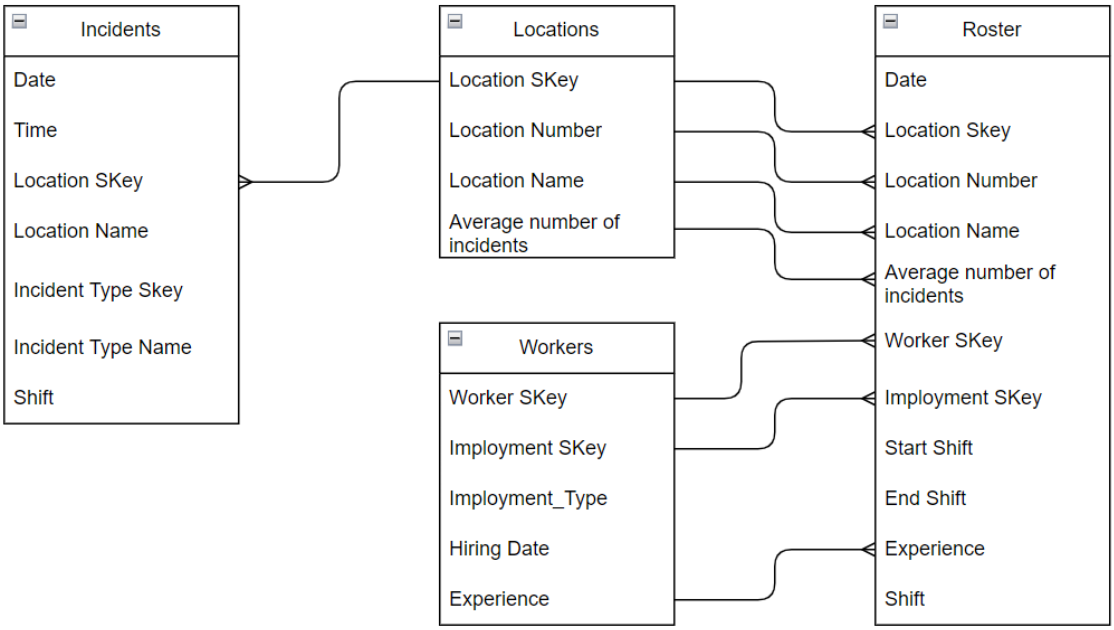

Figure 9.13; Relational model of the structured data

Variable	Type	Description	Purpose
Date	yyyy-mm-dd, Range: 2022-01-01 to 2024-05-28.	Date of the shift.	These variables together uniquely identify any shift at any location of Trajectum where an incident could occur.
Shift	Integer, Range: 1 (morning-shift) to 3 (night-shift).	Type of shift; Morning-shift (start between 06:15 and 14:00), evening-shift (start between 14:00 and 20:00), or night-shift (start between 20:00 and 06:15).	
Location	Integer, Range: 23 to 249.	Number that refers to a unique location of Trajectum where an incident could occur (excludes e.g. office location).	
Average Number of Incidents	Float, Range: 0 to 0.5	Represents the past frequency of incidents at a particular location. The value is the average number of incidents per shift.	Independent variable, meaning that as it is manipulated, we expect the dependent variable (number of incidents) to change accordingly. We expect a negative, negative, and positive relation respectively.
Average Experience	Float, Range: 0 to 38.7.	Average years of experience at Trajectum of all staff members present during specific shifts. Based on the hiring date.	
Worker Ratio	Float, Range: 0 to 1	Represents the share of permanently employed workers among all workers present at any shift.	
Recent Incident	Integer, Range: 0 or 3	Categorical value that represents the time since an incident at that location. 0: Longer than 7 days ago, 1: Within 7 days ago, 2: Within 3 days ago, 3: Within 1 day ago.	Confounding variable, meaning we expect it to influence the strength of the relation between the IVs “Average Experience” and the “Worker Ratio”. We expect the relations to become stronger as “Recent incident” increases.
Number of Incidents	Integer, Range: 0 to 10.	Number of (relevant) incidents that occurred during the specific shift. Relevance is determined in Section 2.3.	Dependent variable, meaning that we aim to predict the value of this variable, based on the values of the independent and the moderating variable.

Figure 9.14; Variables of the final data table

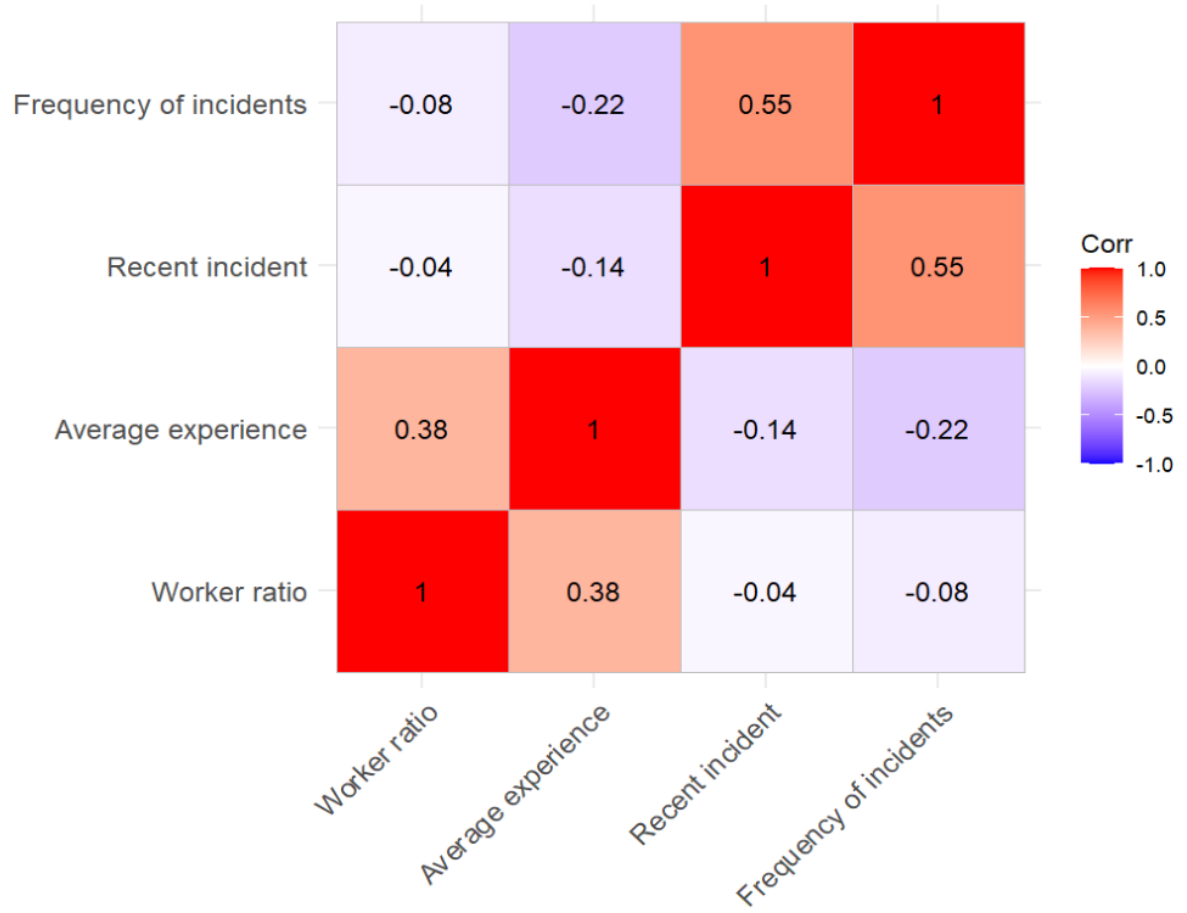

Figure 9.15; Correlation matrix between potential independent variables

	Dependent variable:
	Recent incident coefficient
Constant	0.001
	(0.001)
`Frequency of incidents`	-0.045***
	(0.009)
Observations	81
Residual Std. Error	0.008 (df = 79)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 9.16; Average Experience coefficient

	Dependent variable:
	Recent incident coefficient
Constant	0.001
	(0.001)
`Frequency of incidents`	0.085***
	(0.009)
Observations	81
Residual Std. Error	0.008 (df = 79)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 9.17; Recent Incident coefficient