universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Modelling path-dependent diffusion processes on complex healthcare networks“

verfasst von / submitted by

Michaela Kaleta, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 876
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Physik
Betreut von / Supervisor:	Assoc. Prof. Priv.-Doz. Mag. Dr. Peter Klimek
Mitbetreut von / Co-Supervisor:	Univ.-Prof. Mag. Dr. Christoph Dellago

Contents

Abstract

As the amount of data for the study of complex networks is continuously increasing, many recent studies tried to understand the dynamics of physical, biological or socio-economic systems by means of diffusion processes on empirically measured real-life networks. For many such systems, diffusion as a transport process that seeks to balance a concentration gradient between regions of higher density and lower density as postulated by A. Fick is a model too simplistic to describe transport processes on real-world complex networks. Generalized random walk models (e.g. Katz Prestige, PageRank, ...) often provide a more adequate framework to study such systems. In this work, we will use a random walk model given by the graph Laplacian matrix and statistical tools of logistic regression analysis to study the dynamics of a real-life system, the exchange of patients between different types of healthcare provider in Lower Austria. Using computational means, we investigate the path dependence of patient movements and discuss why the process of patients contacting different specialists on such networks does not satisfy the Markovian property. In addition, we search the system for statistically overrepresented network motifs which are based on connections between specific types of hospital visits of patients and their related contacts to medical specialists. We find a substantial number of such motifs that can be associated with strongly reduced probabilities of re-hospitalization (up to 50% risk reduction for certain diagnoses) if patients have contacts with certain specialists. We discuss sex-specific biases in these treatment paths and we find evidence that the healthcare system tends to amplify existing sex biases in the sense that males (females) typically show greater re-hospitalization risk reduction for male- (female-) dominated diagnoses. Our results quantify for the first time sex-specific re-hospitalization risks in treatment paths and might help in identifying leverage points to improve patient flows in regional healthcare systems.

Da die Menge an verfügbaren Daten für die Analyse von komplexen Systemen kontinuierlich steigt, haben sich viele der neuesten Studien um ein besseres Verständnis der Dynamik in physikalischen, biologischen oder sozio-ökonomischen Systemen als Diffusionsprozess in empirisch beschriebenen reellen Netzwerken bemüht. Für viele solcher Systeme ist Diffusion als Transportprozess, der, so wie von A. Fick postuliert, das Gleichgewicht des Konzentrationsgradienten zwischen Regionen mit höherer Dichte zu niedrigerer Dichte anstrebt, ein zu einfaches Modell um Diffusion in echten komplexen Netzwerken zu beschreiben. Verallgemeinerte Random Walk Modelle (wie z.B. Katz Prestige, PageRank, ...) bieten oft eine bessere Grundlage um solche Systeme zu studieren. In dieser Arbeit verwenden wir ein Random Walk Modell, das durch die Laplace Matrix gegeben ist, und grundlegende Methoden der Statistik wie die logistische Regression um die Dynamik eines echten Systems zu untersuchen, konkret den Austausch von Patienten zwischen verschiedenen Anbietern im Gesundheitswesen der Region Niederösterreich. Mit Hilfe computergestützter Analysen untersuchen wir die Pfadabhängigkeit der Patientenbewegungen und erörtern, warum die Bewegung von Patienten zu Ärzten nicht als ein Markov-Prozess angesehen werden kann. Zusätzlich suchen wir das System nach statistisch überrepräsentierten Netzwerkmotiven ab, welche als Verbindungen zwischen spezifischen Spitalsbesuchen und zugehörigen Ärztekontakten angesehen werden können. Wir stellen fest, dass eine beträchtliche Anzahl solcher Motive mit dem Fakt in Zusammenhang gebracht werden kann, dass das Rehospitalisierungsrisiko stark reduziert wird wenn Patienten Kontakt zu bestimmten Spezialisten hatten (bei einigen Diagnosen bis zu 50% Risikosenkung). Wir erörtern geschlechtsspezifische Tendenzen in diesen Behandlungspfaden und wir folgern, dass das Gesundheitssystem dazu tendiert bereits existierende Geschlechtertrends zu verstärken. Männer (Frauen) zeigen typischerweise genau bei solchen Diagnosen eine größere Risikosenkung, die männer-(frauen-) dominiert sind. Unsere Ergebnisse zeigen zum ersten Mal geschlechtsspezifische Rehospitalisierungsrisiken in Behandlungspfaden und könnten dabei helfen Anhaltspunkte zu finden um Patientenflüsse in regionalen Gesundheitssystemen zu verbessern.

Chapter 1

Introduction

The random movement of particles, such as Brownian motion, is an elementary stochastic process closely connected to the phenomenon of diffusion. The general concept of random motion is used in a vast number of scientific fields like biology [11], chemistry [25] and physics [53]. In a time-varying system, diffusion processes can be interpreted in terms of the probabilities to localize so called random walkers at specific sites in the system. There exists a certain probability to move from one point of the system to another. The decision of a random walker of which step to take is a random process. This kind of diffusion model is often used in the study of complex systems [31]. From a physics point of view, the theory of complex systems can be defined as a quantitative and predictive description of generalized interactions on systems, that can be experimentally tested [52]. Such systems surround our everyday life and have only recently been started to be analysed in a quantitative manner [1]. Using different methods and approaches reaching from physics to social sciences and statistics, the aim of the emerging field of complexity science is to describe properties of large systems that consist of elements i in states $\sigma_i(t)$ that are connected to other elements j of the system through interactions of strength $M_{i,j}(t)$. The states and interactions in complex systems are dynamic and vary over time t like $\frac{d\sigma_i(t)}{dt} \sim F\big(M_{i,j}(t), \sigma_i(i)\big)$, with a function F of present states and present interaction strengths, and $\frac{dM_{i,j}(t)}{dt} \sim G\big(M_{i,j}(t), \sigma_i(i)\big)$, with another function G of the present system states. Note that $\frac{d}{dt}$ does not need to be a differential operator [52]. Applications of this approach reach from the study of innovation diffusion [43] and spread of behaviour in social networks [9] to epidemic spreading in real-world systems [39]. A typical structure within the study of complex systems consists of a dynamical set of nodes and various links connecting them. Diffusion processes can then be modelled as a random walk between nodes using the specific structure and rules connecting the system.

We will be interested in the probabilities of particles, which perform a type of random motion, to be located at specific positions in the network — a diffusion model. In a complex system this type of diffusion process can sometimes be identified with a random walk that is mathematically described by the graph Laplacian Λ. This Laplacian matrix is correlated to the network's adjacency matrix A and leads to a diffusion equation similar to Fick's law. If the diffusion is memoryless, it satisfies the so called Markovian property, meaning that predictions about future states are based exclusively on present states [23] [52]. This

kind of dynamics in complex systems can be efficiently studied by computational means. In particular, different iterative centrality measures can be used to describe diffusion on networks, such as eigenvector centrality or PageRank.

In this work we study diffusion in the interdisciplinary application area of networks describing the clinical history of patients. As was pointed out in the paper from A. Miles [32], the science of complex systems might provide a better understanding of complicated interactions in healthcare systems and cope with difficulties that arise when the individual's conditions need to be considered under the influences of many processes. We therefore make an effort to apply structures known mostly from the fields of physics and complexity science to real-life healthcare data. The nodes of the so obtained network are different healthcare provider (HCP) of the Austrian healthcare system, which are connected by links if patients are exchanged within a certain timespan. These exchanges can be interpreted as a flux between the providers. We can use the information of patients' doctor contacts to create transition probability matrices. Based on the theory of diffusion on networks one can then predict patients' paths in the HCP system. Depending on the number of given previous states, we can analyse the extent to which the underlying diffusion process adheres to the Markovian property. This allows us to make statements about the intrinsic memory(lessness) of a real system.

We are also interested in sex-specific differences of the diffusion of patients, as there exists evidence that females are less likely to diffuse to specialized care settings (e.g. medical specialists, hospitals or high-end medicine [3]), in some cases sex-specific effects can be found for certain diagnoses like cardiovascular diseases [35] or diabetes mellitus [24]. Using a medical claims dataset of about 1.7 million patients of Lower Austria, it is possible to test this hypothesis in terms of different diffusive properties for females and males and to determine their probability density in care networks as well as to examine the memorylessness of the diffusion processes. In addition to the studies of random walks on complex networks, we examine the re-hospitalization risk of patients using logistic regression analysis. We define specific re-hospitalizations as network motifs and test the extent to which they are overrepresented in the system. By identifying outcome (with re-hospitalization) and control (without re-hospitalization) samples in the claims set, we make statements about the sex-specific effects on risk increase of re-hospitalizations due to different contacts that correspond to site visits in the network.

This work is divided up into the following sections: In chapter two, we introduce the theoretical background necessary to understand the connection between diffusion processes known from a classical physics point of view and the process of anomalous diffusion and random walks on networks. We provide an overview of the basic concepts of graph theory and network analysis as well as some general statistical tools used in this work. In the third chapter, we give a detailed description of the data processing and introduce the methodological implementation of the random walk algorithm and logistic regression analysis. In chapter four, we present the main results of the network memory analysis and the network motif analysis as well as a visualization of the underlying system. We discuss the implications of the results together with the limitations of our methods in chapter five.

Chapter 2

Theory and concepts: Random walks on networks

In this chapter we provide an overview on the theoretical concepts used in this thesis. Starting with general definitions of networks and their structural features, we proceed to the concept of network motifs and the statistical tool of logistic regression used to estimate the binary outcome of the re-hospitalization model. We further describe concepts of stochastic processes, in particular the so called Markov processes. Concluding with a brief history of the development of diffusion we create a connection between diffusion typically known in physics and diffusion on networks using the graph Laplacian and random walk theory.

2.1 Networks

Networks, or graphs, find use in many different disciplines [1]. The mathematics of graph theory can be used to make statements about interactions between objects. While in the past focus was set mostly on the properties of single nodes or small structures, latest research is targeting the larger statistical qualities of networks. Many systems in the real world can be represented by complex networks and make them ideal for studying the systems' internal relations [48] [37].
In general, graphs consist of a set of members, called vertices or *nodes*, that are connected through links or *edges*. We will denote a graph with $G = G(N, E)$ where N is the set of nodes and E is the set of edges of the underlying graph. The edges and nodes themselves can have various properties defining the structure and characteristics of the whole network [37]. In the following, we will briefly describe the most important properties.

2.1.1 Nodes and edges

Nodes (sometimes called vertices or also sites) are one of the fundamental building blocks of networks. They can represent a large variety of states or subjects. Just to name a few examples, nodes can stand for micro-grids in the electric power system [44], proteins in human cells [2] or people in a social group [18]. A network can consist of one or more types

of nodes (e.g. as in bipartite networks). Each node can be further characterized by certain properties such as degree or strength.
Edges (or links, bonds) are connections between nodes or in special cases, a connection of a node to itself (self-loops). They describe whether there are interactions in the system. An edge can represent any kind of relationship, a few examples would be ties in a social network like friendships and other relationships or cash flow in markets. Nodes of a network can be connected through more than one type of edge, creating so called multiplex networks [52]. Also, an edge may be directed in the form $X \to Y$ or undirected $X - Y$ or carry weights depending on some chosen characteristic.

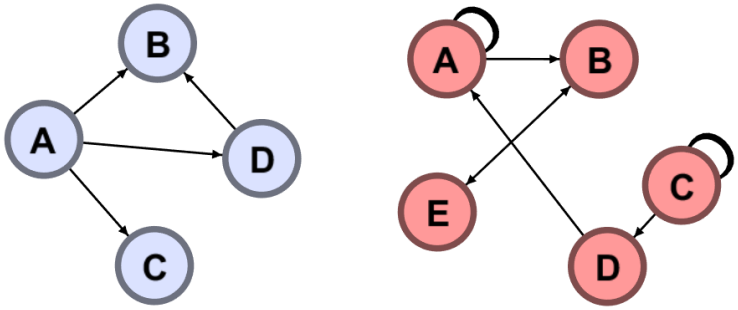


Figure 2.1: Examples of small networks. Left: Simple directed network. Nodes A, B, C and D are connected through directed edges. Right: Directed network of five nodes with self-loops on nodes A and C.

2.1.2 Directed and undirected networks

Depending on the edges, networks can either be directed or undirected. In an undirected network, an edge between two nodes does not specify any direction of interaction. An example could be a tie in a social network, where no favourable direction can be identified. In contrast to that are directed networks, where an edge provides a defined direction of interaction. We distinguish between incoming and outgoing edges for every node. An example of small directed networks is shown in figure 2.1, simple graphs are defined with no self-loops. As a real-life example, a directed network could describe the cash flow in a banking system, where the direction states an important feature about the connection. A graph theoretical description of the structure of networks is provided by adjacency matrices [52].

2.1.3 Adjacency matrix

Adjacency matrices are used as matrix representations of graphs. Even though it might be useful to see a network depiction like in figure 2.1 in some cases, for larger networks and for calculations this representation is not convenient any more. Adjacency describes a property of edges and nodes in a network. In case of edges, adjacency is achieved if two edges share a common node. On the other hand, two nodes are called adjacent, if there is an edge connecting them [54].

In the following, we will denote the adjacency matrix with A. Starting with an undirected simple network with n nodes, A is a $n \times n$ symmetric matrix. The entries of the matrix A_{ij} are defined [36]

$$A_{ij} = \begin{cases} 1, & \text{if nodes } j \text{ and } i \text{ are adjacent,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

This is a simple network without self-interaction and the entries in the diagonal of A are all zero, therefore also the trace of A is zero [4] [5]. This would not be true if the network contained multi-edges or self-loops. In the case of weighted graphs, where an edge carries a specified value, the entries of the adjacency matrix describe the strength of the connection between nodes.

In directed networks, edges point from one node to another and are depicted with arrows in the given direction of the interactions. The entries of the adjacency matrix then change to [36]

$$A_{ij} = \begin{cases} 1, & \text{if an edge points from node } j \text{ to node } i, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

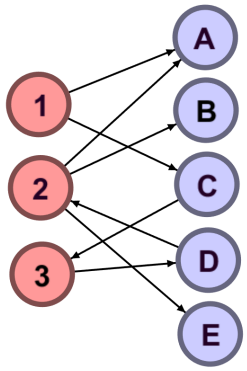
For directed networks, the adjacency matrix is in general not symmetric. As for the undirected graphs, also directed graphs can have multi-edges or self-loops, changing the diagonal entries to non-zero values [36]. As an example, we can write the adjacency matrices of the graphs in figure 2.1 like

$$A_{left} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \qquad A_{right} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Another instance we want to discuss here are *bipartite networks*. Bipartite networks consist of two types of nodes. Edges can only connect nodes of different type. Since the number of nodes per type does not need to be equal, the matrix describing the network is in general not quadratic. Assume the number of nodes of type one is n, the number of nodes of type two is m. We can create the so called incidence matrix B, which is a $n \times m$ matrix. The entries B_{ij} are calculated as [36]

$$B_{ij} = \begin{cases} 1, & \text{if node } j \text{ of type one is connected to node } i \text{ of type two,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

It is possible to write the adjacency matrix of a bipartite network by using the incidence matrix B and the transposed variant B^{T}. The result is a matrix of blocked form, where only off-diagonal blocks contain non-zero elements [52]. A small example graph of a directed bipartite network and the corresponding adjacency matrix can be seen below.



$$A = \begin{bmatrix} 0_{nxn} & B \\ B^{\mathrm{T}} & 0_{mxm} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

2.1.4 Degree and strength

We define the degree of a node as the sum of all edges that are attached to it. If the degree of a node is equal zero, it is called isolated, if the degree is one, it is an end-node [54] [5]. We can use the adjacency matrix to determine the degree k_i for the $i = 1, \ldots, n$ nodes of an undirected network [52],

$$k_i = \sum_{j=1}^{n} A_{ij} \quad . \tag{2.4}$$

In contrast to that, nodes in directed networks have two different types of degree, called *in-degree* and *out-degree*. Since we have in- and outgoing edges in such graphs, the degree of a node can be divided into the number of edges that point to or from the chosen node. Using the definition from (2.2), we can write the in-degree k_i^{in} and out-degree k_i^{out} of node i as [36]

$$k_i^{in} = \sum_{j=1}^{n} A_{ij} \quad , \quad k_i^{out} = \sum_{i=1}^{n} A_{ij} \quad . \tag{2.5}$$

These two distinct degrees can be combined which results in the degree k_i of directed networks: $k_i = k_i^{in} + k_i^{out}$. Usually, the degrees can be calculated by summing over the rows or columns of the adjacency matrix of the network. While in undirected networks the sum over rows and columns are equal, in the directed case the sums are different and the direction of interaction must be respected. Another exception are weighted networks. Here, summing over rows and columns does not yield the degree any more. The entries in the adjacency matrices of such networks no longer describe the number of edges attached, but rather another property of the connections, the so called *strength*. For undirected networks, the strength s_i of a node i is written as [52]

$$s_i = \sum_{j=1}^{n} A_{ij} \quad . \tag{2.6}$$

As for the degree, also the strength varies for directed networks and can be divided into *in-strength* s_i^{in} and *out-strength* s_i^{out} of node i as [52]

$$s_i^{in} = \sum_{j=1}^{n} A_{ij} \quad , \quad s_i^{out} = \sum_{i=1}^{n} A_{ij} \quad . \tag{2.7}$$

2.1.5 Functional networks and motifs

We have already described a few types of networks that depend on some characteristics a network possesses or not (like a direction of edges or weights). There exists another possibility of differentiation that is based on what a link encodes. It divides networks into *directly observable networks* and *functional networks* [52]. We can find networks with edges that represent evident connections of physical and also relational nature. An example would be the subway net with connecting rails or the world wide web with links relating websites to each other. As these edges show an obvious connection, such networks are called directly observable. In contrast to that, there are networks with nodes connected through correlations. In this case we use the term functional network. Here, links describe a process connecting nodes in a way that shows significant correlation of some kind. But despite multiple hypothesis tests, there is always the possibility of false discoveries due to unknown dependencies on variables or simply lack of data [52].

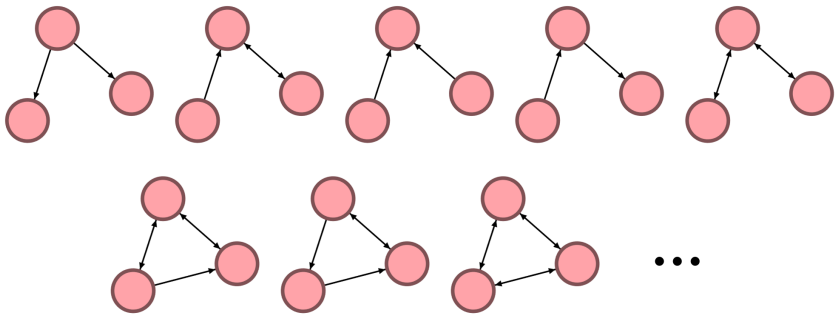


Figure 2.2: Example of network motifs M based on directed edges between three nodes (also called triads). The labelling of nodes is neglected.

For a deeper study of network structures a concept called *network motifs* was introduced by Alon *et al.* in 2002 [33] [47]. By definition, motifs are connection patterns observed in networks appearing with frequencies significantly higher than in random networks (like a Erdös-Rényi network). The concept of motifs was examined in several fields like biology, chemistry or sociology and it was found that some motifs are shared throughout different systems. A more graph theoretical description of connection patterns is provided by subgraphs. Graphs H are called subgraphs if $N(H) \subseteq N(G)$ and $E(H) \subseteq E(G)$ [10]. Motifs, denoted M, are subgraphs of an underlying graph G, so $M \subseteq G$ [33]. They can consist of an arbitrary number of nodes. Figure 2.2 shows some examples for motifs in a three node structure. Notice that the nodes are unlabelled, only the configuration of edges is of importance [52]. Often, the statistical significance of a motif is tested with Z-score analysis [7]. In this work we choose not to use the Z-score as a measure of significance, but rather other statistical methods to show the existence of motifs in the networks, such as odds ratios and logistic regression analysis.

2.1.6 Statistical methods

To find and further analyse network motifs, we will use some simple statistical tools. In most cases, frequencies of occurrences of motifs are counted. This can then be used to calculate odds ratios, outcome probabilities and finally also predict the outcome of a desired binary variable based on generated regression models.

Odds ratios

Odds ratios, or odds, have become a typical tool in statistics, especially when it comes to so called 'case–control studies' in medicine [40]. Usually, we have a binary outcome variable, that can take two possible states, yes or no [51]. Using the frequencies of those two states combined with some potential characteristics, one can create a frequency table as in 2.1. It shows the number of outcomes (or cases) and controls.

	outcome yes	outcome no
characteristics yes	a	b
characteristics no	c	d

Table 2.1: Frequency table of a typical case–control study.

Here, a is the number of outcomes with the given characteristic, b is the number of controls with characteristic, c the number of outcomes without characteristic and finally d is the number of controls without characteristic. We can use the probability of outcome with and without the characteristics to write the relative risk R as [49]

$$R = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \quad . \tag{2.8}$$

The odds represent another ratio of the above numbers. The odds ratio OR is calculated like [51]

$$OR = \frac{a/c}{b/d} \quad . \tag{2.9}$$

Odds describe the ratio of the probability that an outcome occurs with the chosen characteristic to the probability that the outcome does not occur with the characteristic [6]. The odds ratio has a lower limit at zero but has no upper bound. An odds ratio of one means equal probabilities of outcome and no outcome. OR-values smaller than one stand for lower probability that an outcome occurs with characteristic, while values larger than one mean higher probability for an outcome with characteristic.

Another possible way of characterizing the same result as the odds are logarithmic odds ratios, or log-odds $log(OR)$. The logarithm is simply applied to the odds and we obtain [6]

$$log(OR) = log\Big(\frac{a/c}{b/d}\Big) \quad . \tag{2.10}$$

The log-odds have no lower or upper bound and range from $-\infty$ to $+\infty$. One can calculate the standard error $SE_{logodds}$ of the log-odds. It can be further used to calculate confidence intervals and can be written like [6]

$$SE_{logodds} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad . \tag{2.11}$$

As an important fact for this work, log-odds can be associated to logistic regression analysis. By using the exponential function on the coefficients of the regression model one obtains the odds with respect to a unit-increase of the chosen characteristic [51] [34].

Generalized linear models

For modelling data we have to find a model suitable for the given dataset and the required purpose. In case of dichotomous data, binomial distributions provide a good solution. Binomial distributions can be modelled by logistic regression, a widely used form of generalized linear models [29]. In this case, the expected outcome of the response Z is binomial and can only take two values. We write Z as a linear combination of predictor variables $x_1, x_2, \ldots, x_l$, where the predictor variables x_j are independent and have corresponding coefficients β_j with $j = 1, \ldots, l$ like [29]

$$Z = \alpha \ + \ \beta_1 \ x_1 \ + \ldots \ + \ \beta_l \ x_l = \alpha \ + \ \sum_j \beta_j \ x_j \quad . \tag{2.12}$$

As mentioned in the previous section, the coefficients of a logistic regression can be associated with the increase of the log-odds of outcome Z in units of the predictor variables x_j [51]; α is the intercept. We want the probability for an event to vary monotonically with the predictor variables x_j and we therefore use a logistic function that transforms the values accordingly. We obtain the logistic probability $P(Z)$ [13]

$$P(Z) = \frac{exp(Z)}{1 \ + \ exp(Z)} = \frac{\exp(\alpha \ + \ \sum_j \beta_j x_j)}{1 \ + \ exp(\alpha \ + \ \sum_j \beta_j x_j)} \quad . \tag{2.13}$$

$P(Z)$ has a range of $[0, 1]$. Sometimes, the analysis is not based on probabilities but on the previously mentioned odds. In the logistic case, the odds can be written as [13]

$$odds \ Z \ = \ exp(Z) \ = \ exp(\alpha \ + \ \sum_j \beta_j x_j) \quad . \tag{2.14}$$

Again, the odds range from zero to $+\infty$. By applying the logarithm to the odds, one obtains the so called *logit*. The logit of the logistic probability $P(Z)$ is then [13] [21]

$$logit \ P(Z) = log\Big(\frac{P(Z)}{1 \ - \ P(Z)}\Big) = Z = \alpha \ + \ \sum_j \beta_j x_j \quad . \tag{2.15}$$

We use the logit function as the link function in the computational logistic regression model. It will change the range of the probability $P(Z)$ from $[0, 1]$ to the real numbers $\mathbb{R}$. Link functions are used in generalized linear models to create a connection between the linear predictors x_j β_j and the expected values of the response. The link function needs to

be differentiable and monotonously increasing. The logistic probability function forms the inverse to the logit and applying the two functions successively yields the response Z [42]. By using equation (2.13), the probability $P(Z)$ lies within $[0, 1]$. This is the desired outcome of the logistic regression analysis. The generated regression model then provides the corresponding coefficients α and β_j to the variables. We use this to predict the response for a binomial outcome variable for given coefficients α and β_j and for some specified values of predictor variables x_j.

2.2 Stochastic processes

As the word stochastic already suggests, stochastic processes are random processes and are usually formed by a set of random variables X_t selected by random trials. Those variables can take defined values x_t within the state space over time. In case of processes with discrete time steps, the variables X_t provide insight about the random distribution of the state space of the process at a selected time t [20]. It is important to mention, that stochastic processes can posses the property of a *memory*. Some processes, like the Bernoulli process, have no memory, while so called 'path-dependent' processes have memory of previous states of the system [52]. In the following sections we describe a specific type of stochastic process, the *Markov process*, named after the mathematician A. Markov [28]. A depiction of such a Markov process can be seen in figure 2.3.

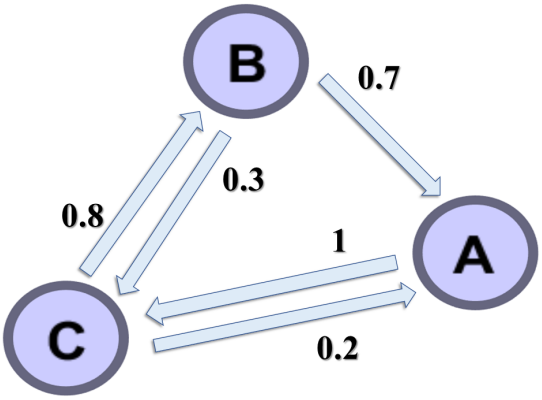


Figure 2.3: Example of a Markov process with three possible states A, B, C. Arrows with numbers stand for the transition probability between states. In sum, the transition probability for each state must be one.

2.2.1 Markov property

The Markov property is a requirement of Markov processes. It states that the future behaviour of a process must be independent of all past states. It is sufficient to know the current state of the system to fully characterize it. In mathematical terms, the Markov property in the discrete time case can be written as [15]

$$P(X_{t+1} = x_{t+1}|x_t, x_{t-1}, \ldots, x_1) = P(X_{t+1} = x_{t+1}|x_t) \quad . \tag{2.16}$$

P is the probability of the system to be in a specific state at a given time. The above property shows the memorylessness of Markov processes. The probability of a future state X_{t+1} to take the value x_{t+1} is only determined by the previous state x_t, no matter of any other previous states [52] [20] [22] .

2.2.2 Markov chains

Markov chains are a type of Markov process, where the number of possible states is finite. In the discrete case, time can take non-negative integer values. In general, one could describe a Markov chain as a process where the Markov property is satisfied. The probability of the system to be in a state $X_t = x_t$ at time t only depends on the last state x_{t-1} at time $t-1$, even if more previous states are known. There is no influence of the past on future steps. Random walks are an example of a Markov chain. Imagine a random walker tossing a coin and taking either a step forward or backward. Each toss is a random trial and the final position is only depending on the next-to-last step. It does not matter how the random walker reached the next-to-last position. The famous Brownian motion of pollutants in water, as a kind of random walk and also a diffusion process, can be described using Markov processes [22] [52].

2.3 Diffusion

The concept of diffusion is widely spread throughout sciences. Nevertheless, they all share a common characteristic. The term diffusion describes in general the movement of some object, either physical or virtual, from a place with higher density to places with lower density. While in social sciences these objects might be people or ideas, the typical subjects studied in fields like chemistry or physics are molecules or particles. We will briefly walk through the history of the study of diffusion processes, a detailed description can be found in [30].

Historically, the most influential experiments for the beginning of diffusion studies in physics were performed already in the early 19th century. Starting with T. Graham, who explored diffusive processes in gases, we have a first description of the observed effect. Graham characterizes diffusion as a spontaneous mixture of gases due to the change of position of small volumes of these gases. He also stated that the size of the volumes is not equal but depends on the gases density [19]. This inspired A. Fick to formulate a continuum theory of diffusion a few years later, leading to today's well-known Fick's law [17]. Later on, botanist R. Brown discovered the disordered movement of small particles in water. This phenomenon was then explained by A. Einstein in 1905, who identified the random molecular motion with the known diffusion equation. Einstein realised that the unordered motion was caused by collisions of particles with the molecules of the solution. The effect of molecular motion was named after its discoverer — Brownian motion [8] [16] [30].

As was said above, the diffusion process is the movement of particles from a position with higher density to a position with lower density. But if we could observe each individual

particle we would notice that every motion happens randomly, meaning that every particle performs a random walk. It may seem, that the picture of random motion and the fact that a direction from higher to lower density is observed, are conflicting. Nevertheless, this can be explained by the fact that with a random motion the fraction of particles moving towards a specific direction is equal in any place. As the number of particles in regions with higher density performing such a random walk is simply larger than in regions with lower density, the fraction of particles moving towards lower density is higher than in opposite direction [14].

Depending on the behaviour of the mean square displacement $\langle r^2 \rangle$, one can distinguish between various types of diffusion processes. If the mean square displacement is linearly dependent on time like $\langle r^2 \rangle \propto t$, the process is called ordinary diffusion, whereas in case of non-linear dependency we speak of anomalous diffusion with $\langle r^2 \rangle \propto t^\alpha$ and $\alpha \neq 1$. While ordinary diffusion is well described by Fick's laws, we use random walks to model anomalous diffusion [55].

2.3.1 Fick's laws of diffusion

Diffusion is mathematically described by the diffusion equation. That was first developed by A. Fick, inspired by the theory of heat conduction derived earlier by Fourier. The equation describes that the transfer rate F of some substance through an unit area is proportional to the gradient of the concentration C. For one dimension, it can be written as [14]

$$F = -\frac{dC}{dx} D \quad . \tag{2.17}$$

This is usually called *Fick's first law of diffusion*. Here, x is the spacial coordinate and D is the *diffusion constant*. The assumption for this equation is an isotropic medium with equal diffusion characteristics in all directions. One can derive a partial differential equation, *Fick's second law of diffusion*, using the first law and by calculating the increase rate of the diffusing material. This then describes the time dependence of the concentration C on the diffusion. We can write the differential equation in three dimensions as [14]

$$\frac{\partial C}{\partial t} = \left(\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) D \quad , \tag{2.18}$$

or in the case of only one dimension as [14]

$$\frac{\partial C}{\partial t} = \frac{\partial^2 C}{\partial x^2} D \quad . \tag{2.19}$$

We assume the diffusion constant D as constant. For a detailed derivation, see [14]. In the case of invariant diffusion constant D the diffusion equation is linear and the effect is called ordinary diffusion. However, there are systems in which we cannot assume ordinary diffusion, e.g. if there are unequal probabilities that particles interact with each other. One then speaks of anomalous diffusion and the diffusion equation may become non-linear [52].

2.3.2 Random walks

Random walks have become part of many models in physics and graph theory. They can be described in a similar way as finite Markov chains. Given certain transition probabilities one can look at Markov chains as a random walker jumping from node to node in a directed graph. This kind of "diffusion" provides new tools for studying properties of graphs. It gives quantitative answers to questions concerning the spreading of information or the time steps needed for information reaching specific points in the system of interest [27].

In order to explain Brown's molecular motion, Einstein assumed each particle performing independent movements on suitable time scales, using random walk theory. In his paper he considered a number of n particles put in a liquid. Each particle would execute independent movements and be displaced by some Δ. He supposed that within a certain time τ the number of particles dn displaced by $\left[\Delta, \Delta + d\Delta\right]$ would follow a probability distribution $p(\Delta)$ yielding [16]

$$dn = n\ p(\Delta)\ d\Delta \quad , \tag{2.20}$$

with the assumption that

$$\int_{-\infty}^{+\infty} p(\Delta)\ d\Delta \ = \ 1 \qquad \text{and} \qquad p(\Delta) = p(-\Delta) \quad . \tag{2.21}$$

Using these definitions and simple series expansion he then investigated the influence of diffusion of particles in a small volume. By assigning the diffusion constant D to a factor of the series expansion as [16]

$$D = \frac{1}{\tau} \int_{-\infty}^{+\infty} \frac{\Delta^2}{2} p(\Delta) d\Delta \quad , \tag{2.22}$$

he obtains an equation of same form as the ordinary diffusion equation in (2.19). For the complete derivation, see [16]. It follows that if step sizes are reduced to infinitesimally small time and position steps, the discrete random walk becomes a continuous diffusion process.

Lets look at a simple example walk in figure 2.4. We start with a random walker at node A and will follow his way until he reaches node E. With a transition probability $p = 1$ he will jump from node A to B in the first time step. In the next step he will move from B to C with the same probability. In the third step, there are two possibilities to move: either directly to node E with probability $p = 0.5$ and ending the walk, or jump first to D with $p = 0.5$ and at last to node E. This random walk fulfils the Markov property — further steps are only depending on the current state the walker is occupying.

The simplest random walk models can be described by two characteristics: unbiased and uncorrelated. The bias concerns the preference of movement in some direction. In an unbiased random walk, the direction of motion is chosen by chance. Similarly, correlation refers to the dependence of direction on previous steps of the walk. In an uncorrelated walk, successive steps are independent of the past, fulfilling the Markov property [11].

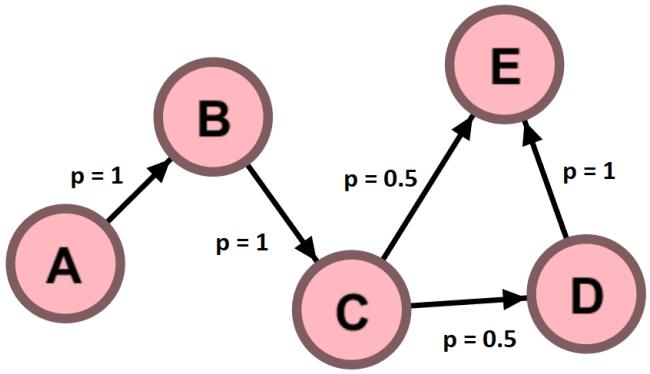


Figure 2.4: Example of a random walk on a weighted directed graph.

2.3.3 Diffusion on networks

The idea of diffusion processes can not only be applied to classical physical systems as gases, it can take place on networks as well. The general concept of diffusion can be modelled efficiently with random walks. We can distribute a number of random walkers on the nodes of a network. By choosing rules for the jump decisions of the walkers, we define specific processes. With a given network topology, it is only possible for the walkers to jump between nodes that are connected. The probability of a jump is usually given by the weights of such networks [52].
One can create a random walk model of diffusion using an update equation with discrete progressive pseudo time. We write w_i^t for the number of random walkers located at node i at time step t. In every time step, walkers from nodes j jump to node i according to some underlying transition probability $dw_{j\to i}$ of the network. Additionally, we assign each node i a property β_i defining if walkers are generated or removed. The diffusion model is started by placing a set of random walkers randomly on nodes of the network. As time passes, the following master equation describes the dynamics of the system as [52]

$$w_i^{t+1} = \sum_j dw_{j\to i}\, w_j^t + \beta_i \quad . \tag{2.23}$$

In many cases, the transition probabilities $dw_{j\to i}$ are determined by the properties of the adjacency matrix of the network. One can deduce certain characteristics from these transitions, called centrality measures. A few examples would be the PageRank or eigenvector centrality, both with specific β_i and $dw_{j\to i}$. Another special case is Laplacian diffusion [52].

2.3.4 Laplacian diffusion

In the Laplacian diffusion model we assume a large number of random walkers initially placed randomly on the nodes of a network. In addition, we set $\beta_i = 0$, so the total number of walkers is constant in the process [52]. In each time step the random walkers jump according to selected rules (often set by the direction of edges and by their weights). The diffusion process can be mathematically described in terms of the so-called graph

Laplacian which stands in relation to the adjacency matrix A. Using the graph Laplacian we obtain a diffusion equation with resemblance to Fick's law of diffusion. The importance of the graph Laplacian does not only lie in the modelling of diffusion on networks, but also on the determination of other network structures such as the existence of densely connected clusters of nodes [23] [36].

In the simplest case, an undirected and unweighed network, we describe the process with the following update equation with the transition probability $dw_{j\to i} = \frac{A_{i,j}}{k_j}$ as [52]

$$w_i^{t+1} = \sum_j \frac{A_{i,j}}{k_j}\, w_j^t \quad . \tag{2.24}$$

Here, the k_j stand for the degree of node j and β_i was set to zero. To generalize this equation for weighted and directed networks, we use the out-strength s_i^{out} instead of the node degree and obtain [52]

$$w_i^{t+1} = \sum_j \frac{A_{i,j}}{s_j^{out}}\, w_j^t \quad . \tag{2.25}$$

This is the update equation for the number of walkers at i and time $t+1$ in the case of directed weighted networks. We will return to this case after we introduce the graph Laplacian and the diffusion equation on networks.

As was mentioned in section 2.3, diffusion can be described as a process of particle movement from a place with high concentration to lower concentrations. By applying this idea to networks, we can describe this process as a flow of walkers between nodes. The time dependency of w_i is then a function of the network's diffusion constant D like [52]

$$\frac{dw_i}{dt} = D\, \sum_j A_{i,j}(w_j - w_i) = D\, \sum_j A_{i,j}\, w_j - D\, w_i\, \sum_j A_{i,j} = D\, \sum_j (A_{i,j} - \delta_{i,j}k_i) w_j \quad . \tag{2.26}$$

Here, we made use of the fact that the sum over the adjacency matrix for a fixed node i is its degree k_i. For a network with n nodes, one can define a new diagonal matrix Δ (not to be confused with the Laplacian operator or the symbol often used for small shifts) with the degrees as entries in the diagonal

$$\Delta = \begin{bmatrix} k_1 & 0 & 0 & \dots & 0 \\ 0 & k_2 & 0 & 0 & \vdots \\ 0 & 0 & \ddots & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & \dots & 0 & 0 & k_n \end{bmatrix} \quad .$$

Finally, we use the diagonal matrix to write the so-called *graph Laplacian* $\Lambda = \Delta - A$. Comparing to equation (2.26) we see that the elements $A_{i,j} - \delta_{i,j}k_i$ form $-\Lambda$. Using the more general vector notation and the newly defined matrices we can now construct a diffusion equation similar to Fick's law [52]:

$$\frac{dw}{dt} = -D\ \Lambda\ w \qquad \text{or} \qquad \frac{dw}{dt} + D\ \Lambda\ w = 0 \quad . \tag{2.27}$$

This corresponds to Fick's law with the exception of a minus sign and the fact that the Laplacian operator is replaced with the graph Laplacian Λ. Now we see the reason for the definition of Λ and for its name. The solution of the differential equation (2.27) can be found using the eigenvalues and corresponding eigenvectors of Λ in combination with given initial conditions [23] [36].

Returning to the case of random walks on directed weighted networks, we can modify the above equations of diffusion for this purpose using the out-strength s_i^{out} instead of degree k_i. The entries in the diagonal matrix Δ simply change to the out-strength. Note that the adjacency matrix and therefore also the Laplacian matrix are not symmetric any more. In the following, we describe a random walk model for Laplacian diffusion described in a project work [23], adapted for the purpose of this thesis. It characterizes diffusion as an iterative process, where future states of w_j are updated based on past calculations like

$$w_i^{new} = \sum_j \frac{A_{i,j}}{s_j^{out}}\ w_j \quad . \tag{2.28}$$

The model assumes a directed weighted network structure with random walkers positioned at its nodes. The weights $\frac{A_{i,j}}{s_j^{out}}$ describe the probability of a walker to jump to adjacent nodes. The algorithm used to model the diffusion process is based on the iterative equation (2.28) and computerized random number generation performed for every single walker of the system. The explicit form of the implemented algorithm will be further described in the next chapter.

Chapter 3

Methods

In this chapter we focus on the methodological description of the two main analysis parts of this work: the random walk model and network motif analysis. We first describe the used claims data and the general data processing. In the random walk model we determine the probabilities of patients to move to specific specialists in the HCP system and use the resulting transition probabilities in a random walk algorithm to characterize the system's memorylessness. To detect network motifs, we define specific re-hospitalization patterns and analyse their sex-specific prevalence. For both analysis parts we describe their basic concepts, data processing, sample definitions, as well as how the analysis results are calculated.

3.1 Data description and processing

First we briefly describe the underlying dataset for both parts of the main analysis. We use an pseudonymized medical claims dataset of Lower Austria that is based on three different subsets: patients' data, diagnosis data and contact data.

- **Patients' data**: identification number (ID), year of birth and sex of 1674266 patients with insurance contract in Lower Austria between years 2006 and 2012. In total, the sample consists of 865125 female and 809141 male patients.
- **Diagnosis data**: list of diagnoses that patients received in hospitals of Lower Austria in the same timespan. We differentiate between 1642 distinct diagnoses, each with a corresponding three-digit ICD10 code. For each code a table with information about diagnosed patients, principal or secondary diagnosis, admission and release date is given.
- **Contact data**: information about patients' contacts to different HCP in Lower Austria in the same timespan. The HCP are divided into 45 separate specialities. For every speciality the IDs of patients who had a contact and the dates of contact are known.

We select data of patients with known year of birth and sex and study the timespan between January 1, 2006 and March 31, 2012. In case of the network motif analysis we are interested

in re-hospitalizations, so the focus is set on patients with at least one hospitalization in the selected timespan. We choose the same patient sample also for the random walk model. Diagnosis and contact data of other patients or out of the selected timespan is sorted out. A sketch of the population selection is shown in figure 3.1.

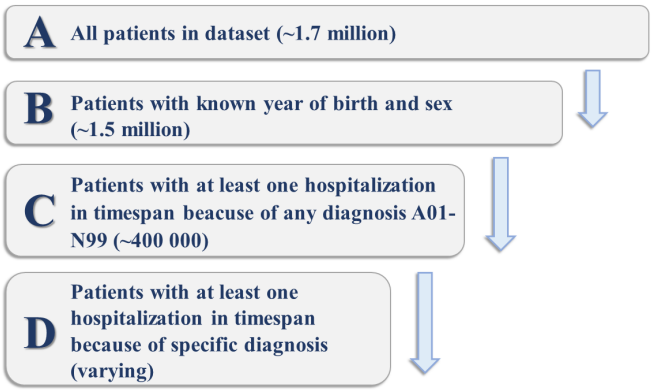


Figure 3.1: Population selection. Starting from the entire patient set A, we delete patients with no meaningful connection to the analysis up to the lowest and smallest sets in D.

The classification of specialities in the contact-data was modified. Some were deleted (because empty or too little data available) or aggregated (because similar disciplines). The diagnosis data is included as another speciality, the hospital, that can also be contacted. We are left with a total of 18 specialities. A list of them is shown in the appendix (table 5.5). In the diagnosis data we only considered diagnoses from a range of ICD10 starting with A01 and ending with N99. All others were considered either too sex-specific (e.g. child births) or too unspecific for re-hospitalization analysis (e.g. private/work accidents).

3.2 Random walk model

3.2.1 Basic concept

Patients move within a system of healthcare providers (HCP). Each patient i has an internal state $\phi_i(t)$ that is given by the type of specialist last contacted. As time progresses, patients move from one speciality to another (or from one state to another) within specific time windows τ_i and form patterns, or paths. An example of such a path is depicted in figure 3.2.

We define a *path* as a sequence of states in the HCP-system, where the time window τ_i between the states is smaller than some maximal value τ_{max}. We are interested in directly connected contacts, referrals of doctors and related conditions of patients. Therefore, if the time window between two consecutive contacts is larger than τ_{max}, the first path stops and a new path starts. A path consists of at least two states and at most of all states a patient was in. We neglect the cases, where patients had contact to some other speciality than one of the chosen 18 and hospitalizations with other than the selected diagnoses.

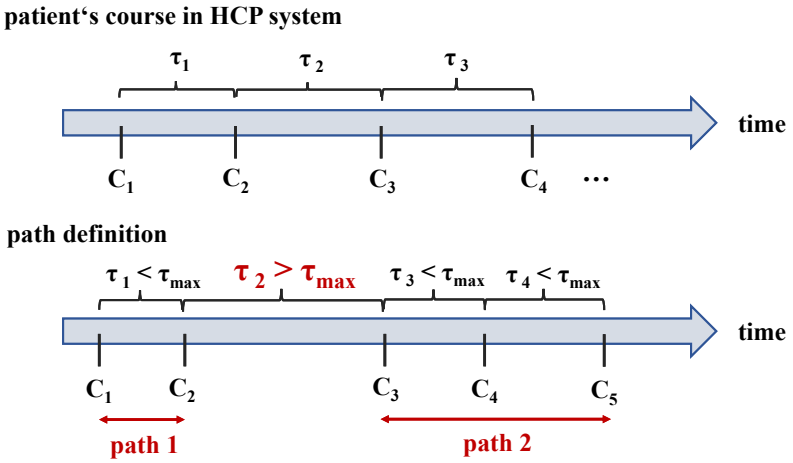


Figure 3.2: Top: Patients are moving in the HCP network. Each contact C stands for a step patients take. Bottom: If the time window τ_i between contacts is too large, paths are split.

By gathering information about paths of a large sample of patients, we can calculate the probability of moving from one specialist to another. Just as in a random walker model, we can use these probabilities to predict a patient's step in the network. Depending on the number of given previous steps, in the following denoted by m, the order of the model prediction changes. A scheme of this prediction model is shown in figure 3.3. We choose to only predict the very last step of patients' paths. This has the advantage that in many cases, there will be enough previous steps that can be used for prediction. It also reduces the computational time spend on the models, as we do not need to generate another random number to select the specific step.

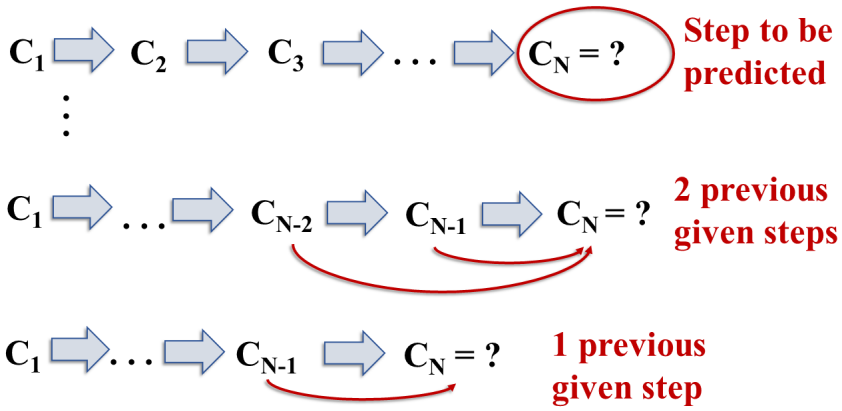


Figure 3.3: Basic concept of predictions. The last step C_N of a patient's path is to be predicted. The Markov model order changes depending on the number of given previous contacts C used for prediction.

One aim of this analysis is to compare the predictions of both sex. If there is any difference in the quality of predictions, it would suggest variations in the diffusion processes of females and males in the network. Apart from the sex-specific analysis, we test the memory of the

processes in the system. In specific: is the process of patients moving to specialists in the system memoryless and therefore Markovian or is it path-dependent? And how does the number of given steps m influence the quality of step prediction? If we can observe an increase of correct predictions with raising number of given steps, the process of moving through the HCP-system is not memoryless and therefore non-Markovian. Another matter of interest is the variation of correct predictions for different age groups. Can we observe any difference in prediction quality for patients of different ages?

3.2.2 Analysis principle

For the random walk model we use information of patients of all ages from patient set C of figure 3.1. In addition we need the patients' contact information in the selected timespan. We further divide the resulting patient sample into female and male patients. For each patient a matrix of all his or her contacts is created. The matrices contain information about date of contact and specialist. Eventually, we will predict patients' steps from one speciality to another and need to construct transition probabilities in the HCP-system. To do so, we loop over all patients (for males and females separately) and all their paths and gradually add weights to the links in the HCP network. In case of $m = 1$ (one given previous step), the network consists of the $n = 18$ chosen specialities (nodes) and the transition matrix therefore has a size of $n \times n$. The resulting network is directed and weighted.
For higher order chains, $m > 1$, the creation of transition matrices is similar, but somewhat more complicated. For every possible permutation of previous steps a separate matrix needs to be determined, providing more information to the transition process. The number of matrices increases with the number of previous steps m like $\propto n^m$, where n is the number of nodes in the system (specialities).

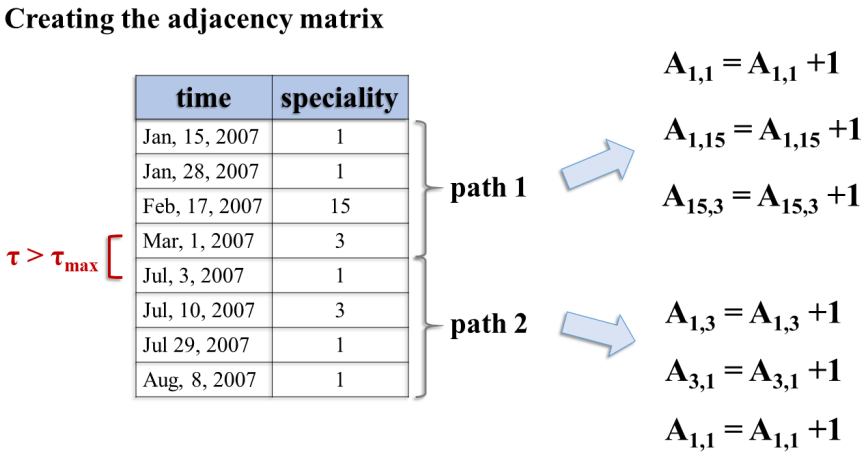


Figure 3.4: Constructing the adjacency matrix. Example for a patient with two separate paths. For every transition between specialists within a path, the adjacency matrix entry is raised.

To compare the quality of prediction for both sexes, we will create three different transition matrices for each model: one based solely on female data, one based solely on male data

and one matrix based on combined data of both sexes. Using these different matrices we can compare the quality of predictions if sex-specific transitions or sex-independent transitions are used in the random walk model.
The process of creating the adjacency matrix is explained in the following example. Imagine some patient of the sample with several contacts C he had in the selected timespan. We order the contacts by time. If more than one contact is dated the same day, the ordering is simply given by the numbering of specialities as shown in the appendix (table 5.5). We then divide the resulting contact history into sections according to the definition of paths. For every correct path (i.e. for paths with a length $\geq m+1$), we count the transitions $C_j \to C_i$. For every such transition, the adjacency matrix element A_{ij} for the transition from j to i is raised by one. This process is repeated for every patient of the sample. The adjacency matrix is updated for each step. We end up with weighted directed network matrices. The meaning of the entries in A is the number of steps taken from some speciality in column j to speciality in row i. The adjacency matrices therefore describe transitions in the HCP network. Figure 3.4 depicts the process of creating adjacency matrices.

3.2.3 Parameters

Several parameters are defined for the memory analysis. To select correct paths, we need a maximal value for the time window in between contacts τ_{max}. We will calculate averages of correct predictions in the models and therefore need to define the number of randomly chosen patients/paths n_{pat} per model run and the number of repetitions per model n_{rep}. The following table 3.1 shows the chosen parameter setting.

Parameter	Setting
maximal time window τ_{max}	21 days
number of repetitions n_{rep}	100
number of patients/paths (system memory analysis) $n_{pat,1}$	1000
number of patients/paths (age dependent analysis) $n_{pat,2}$	100

Table 3.1: Random walk analysis parameter setting.

The maximal time window was defined as 21 days, this provides enough time for contacts to be connected in some way (referrals, related health problems). As the model is based on many predictions of randomly chosen patients, we use averages and the standard deviation of 100 model runs as measures of prediction quality. In each run, patients are randomly selected out of the sample. In case of the system memory analysis part, the underlying patient samples are large enough to provide 1000 random patients per model run. For the age-dependent memory analysis the sample sizes drop and the number of patients per run is lowered to 100.

3.2.4 Random walk model

We briefly describe the basic steps of the random walk model. Small changes will be needed in models with different numbers of given steps m and for the age-dependent analysis, but

the key algorithm to the process is the same.
First, we need to define the weighted directed adjacency matrix A of the current underlying network as explained in section 3.2.2. A depicts the transitions in the system. Next, we create the diagonal matrix Δ. The entries in the diagonal of Δ are the sums over the columns in A and stand for the out-strength s_i^{out} of the nodes (specialists). The matrices have the following structures:

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,n} \\ A_{2,1} & A_{2,2} & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ A_{n,1} & \dots & \dots & A_{n,n} \end{bmatrix} ,$$

$$\Delta = \begin{bmatrix} \sum_{i=1}^{n} A_{i,1} & 0 & \dots & 0 \\ 0 & \sum_{i=1}^{n} A_{i,2} & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ 0 & \dots & \dots & \sum_{i=1}^{n} A_{i,n} \end{bmatrix} = \begin{bmatrix} s_1^{out} & 0 & \dots & 0 \\ 0 & s_2^{out} & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ 0 & \dots & \dots & s_n^{out} \end{bmatrix} .$$

For the random walk model we use an already existing algorithm as described in a project work [23]. It is basically a process of creating random numbers and probabilities and updating the final jump decision of patients in a repetitive manner. Predicted results are then compared to the real outcome. In each model run the following three random numbers ξ need to be created.

- random patient: $\xi_1 \in [1,\ \textit{sample size}]$
- random path: $\xi_2 \in [1,\ \textit{number of paths}]$
- random walk probability: $\xi_3 \in [0, 1]$

First we randomly select a patient whose next step should be predicted. This patient is chosen by generating a random number ξ_1 between 1 and the total number of patients in the current sample. Next, we need to check whether the selected patient has at least one path with a correct length. Correct means for a model with m given previous steps, that a path with length of at least $l = m + 1$ exists. If the selected patient has no such path, a new ξ_1 is generated. If he has at least one correct path (or more), ξ_2 is initialized in a way to choose a random path of this patient. The key part of the random walk model uses the information of the adjacency matrix A and diagonal matrix Δ. We save the destination of the given step in variable k and the real last step to target T. We now generate ξ_3 and initialize an additional variable $r = 1$ for the jump decision which is updated in every loop. The basic scheme of the necessary algorithm is shown in the following.
Given the condition $\sum_{j=1}^{r} A_{ik}/\Delta_{kk} < \xi_3$, the jump decision r is updated in the loop over the possible transitions. Once the while-loop is exited, the jump decision is fixed and is defined by r. We check if the predicted step matches the real last step T. If the prediction is correct, we update the number of correct predictions in the model run.

Algorithm 1: Basic algorithm for the random walk model. The final jump decision is updated while the condition is fulfilled. Subindex i indicates the model run.

```
ξ3 = rand(0,1);
r = 1;
while Σ_{j=1}^{r} A_jk/Δ_kk < ξ3
    r = r + 1;
end
if r == T
    correct_i = correct_i + 1;
end
```

The whole procedure is repeated for $n_{pat,1/2}$ patients' paths (depending on the analysis; age-independent or age-dependent). This yields a percentage of correct predictions per model run. After $n_{pat,1/2}$ paths, a new model run starts. This process is repeated n_{rep} times and results in n_{rep} percentages per model. We calculate the mean μ and standard deviation σ in percentages of the correct predictions $correct_i$ with $i = 1, \ldots, n_{rep}$ as

$$\mu = \frac{\sum_{i=1}^{n_{rep}} \left(\frac{correct_i}{n_{pat,1/2}}\right)_i}{n_{rep}} * 100 \qquad \text{and} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n_{rep}} |\left(\frac{correct_i}{n_{pat,1/2}}\right)_i - \mu|^2}{n_{rep} - 1}} * 100 \quad . \tag{3.1}$$

The procedure described in this section is used in all following memory analysis parts. It will only deviate in the underlying patient samples, number of paths and the creation of transition matrices.

3.2.5 System memory analysis

In the first part of the random walk model we test the memory of the diffusion process of patients in the HCP system. Using varying numbers of m in the models, we can compare the resulting averages of correct predictions and characterize path dependence. We start with one given previous step $m = 1$ and end with five previous steps, $m = 5$. The result are five different models, each with averages out of n_{rep} model runs.

The process of the random walk model for the system memory analysis part has been explained in the previous section. We use $n_{pat,1}$ randomly chosen paths separately for females and males and create n_{rep} repetitions for each model of m. The creation of adjacency matrices of higher order chains is somewhat more complicated than explained in 3.4 but follows the same principle. For first order chains $m = 1$(one given step $\rightarrow$ Markov process), the model works with one matrix as described in 3.2.2. For every m the mean and standard deviation of correct predictions for males and females with separate and combined transition matrices are depicted in the results chapter.

To provide an overview on how many possible paths the models work with, we define the *path length* l as the number of steps in a patients' path. Models with a given number of previous steps m need a sufficiently long path with $l \geq m + 1$. A histogram of the path lengths l in the patient samples of males and females is shown in figure 3.5. Sequences of

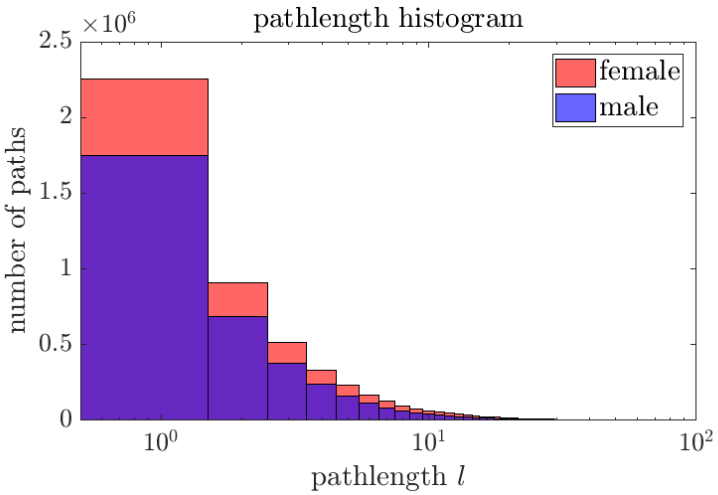


Figure 3.5: Histogram of patients' path lengths. Note that the x-axis uses a logarithmic scale while the y-axis is linear. Paths of length $l = 1$ dominate in the system.

length $l = 1$ are dominant in the system while the number of longer paths decreases. Due to the larger number of female patients in the sample, females tend to have more paths than males throughout all path lengths. As $l = 1$ sequences cannot be considered correct paths in our definition, they do not appear in the models. Yet the sample of all remaining paths is sufficiently large for the random walk model.

3.2.6 Age-dependent memory analysis

The second part of the random walk model deals with the analysis of age dependence of prediction quality. We test whether patients of specific ages have different path behaviour. To do so, we divide the patient samples into age groups. An age group is defined by the year of birth of patients. We choose to work with a 10-year timespan for each age group a to provide an adequate sample size. Starting with the youngest patients in age group a_1 with a maximal year of birth in 2006, patients are distributed into 10 categories up until a_{10} with a minimal year of birth in 1907. A table with all categories and the corresponding years of birth together with the specific numbers of females and males can be seen in figure 3.6 and table 3.2.

As we see in the histogram, some patients are not included in any of the ten categories as their year of birth exceeds the selected limits ('none'). The number of patients available in a_{10} is relatively small compared to all other age groups. The number of patients we randomly sample in the models is therefore lowered to $n_{pat,2} = 100$ due to smaller sample sizes. The random walk models themselves are equivalent to the algorithm described in section 3.2.4. For every age category a we create the three transition matrices: two separate matrices for males and females and one with combined information. The results of the age-dependent memory analysis depict the mean and standard deviation of correct predictions for every age group separately for females and males, for separate and combined transition matrices and for various numbers of given steps, from $m = 1$ to $m = 4$.

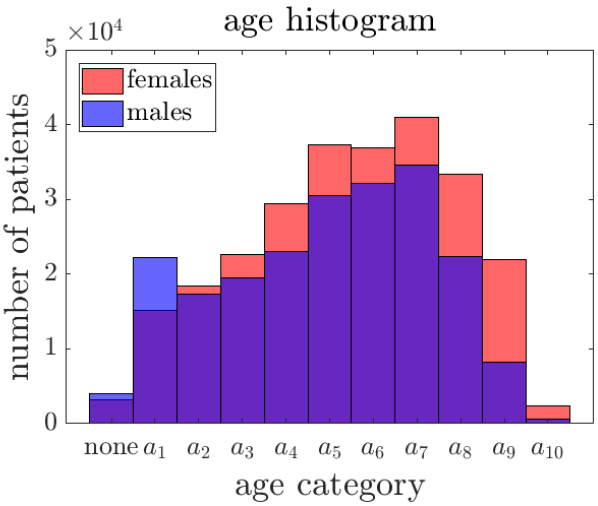


Figure 3.6: Histogram of patients' age categories for females (red) and males (blue). 'None' category shows patients that do not fit into any of the other groups.

age group	year of birth	number of females	number of males
a_1	2006-1997	15195	22216
a_2	1996-1987	18465	17343
a_3	1986-1977	22598	19462
a_4	1976-1967	29432	23054
a_5	1966-1957	37326	30553
a_6	1956-1947	36946	32069
a_7	1946-1937	40935	34564
a_8	1936-1927	33365	22358
a_9	1926-1917	22003	8280
a_{10}	1916-1907	2396	599
none	not listed	3167	3995

Table 3.2: Table of age categories with corresponding years of birth and number of patients in the groups.

3.3 Analysis of network motifs

For further study of the HCP network we analyse selected network motifs. Using regression analysis we will examine several motifs defined as specific re-hospitalizations of patients. Given the underlying dataset, we will focus on the question of sex-specific re-hospitalization risk under certain conditions connected to specialist contacts. The following sections briefly describe the basic idea and methods of this analysis part.

3.3.1 Basic concept

Patients receive various diagnoses in hospitals and visit different specialists depending on their illness. The patterns of diagnoses and contacts can be identified with network motifs. Can we observe such motifs in the data? Given some specified diagnoses and specialists, is there any effect of specialist contacts on the risk of being re-hospitalized?

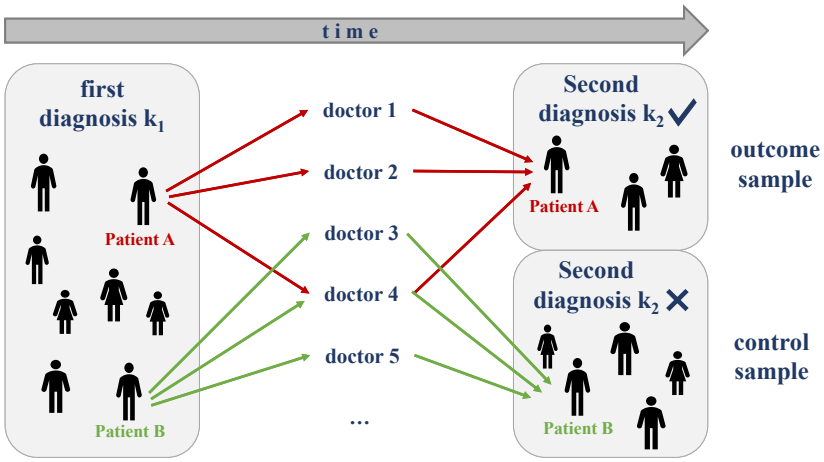


Figure 3.7: Basic analysis idea. Patients with a specific first diagnosis can have contacts to different doctors within a given timespan. Some of these patients then end up with a second diagnosis, some do not. This scheme can be used to define the motifs in the patient-sharing network.

The problem is depicted in figure 3.7. We select patients who received a first diagnosis k_1 and observe their contact behaviour afterwards. The motifs in this system can be described as follows: patients A and B were both hospitalized with the first diagnosis. In case of patient A we observe contacts to several specialists and a received second diagnosis k_2. In contrast to that, patient B also had contacts to doctors, some were the same as A's, some were not. But he did not end up with the second diagnosis. This process clearly represents a directed flow of patients in the network. Using the states of hospitalizations and contacts, we define motifs that occur with different frequencies in the network. In the following we will test the correlation between contacts to doctors and hospitalizations. Is there any effect of the contact on the risk of receiving a second diagnosis?

3.3.2 Analysis principle

The network motif analysis is based on the information we have on patients' contacts between two hospital stays. As an example we look at two patients: patient A gets hospitalized with first diagnosis k_1 at time $t_{A,1}$. Within a specific time window Δt he gets re-hospitalized with second diagnosis k_2 at time $t_{A,2}$. Δt must fulfill certain constraints. In particular, we only consider motifs where Δt is in the range $[\Delta t_{min}, \Delta t_{max}]$. Between his hospitalization times $t_{A,1}$ and $t_{A,2}$ patient A might have had contacts C_i to several specialities (or not). Patient B also gets hospitalized with diagnosis k_1. There is no diagnosis k_2 after the maximal time window Δt_{max}, but within $t_{B,1} + \Delta t_{max}$ he had contact to some specialities. A simplified illustration of these sequences is shown in figure 3.8. Here we see the relevant sampling time for each patient.

We define specific network motifs based on selected first and second diagnoses and specialist contacts. We can create a regression model describing the effects of different factors on the risk of being re-hospitalized. In specific, we study the effects of various first and second diagnoses k_1 and k_2 and contact with different specialities on the re-hospitalization risks for

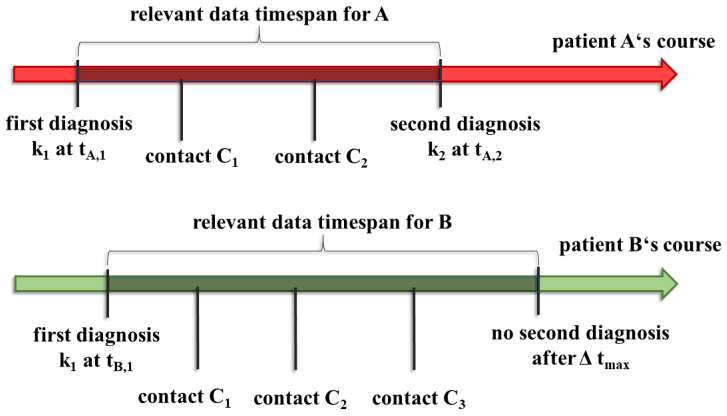


Figure 3.8: Example for relevant data sampling and motifs. Imagine some patient A's and B's trajectories in time. The relevant timespan starts with the first diagnosis and ends with the second (in case of patient A) or after the maximal time window (for patient B). In between there might have been contacts C_i to some doctors. We will use this scheme to define motifs in the network.

females and males, separately. This analysis is divided into two parts: contact-independent (neglecting specialist contacts) and contact-dependent (considering specialist contacts). In the following, we define all parameters used in the analysis, select patient samples together with their relevant contacts and define a sample of diagnoses.

3.3.3 Parameters

Several parameters can be varied in this model: minimal and maximal age of patients, minimal and maximal time window between hospitalizations, minimal number of females and males with a re-hospitalization and minimal number of unique patients per diagnosis. The final list of settings is shown in table 3.3. We choose the year of birth so that the patients' age lies between 50 and 100 years. Most diagnoses are received by patients in this age range. The timespan between minimal and maximal time window is the re-hospitalization window. The minimum is set to 90 days to prevent bias from patients that went to follow-up hospital procedures. The maximum of 1050 days yields enough re-hospitalization cases for a meaningful analysis. We refer to this setting by the main parameter setting. To test the robustness of the results, we vary some of the parameters. The results of the robustness test can be found in the appendix.

3.3.4 Patient samples and relevant contacts

The age range of interest are patients that were born between y_{min} and y_{max}. This means that by 2006 their age lies between 50 and 100 years. For the regression analysis we need an outcome sample and for validation an additional control sample. We select these samples for every diagnosis combination k out of patient set C from figure 3.1. We create a matrix of all patients with specific first and specific second diagnoses k_1 and k_2 and sort the matrix first by ID, then by time. An example for such a matrix is shown in figure 3.9.

Parameter	Setting
minimal year of birth y_{min}	1906
maximal year of birth y_{max}	1956
minimal time window Δt_{min}	90 days
maximal time window Δt_{max}	1050 days
minimal number of females and males per combination cut_{comb}	50
minimal number of patients per diagnosis cut_{diag}	1000

Table 3.3: Main analysis parameters. Some settings will be changed for the robustness test (see appendix).

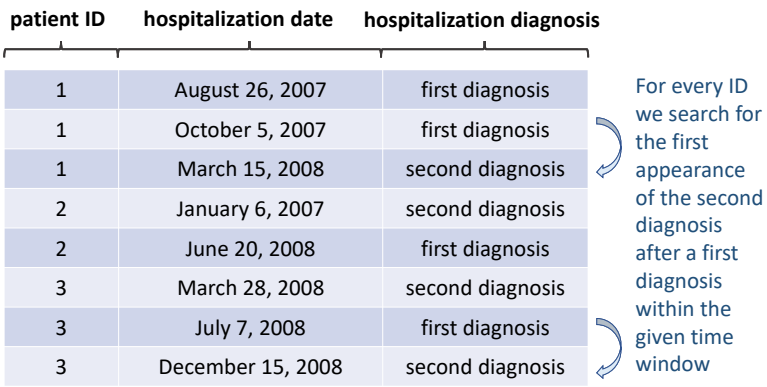


Figure 3.9: Made-up example table and definition of re-hospitalization. Only patients with the correct order of diagnoses are part of the outcome sample.

- **Control sample**: sample of patients in the selected age range who received the first diagnosis k_1 but not the second diagnosis k_2 within the specified time window. For every patient the information about date of first appearance of the first diagnosis is saved. In figure 3.8, patient B would be in the control sample.

- **Outcome sample**: sample of patients in the selected age range with a correct re-hospitalization: first diagnosis k_1 followed by second diagnosis k_2 within the selected time window. Date of appearance of first diagnosis and date of first following second diagnosis are saved for each patient. From the scheme in figure 3.8, patient A would be selected in the outcome sample.

For both, control and outcome sample, we are interested in contacts to specialists patients had after the first diagnosis. The analysis does not depend on the amount of contacts patients had to the same speciality. The importance lies on the information whether a specialist was contacted at all or not. In case of the outcome sample we consider all contacts between k_1 and k_2. For patients in the control sample we start with contacts after the date of k_1 and stop after the maximal time window (see figure 3.8). For both samples we exclude the hospital contacts because of k_1 and k_2. Nevertheless, it is still possible that hospital contacts appear in between the two diagnoses because of other ICD10 codes.

3.3.5 Diagnosis selection

We set a lower limit of cut_{diag} unique patients with a recorded principal or secondary diagnosis and reject all others. Diagnoses can appear as first hospitalizations and as second hospitalizations (re-hospitalizations), resulting in *diagnosis combinations*. We will always denote the first hospitalization diagnosis as k_1, the second as k_2 and the diagnosis combination as k. Diagnosis combinations are further analysed if there are at least cut_{comb} female and male patients in the outcome sample and in the control sample. All other combinations are dismissed.

			d = 1, ..., 1055			
ID	**sex *s***	**year of birth *y***	***d = 1***	***d = 2***	***d = 3***	...
1	0	1909	0	0	1	...
2	0	1954	0	0	0	...
3	1	1962	1	0	0	...
4	1	1934	1	0	0	...
...	...	...	...	...	...	...

Figure 3.10: Made-up example for diagnosis matrix M_{diag}. It contains information about all diagnoses (d_p for every patient and every diagnosis) and basic patient data. We will use the data from this matrix later for calculating diagnosis prevalence.

For simplification and later purposes, we first create a matrix M_{diag} containing information about all diagnoses patients received over time (see example in figure 3.10). Each row in this matrix represents a patient of patient set B from figure 3.1. The first column contains the ID, the second column stands for the sex s of the patients with values $s = 0$ for females and $s = 1$ for males. The third column indicates the year of birth. All following columns stand for one of the diagnoses from ICD10 codes A01 to N99 and are denoted d with $d = 1, \ldots, 1055$. The entry per patient d_p in these diagnosis columns can take the value $d_p = 1$ if the patient received the diagnosis as a principle or secondary diagnosis and $d_p = 0$ if this is not the case.

Matrix M_{diag} contains patients of all ages and can also contain patients who did not receive any diagnosis in the selected timespan. By defining values and parameters like the maximal and minimal year of birth, sex and diagnosis, we can use this matrix to count the number of elements with correct properties. From now on, we will write the subscript M for sex value $s = 1$ (males) and F for $s = 0$ (females). We receive the number of patients with (n_d) or without ($n_{\neg d}$) a selected diagnosis d that can be written as

$$n_{M,d} = |M_{diag}(s = 1 \,\&\, y \leq y_{max} \,\&\, y \geq y_{min} \,\&\, d_p = 1)| \quad , \tag{3.2}$$

$$n_{F,d} = |M_{diag}(s = 0 \,\&\, y \leq y_{max} \,\&\, y \geq y_{min} \,\&\, d_p = 1)| \quad , \tag{3.3}$$

$$n_{M,\neg d} = |M_{diag}(s = 1 \,\&\, y \leq y_{max} \,\&\, y \geq y_{min} \,\&\, d_p = 0)| \quad , \tag{3.4}$$

$$n_{F,\neg d} = |M_{diag}(s = 0 \,\&\, y \leq y_{max} \,\&\, y \geq y_{min} \,\&\, d_p = 0)| \quad . \tag{3.5}$$

The vertical bars |...| stand for the cardinality, which is the number of elements of a given set with selected conditions. We will use these equations later to calculate diagnosis prevalence.

3.3.6 Contact-independent analysis

The first simple network motif of interest consists of the combination of two diagnoses that are connected if a minimum of cut_{comb} male and female patients are involved in the flow. No contacts in between diagnoses are considered. The direction of flow is defined from k_1 to k_2. If such combinations can be found, regression models are created and sex differences are analysed.

Regression model

In the contact-independent case we study the effects of different k_1 and k_2 on the re-hospitalization risk of both sex, independent of contacts that patients might have had. Separate regression models are generated for every diagnosis combination k and for both, males and females. For preparation we create matrices of all female and all male patients in the control and outcome sample for every combination k. These matrices contain information about predictor and response variables. Each row stands for a patient and in this case, it contains only one predictor variable, the year of birth y. The 'response' column of the matrix provides information about possible re-hospitalization as a dummy variable with only two possible states. This means outcome patients were re-hospitalized ($Z = 1$), control patients were not ($Z = 0$). Every matrix represents the patient sample for a diagnosis combination k. An example matrix is shown in figure 3.11.

	Predictor variable	Response variable
patient	year of birth y	rehospitalization
1	1945	1
2	1930	1
3	1940	1
4	1942	0
5	1938	0
...	...	...

Figure 3.11: Made-up example for contact-independent regression matrix. Each row stands for a patient who received k_1. The 'response' column states if k_2 appeared within the selected time window.

We then create computer generated regression models based on the matrices under the condition of the cut-off parameter cut_{comb} in the outcome sample and the control sample. We use a logistic link function (see equation (2.15)), as the distribution of the response variable is binomial.

Regression responses

Given the previously generated regression models we predict the response for specified values with the logistic function (2.13). We use the coefficient of the intercept α and the coefficient for the year of birth β_1 of the model. We want to predict the response for re-hospitalization for a typical patient of the sample and therefore define $\overline{y}$ as the mean year of birth for every sex of the current sample. The results of the predictions are the age-adjusted re-hospitalization risk response $\hat{P}_{M,k}$ for males and $\hat{P}_{F,k}$ for females for each diagnosis combination k. The responses are predicted within a 95% confidence interval and are defined as

$$\hat{P}_{M,k} = \frac{\exp(\alpha_{M,k} \ + \ \beta_{1,M,k} \ \overline{y}_{M,k})}{1 \ + \ \exp(\alpha_{M,k} \ + \ \beta_{1,M,k} \ \overline{y}_{M,k})} \quad \text{and} \tag{3.6}$$

$$\hat{P}_{F,k} = \frac{\exp(\alpha_{F,k} \ + \ \beta_{1,F,k} \ \overline{y}_{F,k})}{1 \ + \ \exp(\alpha_{F,k} \ + \ \beta_{1,F,k} \ \overline{y}_{F,k})} \quad . \tag{3.7}$$

Logarithmic relative risk

To compare the responses of males and females we calculate the logarithmic relative risk $logRR_k$ for each combination k as

$$logRR_k = log\left(\frac{\hat{P}_{M,k}}{\hat{P}_{F,k}}\right) \quad . \tag{3.8}$$

By this means we can easily identify if the predicted re-hospitalization risk is larger for males or for females, depending on the sign of $logRR_k$.

3.3.7 Contact-dependent analysis

In the next part we include specialities to the network motifs and define a flow as a directed link from k_1 to some speciality i to k_2. Each such motif must contain a minimum of cut_{comb} female and male patients to be considered for further analysis.

Regression model

So far, we have neglected the contacts of patients in the time window between hospitalizations. Now we add the information about specialist contacts to the previously created regression matrices to analyse their effect on the re-hospitalization risk. The matrices are constructed in the same way as in the contact independent section, but in addition to the year of birth column we include a column with information about contact to a specialist c_i. An example is given in figure 3.12.

The matrices are created separately for each diagnosis combination k, for every speciality i and for males and females, respectively. A regression model is generated if the cut-off parameter cut_{comb} is satisfied with and without contact, in the control and the outcome sample. Apart from year of birth y, the contact information c_i is the second predictor variable. It is a categorical variable, we write $c = 1$ if a contact occurred, $c = 0$ means no

	Predictor variables		Response variable
patient	year of birth y	c_i	rehospitalization
1	1945	1	1
2	1930	1	1
3	1940	0	1
4	1942	1	0
5	1938	0	0
...	...	...	...

Figure 3.12: Made-up example for contact-dependent regression matrix. An additional column for the contact information is added. It states for each patient of the sample, whether there was at least one visit to specialist i in between diagnoses or within the maximal time window.

contact with this speciality. We assume that the predictor variables are independent from each other.

Regression responses

Based on the obtained coefficients α and β_1, β_2 from the regression models we predict the responses for a chosen setting of predictor variables. For every combination k, every speciality i and both sex we set $\overline{y}$ as the mean year of birth of the patient sample. For the contact variable we can choose between two possibilities: contact or no contact ($c_i = 1$ or $c_i = 0$). For later purposes we predict the responses for both cases. The corresponding coefficient for the year of birth is β_1, while β_2 is the contact coefficient. We obtain the contact-dependent, age-adjusted re-hospitalization risk response $\hat{P}_{M,k,c_i}$ for males and $\hat{P}_{F,k,c_i}$ for females for each diagnosis combination k and speciality i as

$$\hat{P}_{M,k,c_i=1} = \frac{\exp(\alpha_{M,k,c_i} + \beta_{1,M,k,c_i}\,\overline{y}_{M,k,c_i} + \beta_{2,M,k,c_i} \times 1)}{1 + \exp(\alpha_{M,k,c_i} + \beta_{1,M,k,c_i}\,\overline{y}_{M,k,c_i} + \beta_{2,M,k,c_i} \times 1)} \quad , \tag{3.9}$$

$$\hat{P}_{F,k,c_i=1} = \frac{\exp(\alpha_{F,k,c_i} + \beta_{1,F,k,c_i}\,\overline{y}_{F,k,c_i} + \beta_{2,F,k,c_i} \times 1)}{1 + \exp(\alpha_{F,k,c_i} + \beta_{1,F,k,c_i}\,\overline{y}_{F,k,c_i} + \beta_{2,F,k,c_i} \times 1)} \quad , \tag{3.10}$$

$$\hat{P}_{M,k,c_i=0} = \frac{\exp(\alpha_{M,k,c_i} + \beta_{1,M,k,c_i}\,\overline{y}_{M,k,c_i} + \beta_{2,M,k,c_i} \times 0)}{1 + \exp(\alpha_{M,k,c_i} + \beta_{1,M,k,c_i}\,\overline{y}_{M,k,c_i} + \beta_{2,M,k,c_i} \times 0)} \quad , \tag{3.11}$$

$$\hat{P}_{F,k,c_i=0} = \frac{\exp(\alpha_{F,k,c_i} + \beta_{1,F,k,c_i}\,\overline{y}_{F,k,c_i} + \beta_{2,F,k,c_i} \times 0)}{1 + \exp(\alpha_{F,k,c_i} + \beta_{1,F,k,c_i}\,\overline{y}_{F,k,c_i} + \beta_{2,F,k,c_i} \times 0)} \quad . \tag{3.12}$$

Equations (3.9) and (3.10) predict the re-hospitalization risk response in the case that a contact with i took place. Equations (3.11) and (3.12) predict the response if there was no contact with i.

Logarithmic ratio and relative risk ratio

Like in the contact-independent case, we can compare the predicted responses for the risk of males and females. Additionally, in the contact-dependent case we now have two possibilities for the contact state. We are interested in the influence of contacts on the risk compared to not having a contact and we therefore create the logarithmic ratios $logR_{M,k,i}$ for males and $logR_{F,k,i}$ for females for every combination k and every speciality i like

$$logR_{M,k,i} = log\Big(\frac{\hat{P}_{M,k,c_i=1}}{\hat{P}_{M,k,c_i=0}}\Big) \quad \text{and} \tag{3.13}$$

$$logR_{F,k,i} = log\Big(\frac{\hat{P}_{F,k,c_i=1}}{\hat{P}_{F,k,c_i=0}}\Big) \quad . \tag{3.14}$$

Depending on the sign of the logarithm we can measure if the risk with contact is larger or smaller compared to the risk if there was no contact. To compare the effect of contacts on the risk for males to females, the logarithmic relative risk ratio $logRRR_{k,i}$ is calculated using equations (3.13) and (3.14) as

$$logRRR_{k,i} = log\Big(\frac{R_{M,k,i}}{R_{F,k,i}}\Big) = log\left(\frac{\frac{\hat{P}_{M,k,c_i=1}}{\hat{P}_{M,k,c_i=0}}}{\frac{\hat{P}_{F,k,c_i=1}}{\hat{P}_{F,k,c_i=0}}}\right) \quad . \tag{3.15}$$

This value describes the relative risk alteration of males to females due to contacts. If the risk alteration with contact is larger for males, (3.15) is positive, if it is larger for females, the value is negative.

3.3.8 Averaging over diagnosis combinations

Given the selected cut-off parameter cut_{diag} in table 3.3, 193 diagnoses or $193^2 = 37249$ diagnosis combinations remain to be analysed. Single diagnosis combinations provide information about the risk of a specific kind of re-hospitalization. To make general statements about the risk of individual diagnoses and also specialities, we averaged the predicted responses for the following groups.

- fixed first diagnosis k_1
- fixed second diagnosis k_2
- fixed specialities i

The meaning of the fixation is the following: Every diagnosis of the sample can appear as first and as second diagnosis. A fixed first diagnosis k_1 can have many different second diagnoses k_2 as combinations. By fixing $k_1 = d$ we ask for the risk of being re-hospitalized with any k_2, given that a certain d already had appeared. In the same way fixing $k_2 = d$ has multiple first diagnoses k_1. Here, we are interested in the question of what is the risk of being re-hospitalized with a specific d, given that any k_1 preceded. Finally, in case of fixed specialities we analyse the re-hospitalization risk given that contact to some specialist

took place no matter which diagnoses were involved. We use all predicted responses for a given set of fixed diagnoses and specialist contacts to determine the median of the response distribution together with the standard error of the mean.

In the following, we define the medians for all relevant cases of contact independence and dependence used in the results chapter. All medians will be indicated by m (do not confuse with the memory analysis m!) with subscripts and superscripts according to the specific case. Subscripts M and F denote male and female data, superscripts 1 and 2 stand for fixed first and second diagnosis, respectively.

Contact-independent averages

In the contact-independent case we use an average value for the predicted responses for male and female risk. With equations (3.6) and (3.7) we define the medians for fixed k_1 or k_2 separately for males and females as

$$m^1_{\hat{P}_{M,k},d} = median\big(\hat{P}_{M,k}\big)_{k_1 = d} \quad , \tag{3.16}$$

$$m^2_{\hat{P}_{M,k},d} = median\big(\hat{P}_{M,k}\big)_{k_2 = d} \quad , \tag{3.17}$$

$$m^1_{\hat{P}_{F,k},d} = median\big(\hat{P}_{F,k}\big)_{k_1 = d} \quad , \tag{3.18}$$

$$m^2_{\hat{P}_{F,k},d} = median\big(\hat{P}_{F,k}\big)_{k_2 = d} \quad . \tag{3.19}$$

Secondly, we calculate an average value for the relative risk of males to females. We use equation (3.8) and determine the medians for k_1 and k_2 as

$$m^1_{logRR_k,d} = median\big(logRR_k\big)_{k_1 = d} \quad \text{and} \tag{3.20}$$

$$m^2_{logRR_k,d} = median\big(logRR_k\big)_{k_2 = d} \quad . \tag{3.21}$$

Contact-dependent averages

In the contact-dependent case we include the information about specialist contacts. We have created separate regression models for each specialist and need to also average over all specialities to obtain a median for one fixed diagnosis. In equations (3.13) and (3.14) we have used the predicted responses for male and female risk with and without contact to compute logarithmic ratios. Now we average over all i and fixed diagnosis d separately for males and for females and obtain the medians

$$m^1_{logR_{M,k,i},d} = median\big(logR_{M,k,i}\big)_{i,k_1 = d} \quad , \tag{3.22}$$

$$m^2_{logR_{M,k,i},d} = median\big(logR_{M,k,i}\big)_{i,k_2 = d} \quad , \tag{3.23}$$

$$m^1_{logR_{F,k,i},d} = median\big(logR_{F,k,i}\big)_{i,k_1 = d} \quad , \tag{3.24}$$

$$m^2_{logR_{F,k,i},d} = median\big(logR_{F,k,i}\big)_{i,k_2 = d} \quad . \tag{3.25}$$

We repeat the averaging over all i in equation (3.15) to obtain the logarithmic relative risk ratio and calculate the medians

$$m^1_{logRRR_{k,i},d} = median\big(logRRR_{k,i}\big)_{i,k_1\,=\,d} \quad \text{and} \tag{3.26}$$

$$m^2_{logRRR_{k,i},d} = median\big(logRRR_{k,i}\big)_{i,k_2\,=\,d} \quad . \tag{3.27}$$

In contrast to the contact-independent case we can compute medians for a fixed speciality S that can take values $1, \ldots, 18$ for each of the selected specialists. We use equations (3.13) and (3.14) again, but instead of specifying k_1 or k_2 we fix the speciality i to take a specific value $i = S$ and we obtain the medians

$$m_{logR_{M,k,i},S} = median\big(logR_{M,k,i}\big)_{i\,=\,S} \quad \text{and} \tag{3.28}$$

$$m_{logR_{F,k,i},S} = median\big(logR_{F,k,i}\big)_{i\,=\,S} \quad . \tag{3.29}$$

Standard error of the mean

For all previously calculated medians we use the standard error of the mean SE as the measure of uncertainty. It is calculated as the standard deviation σ of the underlying distribution divided by the square root of the size of the sample n_{sample},

$$SE_{mean} = \frac{\sigma}{\sqrt{n_{sample}}} \quad . \tag{3.30}$$

3.3.9 Diagnosis prevalence

Until now, we have estimated the relative re-hospitalization risks between males and females using logistic regression. In addition, we correlate the so predicted re-hospitalization risk to the sex prevalence of diagnoses. We use the number of males and females who did and did not receive specific diagnoses (either as principal or secondary diagnosis), defined in section 3.3.5 using matrix M_{diag}. We define the logarithmic sex odds ratio $logodds_d$ for every d in the diagnosis sample as

$$logodds_d = log\left(\frac{\frac{n_{M,d}}{n_{M,\neg d}}}{\frac{n_{F,d}}{n_{F,\neg d}}}\right) \quad . \tag{3.31}$$

As a measure of uncertainty, we use an approximation of the standard error for the logarithmic odds ratio [34],

$$SE_{logodds} = \sqrt{\frac{1}{n_{M,d}} + \frac{1}{n_{M,\neg d}} + \frac{1}{n_{F,d}} + \frac{1}{n_{F,\neg d}}} \quad . \tag{3.32}$$

Chapter 4

Results

The following chapter presents the results of our analysis. We start with the outcome of the memory analysis of the HCP system which is divided into two parts: the system memory analysis and the age-dependent memory analysis. Secondly, we present the results of network motif analysis. Motifs are further studied with focus on sex-specific re-hospitalization risk and on sex-specific diagnose prevalence. To provide an overview of the underlying network, a graphical visualization of the patient flows in the bipartite network of specialists and diagnoses is presented.
All data processing, calculations and plot visualizations were carried out in the computing environment MATLAB. The network visualization was created using the open source network analysis tool Gephi.

4.1 Random walk model

The random walk model provides an answer to the question of the path dependence of patients' movements in the network. Using the information of the HCP-system, transition matrices for specified patient samples and various numbers of given previous steps were created and used in a random walk model. Prediction quality is measured in terms of percentages of correctly predicted steps out of a series of model runs. First we examine the system's memory in general and then focus on smaller patient samples based on age groups.

4.1.1 System memory analysis

To determine the general system memory we compare predictive quality for varying numbers of given previous steps. In total, three transition matrices are created, two with separate female and male samples and a third with combined data of both sexes. Data of patients of all ages is included. Figure 4.1 shows the analysis results. For each number of past steps m on the x-axis the mean percentage μ of correct predicted last steps out of 1000 random paths is depicted on the y-axis. The percentages are based on averages out of 100 model repetitions. Each point in the figure corresponds to the mean of the distribution together with the standard deviation as a measure of uncertainty.

We observe an increase of correct predictions with raising number of given steps for both sexes. Differences in quality of prediction for combined or separate transition matrices are minor and lie within the standard deviation for both, females and males. For females, we observe a minimum of correct predictions μ for one given previous step with $(49,5 \pm 1,8)\%$ with the separate transition matrix and $(49,5 \pm 1,8)\%$ with the combined transition matrix and maximal correct predictions at five given steps with each $(58,8 \pm 1,6)\%$ and $(58,4 \pm 1,5)\%$. For males, the minimum of correct predictions is again for one given step with $(51,5 \pm 1,7)\%$ with the separate transition matrix and $(51,9 \pm 1,6)\%$ with the combined transition matrix and maxima for five given steps at $(60,3 \pm 1,7)\%$ and $(59,5 \pm 1,6)\%$, respectively. A full list of all results depicted in the plot can be found in the table 4.1.

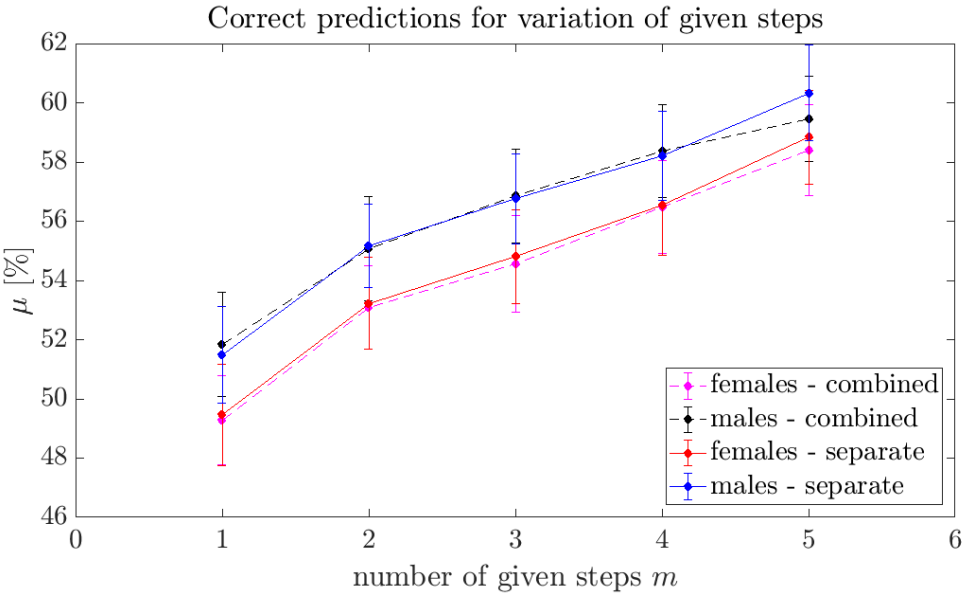


Figure 4.1: Results of system memory analysis. The prediction quality for various given steps m is tested separately for females and males and for different underlying transition matrices. The dashed lines represent the results for a combined underlying transition matrix, full lines for separate transition matrices.

given steps	separate matrices [%]		combined matrix [%]	
	F	M	F	M
1	49,5 ± 1,8	51,5 ± 1,7	49,3 ± 1,8	51,9 ± 1,6
2	53,2 ± 1,6	55,2 ± 1,5	53,1 ± 1,8	55,1 ± 1,5
3	54,8 ± 1,6	56,8 ± 1,6	54,6 ± 1,6	56,9 ± 1,7
4	56,5 ± 1,7	58,2 ± 1,6	56,5 ± 1,6	58,4 ± 1,6
5	58,8 ± 1,6	60,3 ± 1,7	58,4 ± 1,5	59,5 ± 1,6

Table 4.1: : Mean percentage of correct predictions and standard deviation for system memory analysis. Results of figure 4.1. We observe an increase of prediction quality for higher numbers of given previous steps for both sexes.

4.1.2 Age-dependent memory analysis

In the age-dependent memory analysis part we carry out an identical procedure to the system memory analysis only with smaller patient samples. Patient's last steps are predicted using data-based transition matrices for individual age groups. Using the groups a_1 to a_{10} defined in section 3.2.6, we create three transition matrices again, two with separate sex information, one with combined information. The results of the age-dependent memory analysis can be found in figure 4.2.

The mean percentage of correct predictions μ and the standard deviation are depicted on the y-axis separately for females (red) and males (blue) for all age groups shown on the x-axis. The results are divided into tow columns: the results for predictions with separate transition matrices are depicted on the left, results for predictions with combined transition matrices are shown in the right column. We analyse predictions with $m = 1$ to $m = 4$ given steps. We observe a similar effect to the results of the system memory analysis: prediction quality is increased if more previous steps are included in the process of creating the transition matrix.

We see that predictive quality varies for different age groups as well as it differs for females and males in some age groups. Tables with all depicted results can be found in the appendix.

In all above cases we observe that predictive quality increases with the number of given steps. The process of patients moving within the HCP system is therefore path-dependent and non-Markovian. The percentages of correct predictions for separate and combined transitions matrices are similar within the uncertainties, no significant information is gained when using sex-specific transitions. Sex differences in predictive quality can be found for different age groups, mostly for patients born between 1957 and 1996.

4.2 Analysis of network motifs

Next, we identify network motifs as varying sequences of first diagnosis, specialist contact and second diagnosis. We use network motifs to study the sex-specific risk of being re-hospitalized when certain diagnoses or specialists are involved in the motif. First, we use the regression model predictions to correlate the risk of males to the risk of females for fixed k_1 and k_2 and for fixed specialities i. Secondly, we show the correlation between diagnosis prevalence and relative risk of males to females. To demonstrate the effect of contacts on the risk (or relative risk), results of contact-independent and contact-dependent analysis are depicted in combined figures. Notice that in the following all calculated medians will be denoted with m and appropriate super- and subscripts and should not be confused with the number of steps from the memory analysis part.

4.2.1 Motif: Sex-specific re-hospitalization risk

First, we compare male and female re-hospitalization risk. In the contact-independent case we scatter the medians (3.16), (3.17) and (3.18), (3.19) against each other. For the contact-dependent results, we plot (3.22), (3.23) vs. (3.24), (3.25). Results for fixed k_1

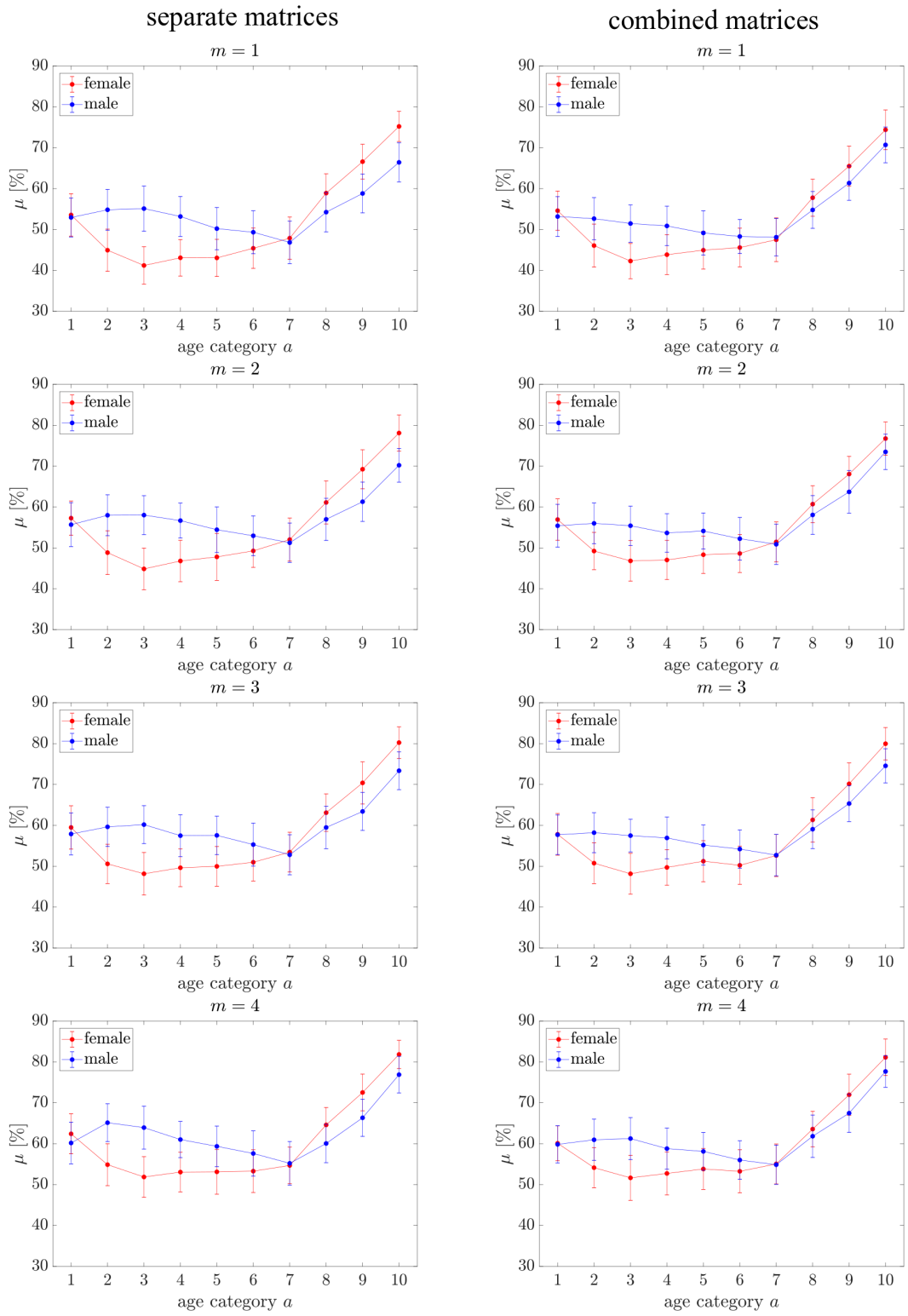


Figure 4.2: Results of age-dependent memory analysis. Female and male mean percentage of correct predictions μ are shown for different age categories a and for varying numbers of given steps m. The left column of plots corresponds to the results with separate transition matrices, the right column to results with a combined transition matrix for females and males.

and k_2 are depicted next to each other in figure 4.3 with joint y-axis. Each point in the plots stands for a specific diagnosis with at least one predicted value from a regression model. Explanatory plots describe the meaning of the quadrants in some of the following cases. The coloured background of the quadrants is equally applied to the real results and simplifies the interpretation of single points.

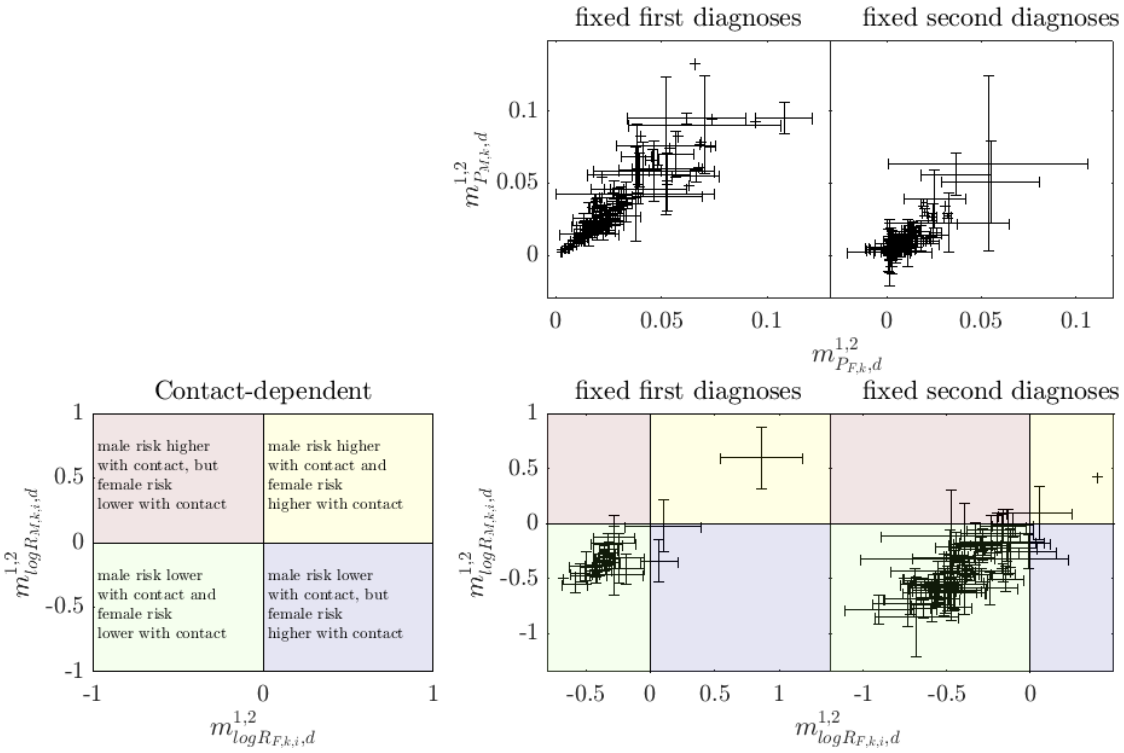


Figure 4.3: Motif analysis: male vs. female re-hospitalization risk. Top: Contact-independent medians for re-hospitalization risk for males vs. females for fixed diagnoses. Bottom: Explanatory plot and contact-dependent medians of logarithmic ratios of males and females. Each point in the scatter plots stands for a fixed first diagnosis (left) or fixed second diagnosis (right).

In the top part of figure 4.3 the contact-independent results are shown. The range for the medians $m^{1,2}_{P_{M,k},d}$ and $m^{1,2}_{P_{F,k},d}$ lies within [0, 1]. Zero stands for no risk at all, one means one hundred percent risk. In the bottom part of the figure, contact-dependent results are shown. Here we compare the effect of having a contact to not having a contact for both sexes. In this case, the medians $m^{1,2}_{P_{M,k,i},d}$ and $m^{1,2}_{P_{F,k,i},d}$ lie in $\mathbb{R}$. Negative values mean lower re-hospitalization risk if patients had a contact to any specialist, positive values mean higher re-hospitalization risk with contact.

To detect effects of specific specialities on the predicted re-hospitalization risk, we fix the speciality $i = S$ and average over all diagnosis combinations k. The resulting medians for male and female logarithmic ratios for fixed specialities $m_{P_{M,k,i},S}$ and $m_{P_{F,k,i},S}$ are shown in figure 4.4. Their values range in $\mathbb{R}$. Notice that there is no distinction between first and second diagnosis, the averaging runs over all combinations. Each speciality is depicted as a color-coded point. The errorbars stand for the standard error of the mean (3.30) of every median value.

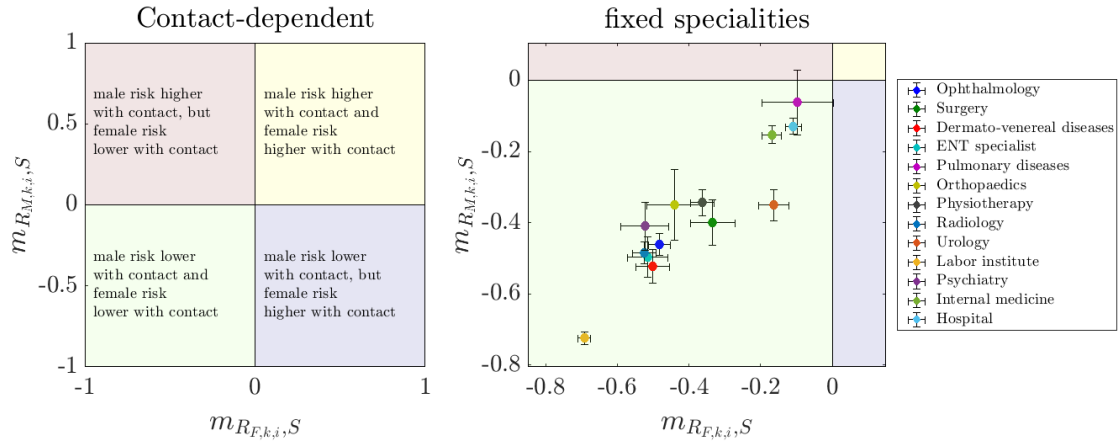


Figure 4.4: Fixed specialities. Explanatory plot (left) and contact-dependent logarithmic ratios for males vs. females with median for fixed specialities. Each color-coded point stands for a specialist from the list on the right.

4.2.2 Motif: Sex-specific diagnosis prevalence

We further specify the properties of diagnoses and assign each diagnosis with an additional value for their sex prevalence. We correlate this prevalence in terms of log-odds $logodds_d$ for every diagnosis d to the relative risk, respectively to the relative risk ratios of males to females. In the contact-independent case the logarithmic ratio $logRR$ simply states whether the male or female risk is higher. In the contact-dependent case, the relative risk ratio $logRRR$ is not a simple ratio of predicted risks any more. In $logRRR$ we compare the risk alteration for males due to contact to the risk alteration for females due to contact. This is explained in more detail in the explanatory plots of figure 4.5. The range of values on both axes in the plots is $\mathbb{R}$. Negative $logodds_d$ mean female-, positive $logodds_d$ male-dominated disease risks. We scatter diagnosis log-odds vs. medians of the relative risks (3.20), (3.21) in the contact-independent case (top part), and vs. the relative risk ratios (3.26), (3.27) in the contact-dependent case (bottom part). Every point in the scatter plots stands for a fixed diagnosis with at least one predicted value from a regression model. The y-errorbars are the standard error of the $logodds_d$ calculated according to equation (3.32) and the x-errorbars depict the standard error of the mean of the distribution using (3.30).

We find that motifs as defined by patterns of various hospitalizations exist in the HCP network. Using regression analysis we find that the re-hospitalization risk behaviour is similar for both sexes. Contact to specialists has a lowering effect on the re-hospitalization risk for almost all diagnoses and all specialities. Diagnosis sex prevalence only shows a direct correlation to relative risk of males to females in case of fixed k_2 and contact independence. In all other cases no clear tendency can be found, but there exist diagnoses with sex prevalence for one sex but higher risk alteration for the opposite sex (fixed first diagnoses in bottom part of figure 4.5). All figures shown in this section represent results for the main parameter setting defined in table 3.3. The robustness test for varying parameter settings can be found in the appendix.

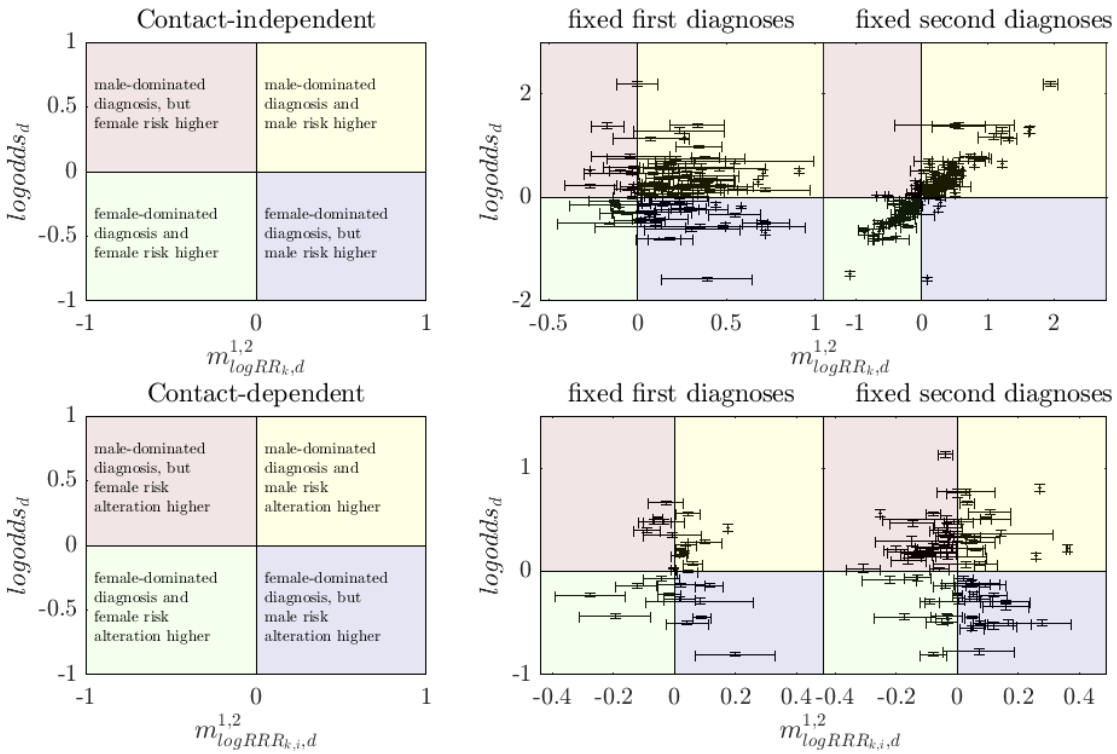


Figure 4.5: Motif analysis: diagnosis prevalence vs. relative risk. Top: Explanatory plot and contact-independent results of diagnosis prevalence vs. logarithmic relative risk. Bottom: Explanatory plot and contact-dependent results of diagnosis prevalence vs. logarithmic relative risk ratios. Each point stands for a fixed diagnosis, fixed k_1 left, fixed k_2 right.

4.3 Graph visualization

4.3.1 Patient flows

For an overview on the bipartite network of diagnoses and specialists we create a visualization of the underlying system. In this network, diagnoses and specialities are connected through patient flows. To define the flux and create the bipartite adjacency matrix, we use the data matrices from the network motif analysis, in specific from the contact-dependent case. The direction is naturally given through the definition of contact-dependent motifs: patients start with a first diagnosis k_1, then visit a specialist and continue to a second diagnosis k_2. The flows are further characterized by weights according to the sex prevalence in each patient flow. To define the weight, we calculate separate percentages for males and females involved in a flow and use their ratio as the final weight.

A detailed description of this analysis part can be seen in the appendix, together with the calculation of sex prevalence of each flow. Colormaps of the percentages of patients per diagnosis with specific contacts to specialities and for various diagnoses provide deeper insights to re-hospitalization processes and to the HCP network in general.

4.3.2 Network visualization

To create a meaningful visualization of the bipartite network, the number of links must be reduced. We are interested in the most relevant connections in terms of strength, but we do not want to loose any node that is somehow connected to the network. We therefore combine two approaches for network backboning: the disparity filter and the maximum spanning tree.
The disparity filter algorithm is used to ensure that nodes with small strength are not eliminated, even though not all nodes must survive the process. By choosing a cut-off parameter, one can selectively filter out links [46] [12]. To avoid isolated nodes, the maximum spanning tree is used in combination with the disparity filter. Each link, that either fulfils the disparity criteria or is part of the maximum spanning tree, is part of the final network visualization. By this means the visualization complexity can be reduced.
The disparity filter is applied with a cut-off at significance level $\alpha = 0.2$. The resulting bipartite network visualization is shown in figure 4.6. Link colors describe the sex prevalence of each link. Values of sex prevalence are normalized to a symmetric distribution around zero. We group single diagnosis nodes by their first letter of the ICD10 code and observe several categories like mental diseases or diseases of the respiratory system. For the purpose of simplification, the general practitioner is excluded in this analysis, as there is no specialization linking him to any class of diseases (and re-hospitalizations). For almost all diagnoses, nearly 100% of patients had contact to the general practitioner. Naturally, gynaecologists are not presented in this visualization as well, as there are 0% of males (apart from very few incidents) engaging in this contact and no ratio can be calculated. Similarly, other specialists do not show up in the network visualization as well, as there are not enough patients who had contact and the cut-off parameters were therefore not satisfied.

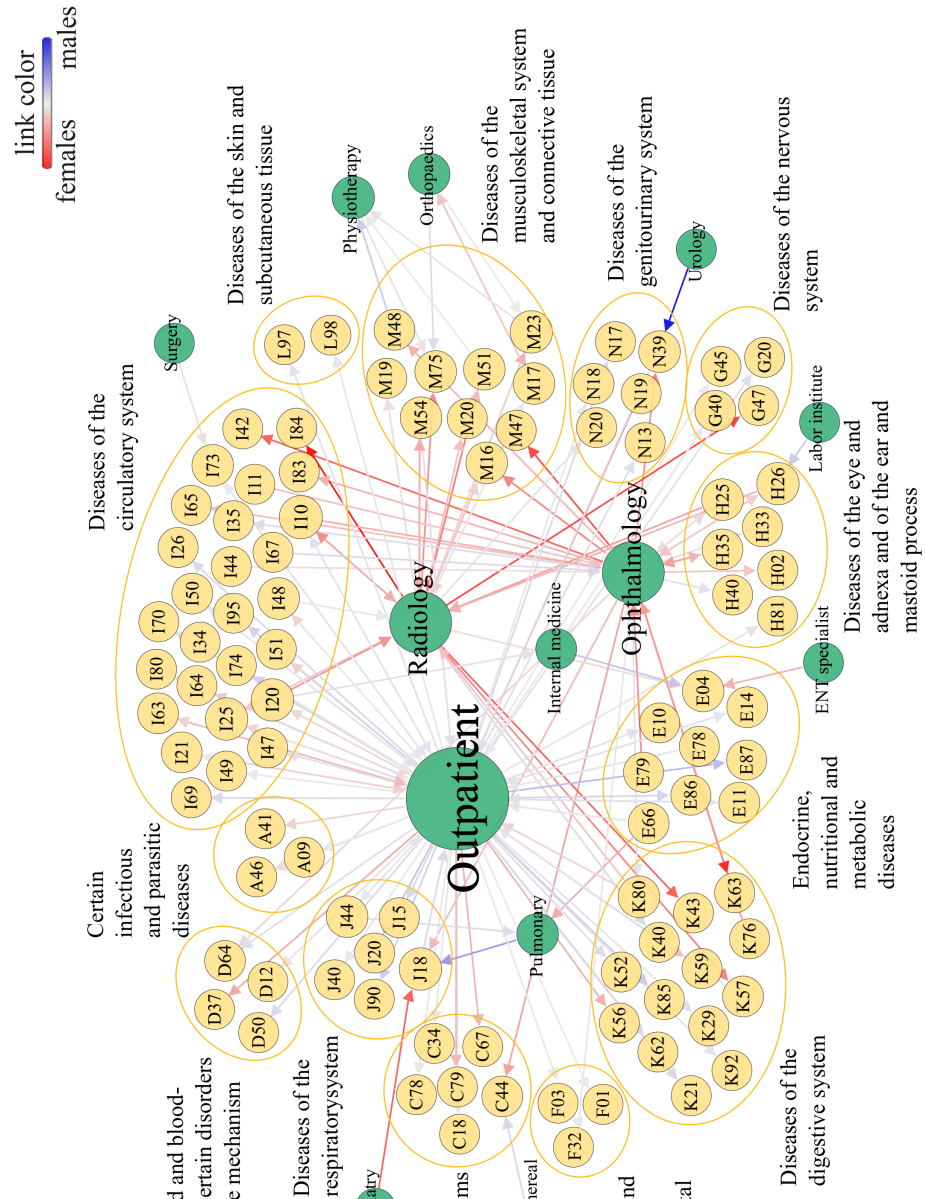


Figure 4.6: Visualization (created in Gephi) of the directed bipartite network of specialists and diagnoses. Links describe patient flows between the two types of nodes. Links are coloured depending on the ratio of male percentage to female percentage of patients with a connection. Node sizes scale with the in-degree of nodes. Link colors were scaled to a symmetric distribution around zero. The outpatient node stands for the contact to the hospital.

Chapter 5

Discussion

The irrevocable complexity of interactions in healthcare systems [50] and the importance of care coordination (e.g. when and how patients receive therapy) has only lately been recognized. So far two bottlenecks existed: One was the missing scope of data concerning patient and diagnose information, the other one was the lack of methods to deal with path-dependent processes on complex networks. Now the data exists and we can develop empirical methods for the first time. By identifying a variety of interacting structures in healthcare systems, we can show how improved care coordination can result in better population health. Using concepts for the study of complex systems, that evolve dynamically depending on relationships and interaction patterns, improvements can be made to patient-centred care [50] and in understanding the interplay of multiple care providers with respect to their impact on population health [32]. In this work, we attempted to show that new approaches from the study of complex systems, combined with physics inspired modelling of dynamical processes on networks, can lead to new insights in healthcare systems.

Can the diffusion process of patients to physicians be described as a Markovian process? Or do we observe path dependence? And can we find specific network motifs suggesting sex biases in patients' treatment paths? Several studies were performed in the last years to analyse the structure and features of various healthcare systems of the world [41] [26]. Examples range from the patient sharing networks in the US, suggesting that females tend to have deviating access to medical treatments than males [3], to big data exploration of an Austrian health claims data set, used to create networks and find so far unknown correlations between diseases and age and sex factors [45] [24]. In an attempt to tackle similar questions and using a medical claims data set of the Austrian population, we modelled the diffusion of patients as a random walk and evaluated sex-specific re-hospitalization risks associated with certain paths in this network by means of a regression analysis. In the following we will discuss the main results in detail and discuss their implications.

First we focus on the path dependence as a diffusive property of patients in the Austrian healthcare system. By comparison of predictive quality of random walk models of varying order we can say that the process of patient diffusion in the HCP network is non-Markovian (see figure 4.1). We observe an increase in mean correct predictions μ of patients' last steps when more than one previous step is used for creating the transition matrix. This is the case for both sexes. In addition, we see that μ for male patients is higher for all

tested numbers of steps. The results for separate male and female transition matrices and for combined transition matrices with information of both sexes indicate that we do not gain much information by using sex-specific transitions for prediction, even though the differences seem to be larger when we attain higher numbers of given steps. A similar memory analysis was realized also for separate age groups of the sample. In figure 4.2 we see that the predictive quality varies for different ages of patients, for increasing number of given steps and it also varies for females and males. Interestingly, we observe that μ is very similar for both sexes of young age. For older patients correct predictions for males outperform predictions for females until an age of around 60 to 70, where the curves intersect again. For ages 70 and higher, prediction quality is higher for females than for males and reaches the highest values of μ for the oldest patients. Also in the age-dependent memory analysis we see path dependence of the diffusive process. For raising numbers of given steps we observe an increase of μ. The differences between results for separate and combined transition matrices are minor, but it appears as if prediction quality for males using a combined matrix decreases with respect to the sex-specific analysis, especially between ages 20 and 50.

As a second part of this thesis we analysed the existence of network motifs, i.e. frequently observed patterns of patient flows. Using statistical tools we examined specific re-hospitalization patterns as motifs. More precisely this means finding directed connections of first diagnoses, specialist contacts and second diagnoses separately for both sexes. We find that such patterns exist for various combinations of diagnoses and specialities in the created network. To further examine these motifs we chose logistic regression analysis, which allowed us to analyse correlations between specified predictor variables and potential responses, adjusting for age. Using the motif analysis we tried to tackle two questions: 1) How do contacts to specialists affect the re-hospitalization risk? 2) Can we detect sex biases in treatment paths? The variety of possibilities to approach these questions is immense. We focused mainly on distinguishing between contact-dependent and -independent effects, as well as on effects of first hospitalization diagnoses and readmission diagnoses. We need to say in advance, that it was not objective of this analysis to describe effects of only a few chosen diseases, but rather to find general tendencies. To answer question 1), we first look at the bottom part of figure 4.3, where patients' contacts in between hospitalizations are considered. The correlation of contact-dependent re-hospitalization risks for males and for females shows linear behaviour. We observe that the majority of diagnoses (both k_1 and k_2) lies within the quadrant indicating lower re-hospitalization risk if patients had contact to specialists. In figure 4.4 the same tendency can be found if not diagnoses but specialities are fixed. Contacts to all specialists show a risk-reducing effect for both sexes. This means that seeing a doctor seems to indeed have positive effects on patients' health, at least when it comes to re-hospitalization risk. The strongest risk reduction effect is found for the labor institute, while the weakest risk reduction is assigned to pulmonary specialists, hospital and internal medicine, which, due to their nature, are more likely to lead to a re-hospitalization. Concerning question 2), we look at the results in figure 4.3. We see that without taking contact-dependent effects into account (top of figure), the predicted

risk for males correlates approximately linearly to the risk for females. The same tendency can be found in the contact-dependent case (bottom part of figure). Uncertainties of risk seem to be much smaller in the case of fixing diagnose k_1 and asking for the risk of being re-hospitalized because of any other diagnosis.

In another part of the motif analysis we want to examine the correlation of diagnosis sex prevalence to sex-specific re-hospitalization risk. Starting with the contact-independent part of the results shown in figure 4.5 (top), we see that effects for fixed k_1 clearly differ from the effects seen when fixing k_2. While for fixed k_2 the linear trend is maintained, for fixed k_1 there is no evident tendency. We interpret these results as follows: For diagnoses with prevalence for one sex, the risk of being re-hospitalized with these diagnoses is higher for this sex. Notice however, that uncertainties for first diagnoses are larger. The probably most complicated results are shown in the bottom part of figure 4.5 as it condenses information about contacts, prevalence and relative risk for both sexes into one picture. Although no clear tendency can be found, one can say that for fixed k_1 we observe a few diagnoses that are dominated by one sex but show a larger risk alteration for the opposite sex, suggesting an inverse correlation. For fixed k_2 the diagnoses form a cloud with no recognizable trend. In conclusion, results from the network motif analysis show contact-dependent effects as well as sex differences in treatment paths for different types of diagnoses, but no statement can be made about their significance and relevance to the quality of treatment patients of different sex receive.

An often not sufficiently acknowledged limitation in such data-driven work concerns data quality. Real-life big datasets are constructs of enormous amounts of messy information that need to be sorted through, cleaned and rearranged just as experimental lab data. Only after several unsuccessful attempts to give meaningful answers, we reached the point of defining concrete and very specific research questions concerning the network memory and network motifs. Despite of the large dataset provided, there exist certain boundaries and flaws. Many more characteristics of patients would need to be considered to rule out otherwise unnoticed effects. This drawback was for example also described in another sex bias health study for CVD (cardiovascular disease) prevention, stating that factors like a stressful life style and hormone levels might have impacts on the analysis [38]. The only attributes known in our dataset are age and sex of patients, together with time-stamped contacts and diagnoses. This needs to be considered when interpreting the results and infer conclusions.

In this work we developed a new method to detect sex bias in healthcare data including patients' contacts to doctors. As no reports of comparable methods was available before, the results cannot be related to any other analysis. In contrast to studies like [24] or [3], where the focus is either set on specific diseases or on more detailed information about physicians and patients, our work examines the global nature of the underlying network and seeks to identify system-level effects in the data. We can point out to possible disadvantages and criticism of the implemented analysis. All results were calculated using data and samples that obey some chosen cut-off parameters. We are aware that choosing different cut-off settings for age range, time windows, number of patients per diagnose etc. might

affect the results. We therefore performed a robustness test for three varying parameters. As can be seen in the appendix, most effects remain robust across the considered cut-off settings. Obviously the choice of specialities and diagnoses for analysis could be varied as well. To provide medicinally relevant statements to the shown results, we collaborated with a team of specialists for internal medicine. A publication concerning the here presented method and results and providing a detailed medical interpretation is in progress. We hope that the methods and ideas used in this work can be applied to similar problems of big data analysis and improve the understanding of complex structures in & beyond healthcare.

Appendix

Random walk model predictions

The following tables show results of mean correct percentage of predictions and standard deviations for the modelling of last steps for the age-dependent memory analysis.

age category	separate matrices [%]		combined matrix [%]	
	F	M	F	M
1	54,4 ± 5,0	53,5 ± 5,1	54,0 ± 4,7	53,6 ± 5,3
2	44,7 ± 5,0	54,0 ± 4,7	46,3 ± 5,1	52,5 ± 5,5
3	40,4 ± 4,8	53,6 ± 4,6	43,0 ± 5,1	51,0 ± 5,0
4	41,7 ± 4,4	52,6 ± 5,2	43,0 ± 5,1	50,0 ± 5,2
5	43,9 ± 5,3	49,5 ± 5,3	44,0 ± 4,9	49,1 ± 5,2
6	44,1 ± 5,5	49,9 ± 5,5	46,0 ± 5,6	47,9 ± 4,9
7	47,6 ± 5,1	47,5 ± 5,8	48,8 ± 5,3	48,0 ± 5,2
8	57,1 ± 4,6	53,4 ± 4,8	58,2 ± 5,3	54,4 ± 4,9
9	66,2 ± 4,7	58,9 ± 5,2	65,5 ± 4,3	62,5 ± 4,6
10	76,0 ± 4,5	66,6 ± 4,8	74,7 ± 4,6	70,1 ± 4,9

Table 5.1: : One given step. Mean percentage of correct predictions and standard deviation for age dependent memory analysis with one given step. Results of first row in figure 4.2

age category	separate matrices [%]		combined matrix [%]	
	F	M	F	M
1	57,2 ± 5,0	56,0 ± 4,8	57,2 ± 5,0	56,2 ± 5,4
2	48,3 ± 5,0	57,8 ± 4,8	49,6 ± 5,1	55,9 ± 4,4
3	45,5 ± 4,8	58,6 ± 5,4	46,5 ± 5,3	55,5 ± 5,0
4	46,5 ± 5,3	56,4 ± 4,7	48,0 ± 5,0	54,0 ± 5,3
5	47,1 ± 4,3	55,6 ± 4,4	49,0 ± 4,8	54,3 ± 4,7
6	48,6 ± 4,9	53,5 ± 4,5	49,0 ± 5,6	52,3 ± 5,0
7	51,2 ± 5,0	52,1 ± 5,7	52,1 ± 5,5	51,4 ± 4,7
8	60,4 ± 5,0	57,2 ± 4,9	59,9 ± 5,9	57,9 ± 5,0
9	68,7 ± 4,7	62,1 ± 4,4	68,3 ± 4,5	64,3 ± 4,5
10	77,4 ± 4,8	70,5 ± 4,4	77,5 ± 4,8	72,3 ± 4,9

Table 5.2: : Two given steps. Mean percentage of correct predictions and standard deviation for two given steps. Results of second row in figure 4.2

age category	separate matrices [%]		combined matrix [%]	
	F	M	F	M
1	59,4 ± 5,1	57,6 ± 4,6	57,7 ± 5,1	58,5 ± 5,5
2	50,9 ± 5,1	60,8 ± 4,9	51,8 ± 4,9	58,2 ± 5,7
3	47,9 ± 4,9	59,7 ± 5,2	48,3 ± 5,1	57,9 ± 4,4
4	49,4 ± 5,0	58,2 ± 4,7	50,1 ± 5,2	56,5 ± 5,0
5	49,8 ± 4,7	57,4 ± 5,3	50,1 ± 4,9	55,8 ± 4,5
6	51,5 ± 5,1	56,0 ± 5,6	50,8 ± 5,5	54,5 ± 5,2
7	53,3 ± 5,4	52,2 ± 5,3	53,2 ± 5,2	53,0 ± 5,5
8	61,8 ± 4,8	59,0 ± 4,8	61,5 ± 5,2	58,2 ± 4,7
9	69,9 ± 5,1	63,9 ± 4,8	69,5 ± 4,5	65,5 ± 4,2
10	80,4 ± 4,0	73,9 ± 4,5	80,4 ± 4,3	75,1 ± 4,8

Table 5.3: : Three given steps. Mean percentage of correct predictions and standard deviation for three given steps. Results of third row in figure 4.2

age category	separate matrices [%]		combined matrix [%]	
	F	M	F	M
1	63,3 ± 4,7	60,8 ± 4,8	60,9 ± 4,9	60,0 ± 4,9
2	54,1 ± 5,2	64,1 ± 5,0	54,8 ± 4,8	61,0 ± 5,3
3	51,2 ± 4,9	64,1 ± 4,7	51,3 ± 5,0	61,6 ± 5,1
4	53,8 ± 5,6	62,6 ± 5,0	52,6 ± 5,2	58,3 ± 4,7
5	53,3 ± 4,9	59,5 ± 4,7	53,3 ± 5,0	57,9 ± 5,1
6	54,3 ± 4,4	58,3 ± 5,0	53,5 ± 5,4	56,6 ± 5,0
7	55,1 ± 4,4	55,3 ± 5,6	54,6 ± 4,4	55,0 ± 5,3
8	65,5 ± 4,8	59,4 ± 5,3	63,5 ± 5,3	60,7 ± 4,9
9	72,7 ± 4,4	65,9 ± 5,1	71,2 ± 4,3	67,6 ± 4,4
10	82,0 ± 3,8	77,5 ± 4,3	81,6 ± 4,3	77,6 ± 4,2

Table 5.4: : Four given steps. Mean percentage of correct predictions and standard deviation for four given step. Results of forth in figure 4.2

Boxplots

In the contact-dependent motif analysis, we can depict the response distributions for fixed diagnoses with boxplots. For every fixed diagnosis d on the x-axis the median of the distribution is shown as a red line, the box corresponds to the 25th and 75th percentile and the whiskers reach to all other data points not considered outliers. Red crosses mark outliers to the distribution (see figures 5.1 and 5.2).
We carried out a sign test to verify if the median of distributions significantly differs from zero. A star above diagnoses indicates that they have distributions with a median significantly (p-value = 0.05) different from zero. All other diagnoses failed to reject the null hypothesis.

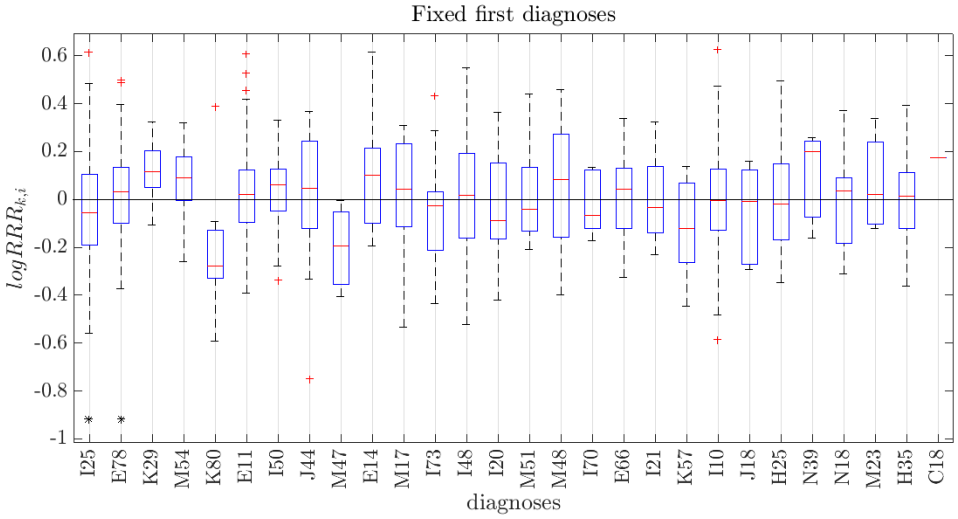


Figure 5.1: Boxplot of $logRRR_{k,i}$-distributions for fixed first diagnoses.

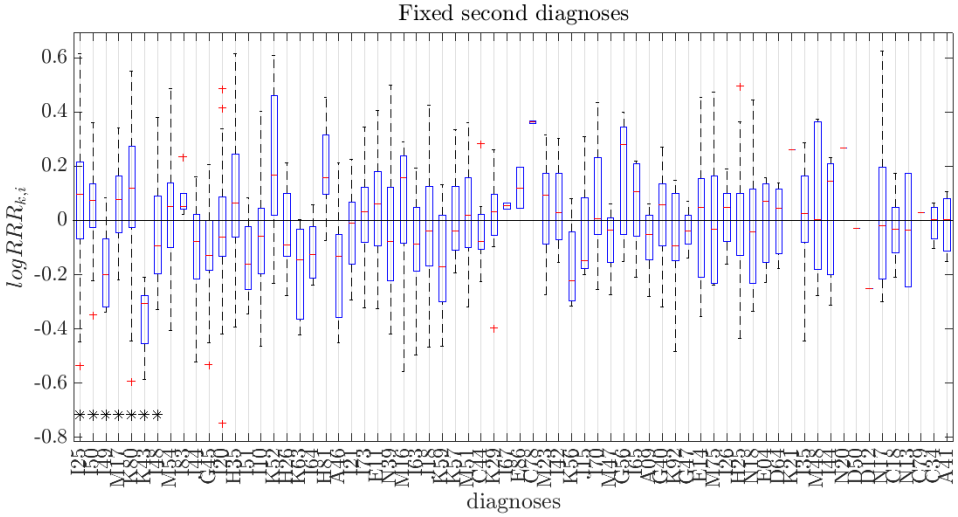


Figure 5.2: Boxplot of $logRRR_{k,i}$-distributions for fixed second diagnoses.

Physician specialities

From a total of 45 different specialities in the original dataset, we worked with 17. The general practitioner and paediatrics were combined as they provide primary healthcare to patients of different age. The three separate specialities psychotherapy, clinical psychology and psychotherapy and clinical psychology were united because of their similar character. The other 15 specialities were selected for their importance in re-hospitalization analysis and for sufficient amount of data available. The hospital, as a tertiary healthcare provider, provides information about diagnoses and is included in the list of specialities as number 18.

Number	Specialty
1	General practitioner and paediatrics
2	Ophthalmology
3	Surgery
4	Dermato-venereal diseases
5	Obstetrics and gynecology
6	ENT specialist
6	Pulmonary diseases
8	Neurology
9	Orthopaedics
10	Physiotherapy
11	Radiology
12	Accident surgery
13	Urology
14	Labour institute
15	Psychotherapy and clinical psychology
16	Psychiatry
17	Internal medicine
18	Hospital

Table 5.5: : List of all remaining specialities.

Colormap networks

For an overview on the patient flows in the bipartite diagnosis–speciality network, the percentages of females $p_{F,d,i}$ and males $p_{M,d,i}$ involved in the flows are calculated. For every fixed diagnosis d and speciality i we sum the number of males and females with and without contact of all compatible combinations k. Compatible in this case means either $d = k_1$ for fixed first diagnosis or $d = k_2$ for fixed second diagnosis. In the end we compute the mean percentage $\overline{p}_{d,i}$ of patients (males and females) involved in the flow with speciality i and diagnosis d as

$$\overline{p}_{d,i} = \frac{p_{M,d,i} + p_{,Fd,i}}{2} \quad .$$

Only diagnosis combinations k that fulfil the selected conditions for cut-off (main parameter setting; this means at least 50 male and 50 female patients per flow) are included in this estimation. For the sex prevalence of each patient flow we use the above percentages of females and males and obtain

$$preval_{d,i} = log\Big(\frac{p_{M,d,i}}{p_{F,d,i}}\Big) \quad .$$

The processes in this network can be divided into two stages: In case of fixed first diagnoses the direction of the flow is from k_1 to specialist, for fixed second diagnoses from specialist to k_2. Only diagnoses with at least one non zero $\overline{p}_{d,i}$ flow are depicted in the following plots. The resulting colormaps for first and second diagnoses are shown in figures 5.3 and 5.4 (the colouring of the matrices should not be confused with the network visualization as it is not normalized to a symmetric distribution around zero). Using these colormaps we see how patients with different diagnoses move to specialities. For fixed k_1 we observe that patients with diagnosis I10 (Essential (primary) hypertension) have the highest variety of specialists they contact after the diagnosis. In general, the ophthalmologist and radiologist are the most visited specialists, together with contacts to the hospital. Looking at the sex prevalence of flows we see most prominently the male dominated patient flows to the urologist and the female dominated flows to the psychiatrist and radiologist. For fixed k_2 the flow is defined in the direction from speciality to second diagnosis. In percentages of patients involved, radiology and hospital contacts stand out. In terms of sex prevalence, most of the connections are female dominated, except for flows from the urologist and pulmonary diseases.

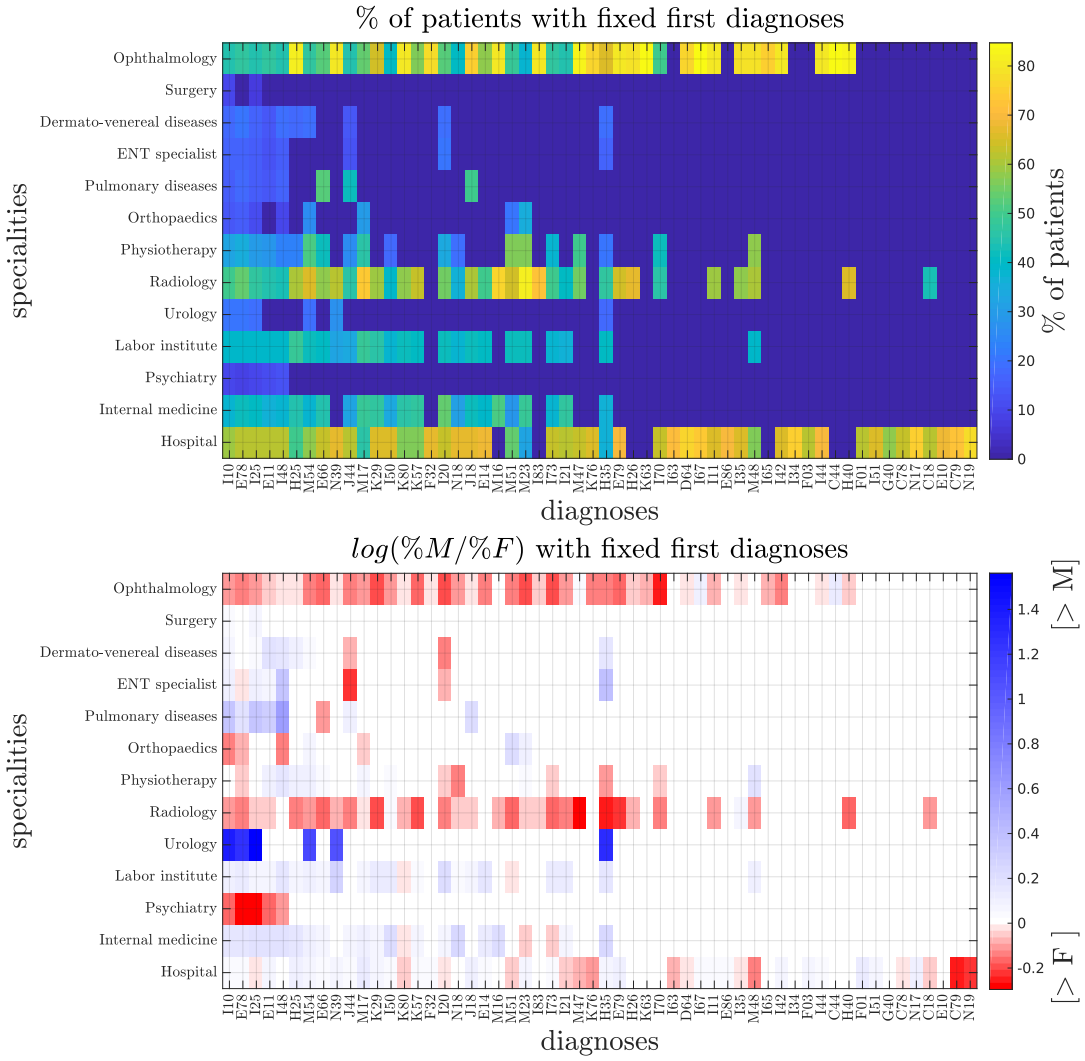


Figure 5.3: Colormap for fixed first diagnosis. Direction of flow: from diagnosis to speciality. Top map shows the percentage of patients involved in the flow from first diagnosis to a speciality, bottom map the sex prevalence of each link.

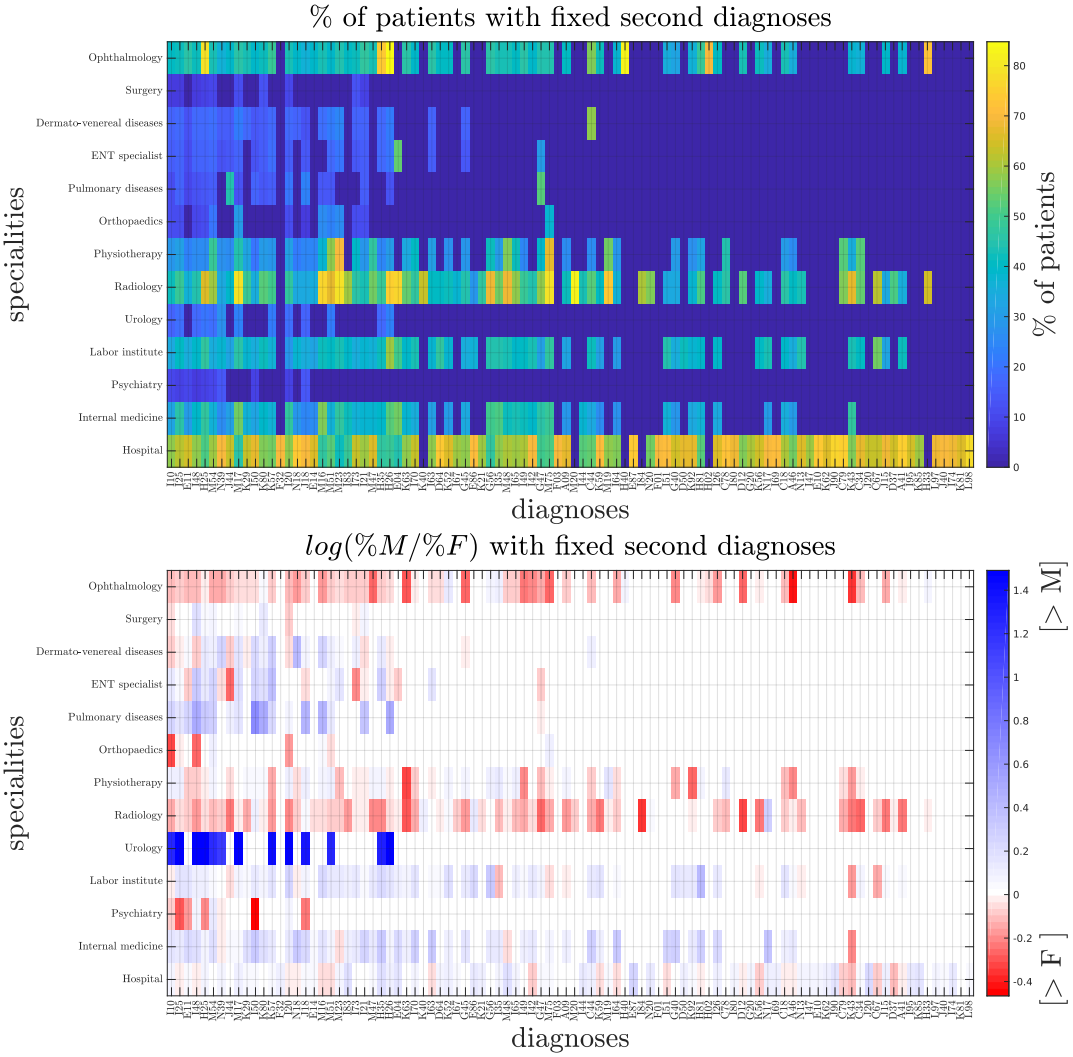


Figure 5.4: Colormap for fixed second diagnosis. Direction of flow: from speciality to diagnosis. Top map shows the percentage of patients involved in the flow from speciality to second diagnosis, bottom map the sex prevalence of each link.

Robustness test

Finally, we test the robustness of the results received in the logistic regression analysis. To do so, we vary three of the parameters set before. For each test one parameter is modified while keeping all others constant to the main parameter setting. First we vary the minimal age (or year of birth) of patients, then the time window in between hospitalizations and finally the minimum cut-off parameter for females and males with a diagnosis combination/contact.
In the following sections we will only show the defined parameters and resulting plots, all steps were already described in the methods chapter and are applied in the same way.

Age variation

We start with varying the setting of minimal year of birth, so that we allow patients of lower age in the samples. This leads to a larger sample size of diagnoses and patients involved in the regression model. We only vary minimal year of birth, as it is not meaningful to analyse the small patient sample with age over 100 years (above the age of 100, the dataset might include miss-registered patients). Table 5.6 shows the parameter settings, figures 5.5, 5.6 and 5.7 depict the resulting plots.

Parameter	Setting
minimal year of birth y_{min}	1906
maximal year of birth y_{max}	2006
minimal time window Δt_{min}	90 days
maximal time window Δt_{max}	1050 days
minimal number of females and males per combination cut_{comb}	50
minimal number of patients per diagnosis cut_{diag}	1000

Table 5.6: Analysis parameters for robustness test with age variation.

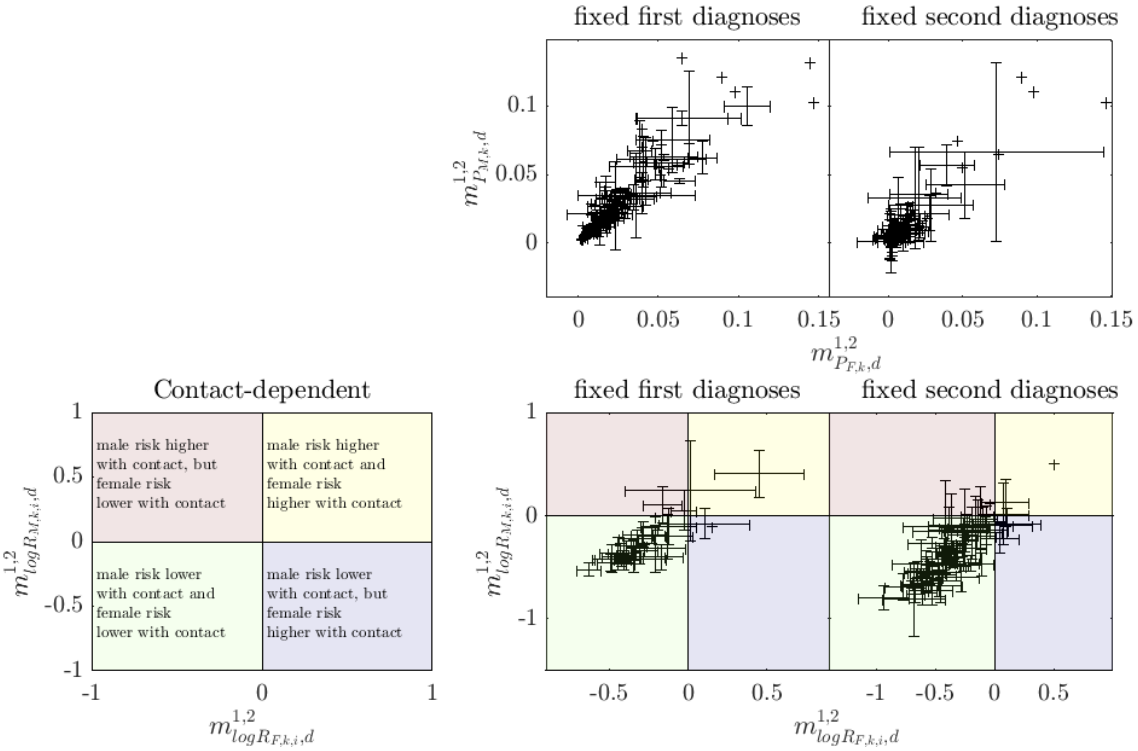


Figure 5.5: Male vs. female risk. Top: Contact-independent medians for re-hospitalization risk for males vs. females for fixed diagnosis. Bottom: Explanatory plot and contact-dependent medians of logarithmic ratios of males and females. Each point in the scatter plots stands for a fixed first diagnosis (left) or fixed second diagnosis (right).

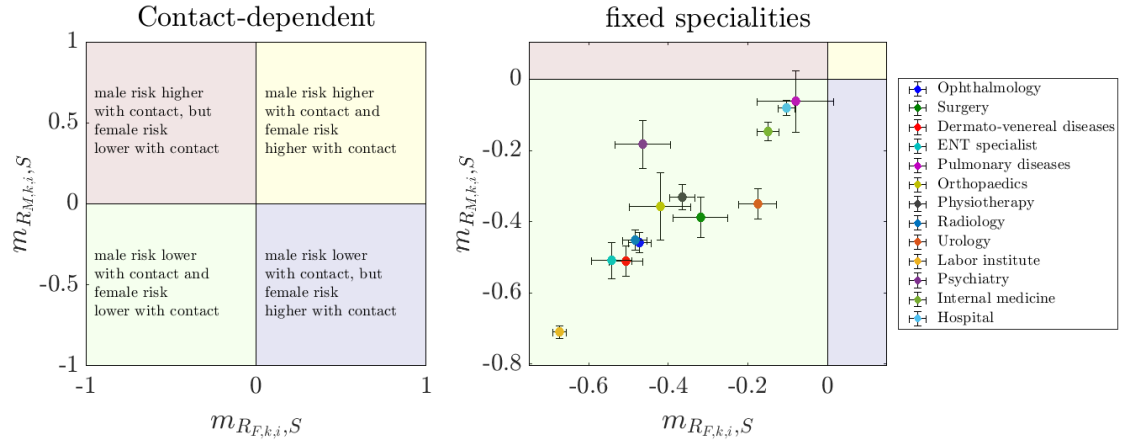


Figure 5.6: Fixed specialities. Explanatory plot (left) and contact-dependent logarithmic ratios for males vs. females with median for fixed specialities. Each color-coded point stands for a speciality from the list on the right.

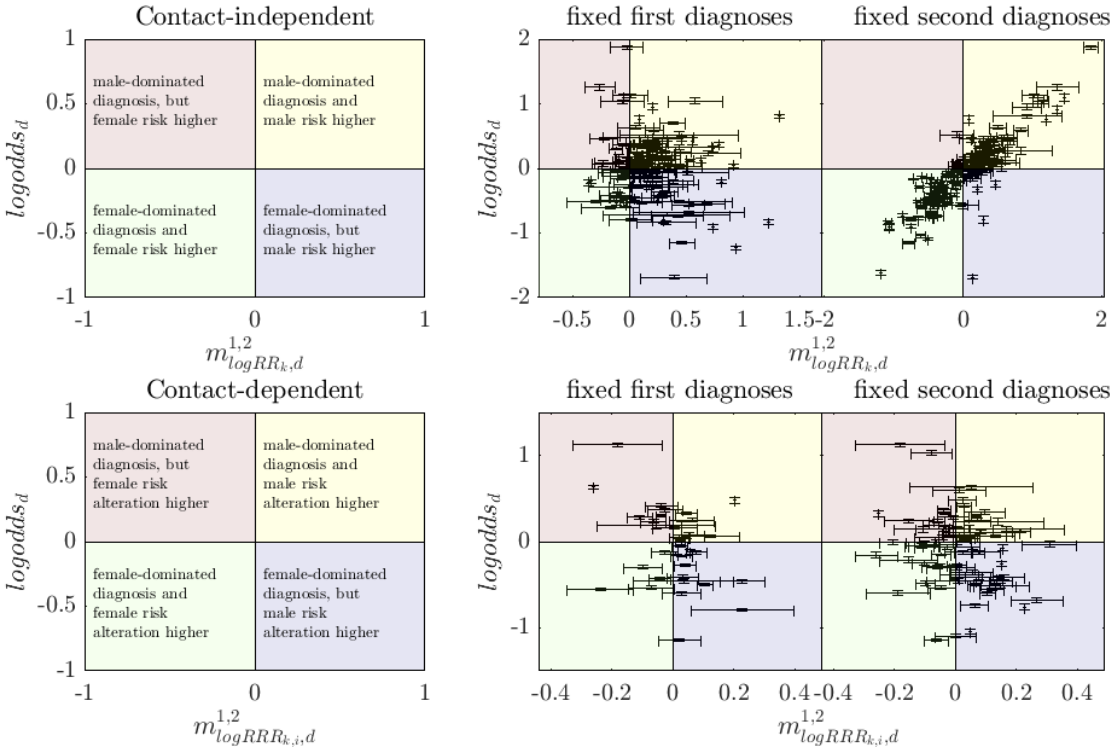


Figure 5.7: Top: Explanatory plot and contact-independent results of diagnosis prevalence vs. logarithmic relative risk. Bottom: Explanatory plot and contact-dependent results of diagnosis prevalence vs. logarithmic relative risk ratios. Each point stands for a fixed diagnosis.

Time window variation

Next we change the maximal time window between hospitalizations. The minimal time window is not changed as a minimum of 90 days is the only meaningful setting to exclude follow-up procedures in hospitals. The maximal time window is reduced in this analysis, leading to smaller patient samples and less diagnosis combinations. Settings and figures can be seen below (5.7, 5.8, 5.9, 5.10).

Parameter	Setting
minimal year of birth y_{min}	1906
maximal year of birth y_{max}	1956
minimal time window Δt_{min}	90 days
maximal time window Δt_{max}	525 days
minimal number of females and males per combination cut_{comb}	50
minimal number of patients per diagnosis cut_{diag}	1000

Table 5.7: Analysis parameters for robustness test with time window variation.

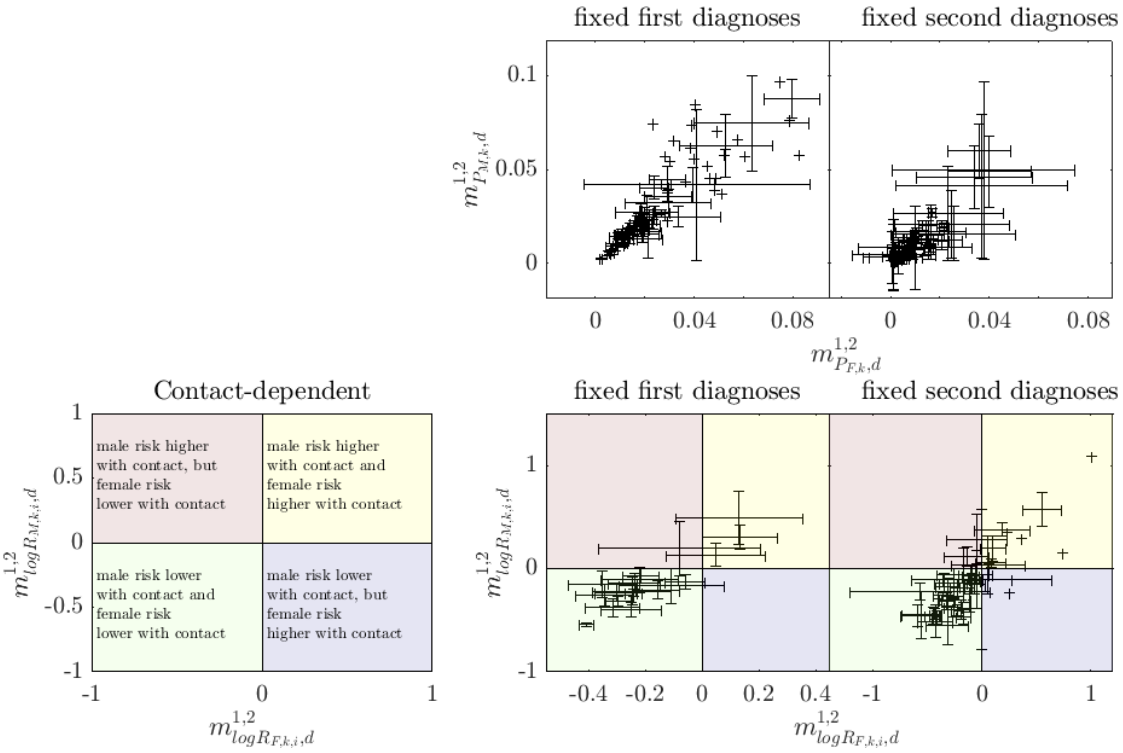


Figure 5.8: Male vs. female risk. Top: Contact-independent medians for re-hospitalization risk for males vs. females for fixed diagnosis. Bottom: Explanatory plot and contact-dependent medians of logarithmic ratios of males and females. Each point in the scatter plots stands for a fixed first diagnosis (left) or fixed second diagnosis (right).

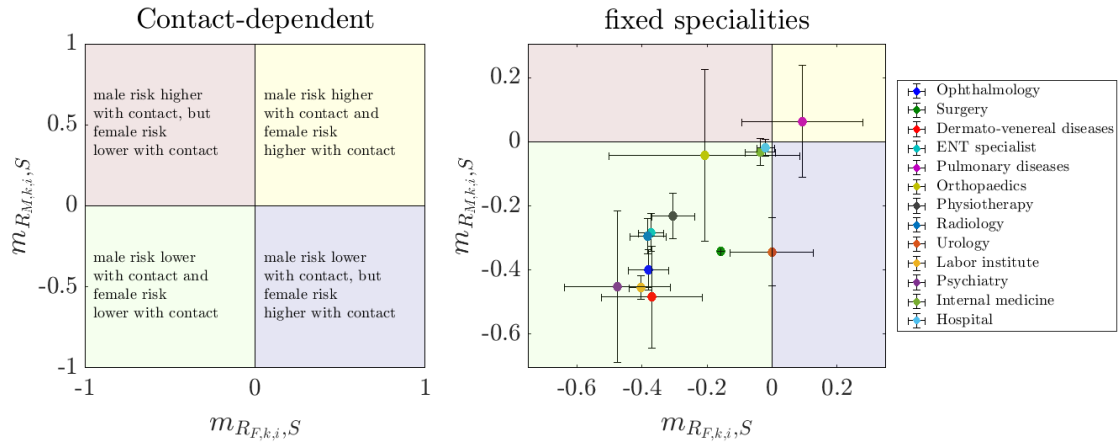


Figure 5.9: Fixed specialities. Explanatory plot (left) and contact-dependent logarithmic ratios for males vs. females with median for fixed specialities. Each color-coded point stands for a speciality from the list on the right.

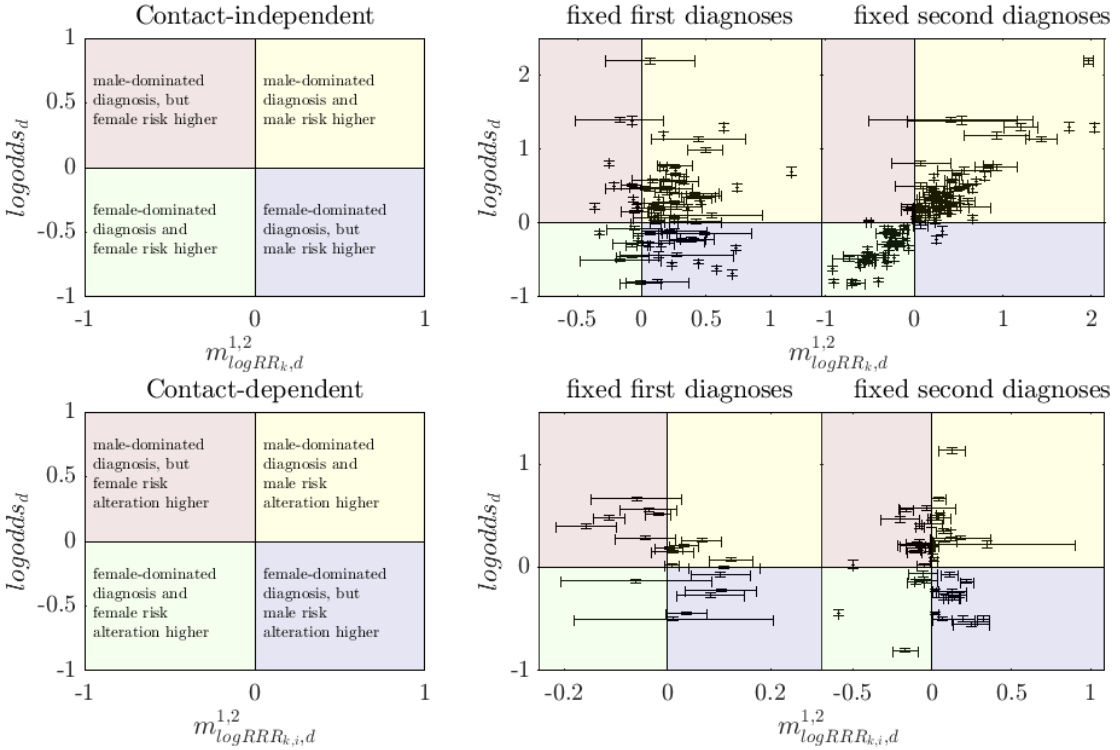


Figure 5.10: Top: Explanatory plot and contact-independent results of diagnosis prevalence vs. logarithmic relative risk. Bottom: Explanatory plot and contact-dependent results of diagnosis prevalence vs. logarithmic relative risk ratios. Each point stands for a fixed diagnosis.

Cut-off variation

The final robustness test varies the cut-off parameter for minimal number of females and males per diagnosis combination (or in case of contact dependence, per diagnosis combination and speciality). The minimum is lowered to 25, as an even larger minimum than 50 is unreasonable and eliminates too many combinations. By lowering the cut-off, the sample sizes and number of diagnosis combinations increase .The parameter table and resulting plots are shown in table 5.8 and figures 5.11, 5.12 and 5.13.

Parameter	Setting
minimal year of birth y_{min}	1906
maximal year of birth y_{max}	1956
minimal time window Δt_{min}	90 days
maximal time window Δt_{max}	1050 days
minimal number of females and males per combination cut_{comb}	25
minimal number of patients per diagnosis cut_{diag}	1000

Table 5.8: Analysis parameters for robustness test with cut-off variation.

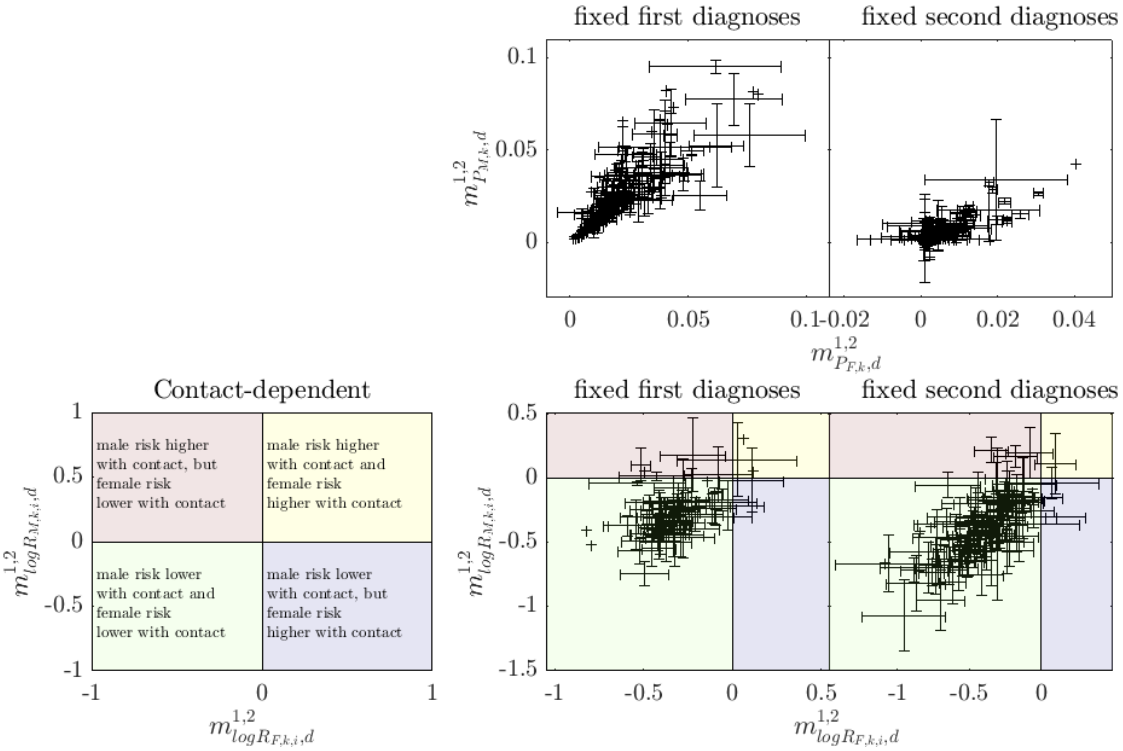


Figure 5.11: Male vs. female risk. Top: Contact-independent medians for re-hospitalization risk for males vs. females for fixed diagnoses. Bottom: Explanatory plot and contact-dependent medians of logarithmic ratios of males and females. Each point in the scatter plots stands for a fixed first diagnosis (left) or fixed second diagnosis (right).

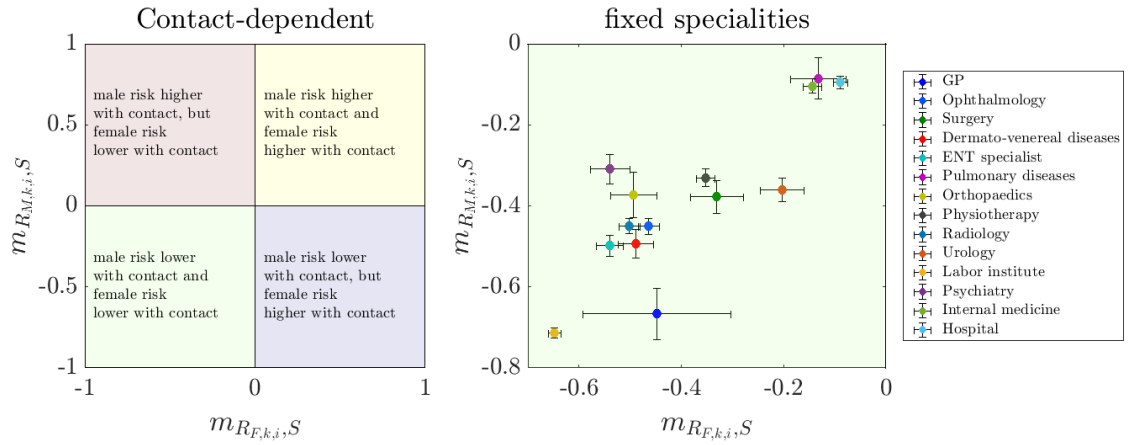


Figure 5.12: Fixed specialities. Explanatory plot (left) and contact-dependent logarithmic ratios for males vs. females with median for fixed specialities. Each color-coded point stands for a speciality from the list on the right.

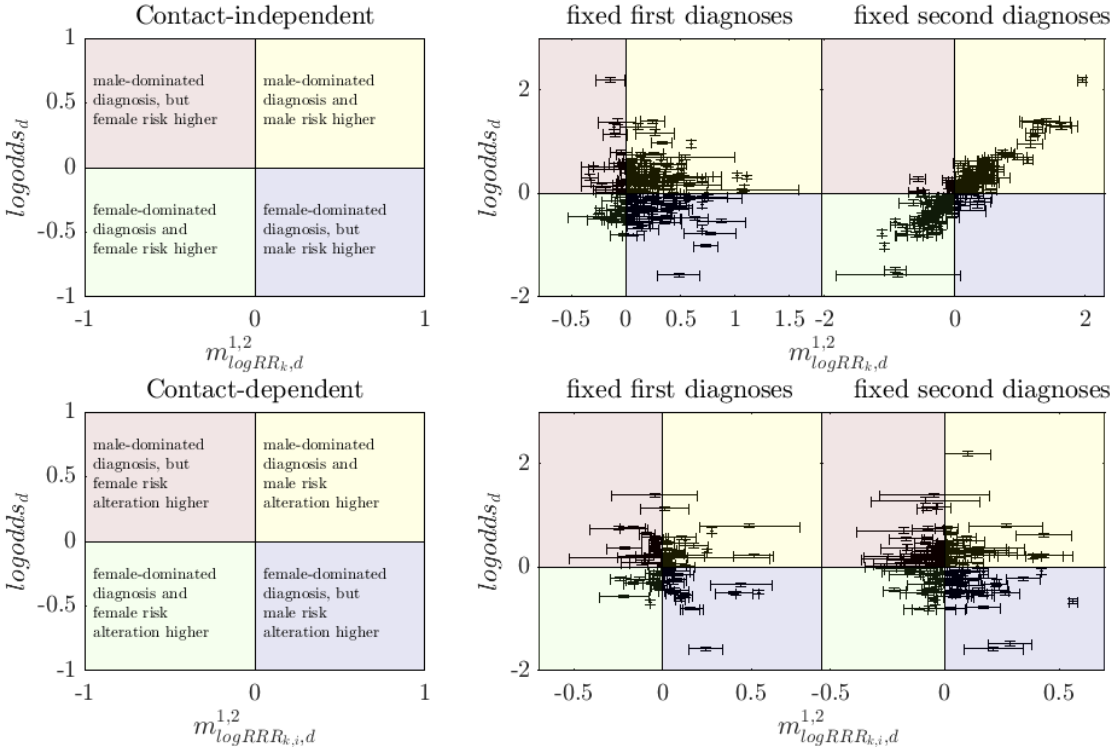


Figure 5.13: Top: Explanatory plot and contact-independent results of diagnosis prevalence vs. logarithmic relative risk. Bottom: Explanatory plot and contact-dependent results of diagnosis prevalence vs. logarithmic relative risk ratios. Each point stands for a fixed diagnosis.

Bibliography

[1] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Reviews of Modern Physics* 74.1 (2002), p. 47.

[2] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease". In: *Nature Reviews Genetics* 12.1 (2011), p. 56.

[3] Michael L Barnett et al. "Physician patient-sharing networks and the cost and intensity of care in US hospitals". In: *Medical Care* 50.2 (2012), p. 152.

[4] Norman Biggs. *Algebraic graph theory*. Vol. 67. Cambridge, England: Cambridge University Press, 1993.

[5] Norman Biggs, E Keith Lloyd, and Robin J Wilson. *Graph Theory, 1736-1936*. Oxford, England: Oxford University Press, 1986.

[6] J Martin Bland and Douglas G Altman. "The odds ratio". In: *BMJ* 320.7247 (2000), p. 1468.

[7] Stefano Boccaletti et al. "Complex networks: Structure and dynamics". In: *Physics Reports* 424.4-5 (2006), pp. 175–308.

[8] Robert Brown. "On the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies". In: *Edinburgh New Philosophical Journal* 5 (1828), pp. 358–371.

[9] Damon Centola. "The spread of behavior in an online social network experiment". In: *Science* 329.5996 (2010), pp. 1194–1197.

[10] Gary Chartrand and Ping Zhang. *A first course in graph theory*. Mineola, New York: Courier Corporation, 2013.

[11] Edward A Codling, Michael J Plank, and Simon Benhamou. "Random walk models in biology". In: *Journal of the Royal Society Interface* 5.25 (2008), pp. 813–834.

[12] Michele Coscia and Frank MH Neffke. "Network backboning with noisy data". In: *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE. 2017, pp. 425–436.

[13] Jan Salomon Cramer. *Logit models from economics and other fields*. Cambridge, England: Cambridge University Press, 2003.

[14] John Crank et al. *The mathematics of diffusion*. Oxford, England: Oxford University Press, 1979.

[15] Rick Durrett. *Probability: theory and examples*. Vol. 49. Cambridge, England: Cambridge University Press, 2019.

[16] Albert Einstein. "On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat". In: *Annalen der Physik* 17 (1905), pp. 549–560.

[17] Adolph Fick. "On liquid diffusion". In: *Journal of Membrane Science* 100.1 (1995), pp. 33–38.

[18] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences* 99.12 (2002), pp. 7821–7826.

[19] Thomas Graham. "XXVII. On the law of the diffusion of gases". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.9 (1833), pp. 175–190.

[20] Holger Hermanns. "Markov Chains". In: *Interactive Markov Chains*. Berlin: Springer, 2002, pp. 35–55.

[21] Joseph M Hilbe. *Logistic regression models*. London, England: Chapman and Hall/CRC, 2009.

[22] Lai Chung Kai. *Markov Chains: With Stationary Transition Probabilities*. Berlin: Springer-Verlag, 1967.

[23] Stefan Kitzler. *Comparison of methods to calculate diffusion in complex media and their application to the spreading of diseases*. Project work. Institute of Applied Physics at the Vienna University of Technology and Section for Science of Complex Systems at the Medical University of Vienna, 2016.

[24] Peter Klimek et al. "Quantification of diabetes comorbidity risks across life using nation-wide big claims data". In: *PLOS Computational Biology* 11.4 (2015), e1004125.

[25] Tibor Kudernac et al. "Electrically driven directional motion of a four-wheeled molecule on a metal surface". In: *Nature* 479.7372 (2011), p. 208.

[26] Bruce E Landon et al. "Using administrative data to identify naturally occurring networks of physicians". In: *Medical Care* 51.8 (2013), p. 715.

[27] László Lovász et al. "Random walks on graphs: A survey". In: *Combinatorics, Paul Erdos is Eighty* 2.1 (1993), pp. 1–46.

[28] Andrei Andreevich Markov. "The theory of algorithms". In: *Trudy Matematicheskogo Instituta Imeni VA Steklova* 42 (1954), pp. 3–375.

[29] Peter McCullagh and John A Nelder. *Generalized linear models*. London, England: Chapman and Hall, 1989.

[30] Helmut Mehrer. *Diffusion in solids: fundamentals, methods, materials, diffusion-controlled processes*. Vol. 155. Berlin: Springer Science & Business Media, 2007.

[31] Ralf Metzler and Joseph Klafter. "The random walk's guide to anomalous diffusion: a fractional dynamics approach". In: *Physics Reports* 339.1 (2000), pp. 1–77.

[32] Andrew Miles. "Complexity in medicine and healthcare: people and systems, theory and practice". In: *Journal of Evaluation in Clinical Practice* 15.3 (2009), pp. 409–410.

[33] Ron Milo et al. "Network motifs: simple building blocks of complex networks". In: *Science* 298.5594 (2002), pp. 824–827.

[34] Julie A Morris and Martin J Gardner. "Statistics in medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates". In: *British Medical Journal (Clinical Research Ed.)* 296.6632 (1988), p. 1313.

[35] Lori Mosca, Elizabeth Barrett-Connor, and Nanette Kass Wenger. "Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes". In: *Circulation* 124.19 (2011), pp. 2145–2154.

[36] Mark Newman. *Networks*. Oxford, England: Oxford University Press, 2018.

[37] Mark EJ Newman. "The structure and function of complex networks". In: *SIAM Review* 45.2 (2003), pp. 167–256.

[38] Pamela Ouyang et al. "Strategies and methods to study female-specific cardiovascular health and disease: a guide for clinical scientists". In: *Biology of Sex Differences* 7.1 (2016), p. 19.

[39] Romualdo Pastor-Satorras et al. "Epidemic processes in complex networks". In: *Reviews of Modern Physics* 87.3 (2015), p. 925.

[40] Neil Pearce. "What does the odds ratio estimate in a case-control study?" In: *International Journal of Epidemiology* 22.6 (1993), pp. 1189–1192.

[41] Hoangmai H Pham et al. "Primary care physicians' links to other physicians through Medicare patients: the scope of care coordination". In: *Annals of Internal Medicine* 150.4 (2009), pp. 236–242.

[42] Philippe Rigollet. *Statistics for Applications. Chapter 10: Generalized Linear Models (GLMs)*. Massachusetts Institute of Technology: MIT OpenCourseWare. 2016. URL: https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

[43] Everett M Rogers et al. "Complex adaptive systems and the diffusion of innovations". In: *The Innovation Journal: The Public Sector Innovation Journal* 10.3 (2005), pp. 1–26.

[44] Mahmoud Saleh, Yusef Esa, and Ahmed Mohamed. "Applications of Complex Network Analysis in Electric Power Systems". In: *Energies* 11.6 (2018), p. 1381.

[45] Simone Katja Sauter et al. "Analyzing healthcare provider centric networks through secondary use of health claims data". In: *2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2014, pp. 522–525.

[46] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. "Extracting the multiscale backbone of complex weighted networks". In: *Proceedings of the National Academy of Sciences* 106.16 (2009), pp. 6483–6488.

[47] Shai S Shen-Orr et al. "Network motifs in the transcriptional regulation network of Escherichia coli". In: *Nature Genetics* 31.1 (2002), p. 64.

[48] Ingve Simonsen et al. “Diffusion on complex networks: a way to probe their large-scale topological structures”. In: *Physica A: Statistical Mechanics and its Applications* 336.1-2 (2004), pp. 163–173.

[49] Christopher L Sistrom and Cynthia W Garvan. “Proportions, odds, and risk”. In: *Radiology* 230.1 (2004), pp. 12–19.

[50] Joachim P Sturmberg and Carmel M Martin. “Complexity and health–yesterday’s traditions, tomorrow’s future”. In: *Journal of Evaluation in Clinical Practice* 15.3 (2009), pp. 543–548.

[51] Magdalena Szumilas. “Explaining odds ratios”. In: *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 19.3 (2010), p. 227.

[52] Stefan Thurner, Rudolf Hanel, and Peter Klimek. *Introduction to the theory of complex systems*. Oxford, England: Oxford University Press, 2018.

[53] Doug Toussaint and Frank Wilczek. “Particle–antiparticle annihilation in diffusive motion”. In: *The Journal of Chemical Physics* 78.5 (1983), pp. 2642–2647.

[54] Robin J Wilson. *Introduction to graph theory*. London, England: Pearson Education India, 1979.

[55] Damian H Zanette and Pablo A Alemany. “Thermodynamics of anomalous diffusion”. In: *Physical Review Letters* 75.3 (1995), p. 366.