Exploring the Danish Diseasome

Jensen, Anders Boeck; Brunak, Søren; Jensen, Lars Juhl

Publication date:
2013

Document Version
Peer reviewed version

Exploring the Danish Diseasome

PhD thesis
Anders Boeck Jensen

Supervisors:
Søren Brunak
Lars Juhl Jensen

October 20th 2013

Center for Biological Sequence Analysis
Department of Systems Biology
Technical University of Denmark
Kemitorvet, Building 208
DK-2800 Lyngby
Denmark

Preface

This work was submitted as the requirements of obtaining a PhD degree at Center for Biological Sequence Analysis, Department of Systems Biology, DTU. It has been funded by a grant given by Det Strategiske Forskningsråd (the Danish Stragical Research Council) to the ESICT-consortium (Experience-oriented Sharing of health knowledge via Information and Communication Technology).

The majority of the work has been carried out at the Novo Nordic Foundation Centre for Protein Research at KU.

The thesis work has been supervised by Søren Brunak and Lars Juhl Jensen.

Ethical statement
This study is based on registry data from the National Patient Registry. Use of data has been approved from the Danish Data Registration agency (j.nr. 2010-54-1059).

Abstract

In the recent decades, there has been a shift in biological research towards data-driven analyses, where entire biological systems are investigated. System level analysis is applied in medical research with approaches such as personalized medicine and predictive medicine, where two of the goals are to predict diseases predispositions and the outcomes of medical treatments for each patient. In traditional medical and epidemiological research, medical conditions are investigated one at a time. This is done to eliminate predisposing effects of confounding factors such as other diseases and environmental components. In contrast, systems level research is often performed in a data-driven manner, where the aim is to analyze the impact of the full diseaseome instead of analyzing a disease as an isolated entity.

This thesis presents four studies of health registry data, all with the aim to characterize how diseases correlate and develop throughout the entire population of Denmark. Three of the analyses have been carried out at a systems level, where groups of diagnoses have been examined to better understand disease relationships. It has been shown how disease progression over time can be analyzed with data-driven methods. This was done by identifying pairs of diagnoses that show strong temporal correlation and analyzing how patients progress in different trajectories of these diagnoses. Rather than focusing on a trajectory of a single disease, patterns of disease development across the full spectrum of pathology were identified.

In another study included in this thesis, the hypothesis that gut bacteria plays a role in cardiovascular diseases is analyzed by comparing patients who have undergone full colectomy with patients who have their colon intact. Finally, a study shows how health registry data can be used to examine genetic diseases. The correlations between Mendelian and complex diseases have been analyzed to identify diseases that might share genetic etiology.

The disease trajectories presented here is a step towards predicting outcome of medical treatment, by unraveling temporal disease correlations. While the correlations and trajectories are descriptive, the results represent ideal input to predictive models. Additional data concerning medical treatment and surgery, drug prescriptions and genotype can also be incorporated into such models. Thus, they can aid in the further development of personalized and predictive medicine.

Dansk resumé

Der har i de seneste årtier været et skift indenfor forskning i biologi, hvor data-drevne metoder, der analyserer fulde biologiske systemer i stedet for isolerede delsystemer, er blevet taget i brug. Analyser foretaget på system niveau bliver brugt indenfor *personalized and predictive medicine*. Her er to af målene at kunne lave statistiske forudsigelser af virkningen af medicinsk behandling for den enkelte patient og forudsige hvilke fremtidige sygdomme, den enkelte patient er disponeret for. I traditionelle medicinske og epidemiologiske studier analyseres sygdomme én ad gangen. Dette gøres, så den prædisponerende effekt fra andre sygdomme og miljøets påvirkning af patienterne kan elimineres. I modsætning til dette analyserer undersøgelser på system niveau ofte det fulde sygdomsspektrum, i stedet for at analysere en sygdom som en isoleret enhed.

Denne afhandling præsenterer fire register baserede studier, hvor målet er at karakterisere korrelationer mellem sygdomme og sygdomsudvikling for hele Danmarks befolkning. Tre af analyserne er blevet foretaget på system niveau, hvor grupper af sygdomme er blevet analyseret for at få bedre forståelse for sygdomssammenhængen. Det er blevet vist hvordan tidslig sygdomsudvikling kan analyseres med data drevne metoder. Dette er blevet gjort ved at identificere sygdomspar, som udviser stærk tidslig korrelation og analysere hvordan patienter følger forskellige baner af disse sygdomme. Der blevet fundet mønstre på tværs af det fulde spektrum af sygdomme, i stedet for at fokusere på en enkel sygdom.

Ydermere indeholder afhandlingen et studie, der analyserer en hypotese om tarm bakteriernes rolle i hjerte-karsygdomme. Studiet sammenligner patienter, der har fået foretaget fuld kolektomi med patienter, der stadigvæk har en intakt tyktarm. Slutligt præsenteres et studie, der viser hvordan registerdata kan bruges til at undersøge genetisk betingede sygdomme. Her vises hvordan korrelationer mellem Mendelske og komplekse sygdomme kan bruges til at identificere delt genetisk ophav mellem sygdomme.

Analysen af sygdomsudvikling, der præsenteres her, er et skridt på vejen mod at forudsige effekten af medicinsk behandling. I deres nuværende form er analyserne af sygdomskorrelationer og sygdomsudvikling deskriptive, men kan bruges som input til forudsigende modeller. På denne måde kan de bruges i den fremtidige udvikling af *personalied and predictive medicine*.

Acknowledgement

As a PhD student you are luckily not a one-man army. This thesis would never have been written without the help from many people, to whom I owe big graditude.

First I would like to thank my officemate Peter who urged me to apply for the PhD position and without whom I would not have been PhD here. We have had many good discussions and helped each other on many occasions.

Secondly I owe the greatest thanks to Pope, Tudor, Thorkild and Teresa for interpreting my results and giving qualified input for further analysis. Without the excellent collaboration with them, I would have spent endless hours trying to figure out what is interesting in the results that I made. Perhaps the biggest lesson from the PhD is not to venture in the field of medical research without the backup from a medical doctor. I also want to give a thanks to Sabrina, Robert and Henriette who helped me with the medical knowledge, especially during the first years when I was without aid from the first mentioned people.

Thanks to my family, Gert, Bente and Torben who have supported and inspired me throughout all my studies. And to my very special friend, Salman *Mera Pyar*, whose love and support have made the last year of my studies very enjoyfull. You have lifted many burdens for me during the last time of my PhD.

I would also like to thank my entire group in CPR. It would not have been three good years without a good group of nice people. Especially also thanks to my officemates Albert and Andreas and Peter. I have greatly appreciated your company during these years. And a very special thanks to Damian who introduced me to Salman.

I would also like to thank Kalliopi, Sabrina and Salman, who have done proof reading of this thesis. The reader will probably also appreciate that many grammatical and splicing errors have been dealt with by them. And thanks to Albert, Alberto and Robert for beta testing parts of the thesis.

Also thanks to the ESICT group and the HEXANORD group.

Second to last, but not least, thanks to my two supervisors Lars and Søren. Thanks to Søren for providing me with ideas and input and to Lars for being able to capture some of the ideas and helping to turn them into practical solutions. It is needless to say that without the help from these two people, there would have been no thesis.

My last thanks goes to all other who contributed with large or small things, but did not make it to this page.

Publications

Manuscript I
Anders Boeck Jensen, Pope L. Moseley, Tudor I. Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, Søren Brunak. **Patient-specific disease trajectories condensed from population-wide registry data** (manuscript ready for submission)

Manuscript II
David R. Blair, Christopher S. Lyttle, Jonathan M. Mortensen, Charles F. Bearden, Anders Boeck Jensen, Hossein Khiabanian, Rachel Melamed, Raul Rabadan, Elmer V. Bernstam, Søren Brunak, Lars Juhl Jensen, Dan Nicolae, Nigam H. Shah, Robert L. Grossman, Nancy J. Cox, Kevin P. White, Andrey Rzhetsky, **A Non-Degenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk,** Cell 155:1, p. 70-80, 2013

Work in progress

Work in progress I
Analysis of non-temporal correlations

Work in progress II
Risk of cardio-vascular disease following total colectomy
Made in collaboration with Teresa A. Ajslev, Thorkild I. A. Sørensen.

Thesis overview

This thesis presents retrospective data-driven analysis of disease correlations using the Danish National Patient Registry. **Chapter 1** introduces the research field. The data used and other health related data sources are discussed in **chapter 2**. The disease correlations have been analyzed in a non-temporal manner described in **work in progress I** and discussed in **chapter 3**; and in a temporal manner that is extended to cover longer disease trajectories described in **manuscript I** and discussed in **chapter 4**. **Work in progress II** presents a thesis regarding the role of gut-bacteria in cardiovascular diseases, which is covered in **chapter 5**. **Chapter 6** covers **manuscript II** that addresses how genetics knowledge can be inferred from disease correlations. **Chapter 7** contains concluding remarks on the analyses.

Table of contents

Chapter 1:
General introduction

1.1 Background

Humans have attempted to fight diseases for thousands of years - today using modern medicine based on scientific studies. Epidemiological studies are among the methods for inferring knowledge for new medical treatments. In the recent decades, genomics, proteomics and other fields of cellular biology have also become important contributors to knowledge that can lead to new medical treatments [1,2]. Systems Biology brought an important shift in paradigm within the biological sciences: Instead of examining a reduced and simplified part of a biological system, the whole system is now the target for examination [3]. This shift has now reached epidemiology with data-driven studies that analyze correlations for the entire system of diseases [4,5]. Such studies are an important part of the progress to characterize the complex interaction between diseases. It has already been shown that many genetic diseases are correlated through shared genes [6,7]. The effect of genes in diseases is not only limited to genetic (hereditary) diseases [7]. For example, genes that can limit the effect of AIDS have been identified [8]. It is also well known from epidemiological studies that diseases are also correlated through common predisposing factors such as the environment [9]. The effect of environmental factors can in some cases be explained by molecular mechanisms [7]. For example, a study shows it is possible that chronic stress affect cancer through a protein pathway [10].

Since the sequencing and mapping of the human genome, there has been a large effort to create personalized medicine tailored to each patient's genome [9]. P4 medicine is the development of Personalized medicine that furthermore focuses on Prediction, Prevention, and Participatory [2,11]. One aim of P4 medicine is predicting the effects of treatments on individual patients [12]. Another aim is to predict future diseases from the current medical history of the patient in order to prevent them. The majority of genetic diseases are not caused by a single gene, but by a complex interaction between multiple genes, where each gene can contribute to many diseases [7,13]. Therefore, it important to study correlations between diseases to reach the goals of P4 medicine [14]. Correlation analysis can both uncover diseases that possibly share genetic etiology and diseases that correlate through a disease mechanism or common risk factor. This can assist in prognostics and clinical decision-making. Here temporality is an essential factor. Resolving the temporal directionality of disease correlations and analyzing longer disease progression patterns can serve as a basis for predicting future outcome.

While there has been much focus on analyzing protein interaction networks and other molecular data, disease registries remain an underutilized source for uncovering correlations between diseases and investigating longer disease progression on a large scale [4]. The full population data available in these registries is useful to achieve the goals of P4 medicine, and methods from population statistics should be incorporated into all aspects of P4 medicine [15]. In recent years a few studies have started focusing on these registries as a source of large scale disease correlation analysis [4,16]. The work presented in this thesis is a continuation and extension of these initial steps.

1.2 Epidemiological research

Patterns of disease development in the population and the effects of demographics and risk factors on diseases are analyzed in epidemiology. This thesis presents epidemiological work on disease correlations, where one goal is to identify correlations and disease trajectories that can be explained by genetics. However, it is complicated to infer genetic knowledge from disease registry data as discussed in chapter 6. Therefore, this thesis focuses more on epidemiology than genetics.

Epidemiological analyses that uses registry data to analyze correlations between pairs of diseases and between diseases and risk factors are very common [17]. For example, the Danish Cancer Registry has been integrated with the Registry of Induced Abortions to investigate risk of cancer after abortion [18] and with the Registry of Blood Donors to investigate risk of cancer after repetitive blood donation [19]. In both the studies, number of patients who were diagnosed with cancer after being subjected to the risk factor was counted and different groups of patients compared. No correlation between the two risk factor and cancer was found in these studies.

These are examples of traditional epidemiological research, where a hypothesis is required before the study can be carried out. The abortion-cancer study [18] investigated a prior stated hypothesis [20]. A way of identifying new hypotheses is to focus on common severe diseases such as cancer and systematically study all the risk factors one by one. A risk in using this approach is the lack of correction for multiple testing (unless they are all studied simultaneously). Several different labs can study the same non-significant risk factor until it is found to be significant by chance.

It is an important aspect of epidemiology to control for confounding factors. In the study of blood donation and cancer, factors such as gender, age and number of blood donations were controlled for in the statistical analysis [19]. Ideally, studies should be designed so patients are compared to cases with same pre-disposing risks. For example, in cancer studies, smoking status should be part of the criteria for matching cases and controls. Age and gender are two other very important confounding factors. Elderly people are more prone to a variety of diseases as they are weakened by age. This can lead to an increased number of age-related diseases that co-occur in the same patient because the patient is old. Likewise, gender is important as diseases such as prostate cancer only occur in one gender. Other confounding factors include weight, lifestyle related factors (e.g. physical activity and alcohol consumption) and social status.

1.3 Data driven research

In the data driven approaches, the process is the inverse of hypothesis driven approaches: The data is collected and tested to form hypothesis a posteriori in an explorative manner. This approach is often used in studies that seek to illuminate entire systems rather than analyze specific subsets (such as a single disease). For data driven research, the objective is often to explore all possible correlations in a large data set (e.g. [4,5]. An important task here is to identify interesting hypotheses that can be translated into usage. For studies generating thousands or millions of hypotheses, this can be a big challenge requiring extensive amounts of manual work. However, because the full system (e.g. all diseases) is investigated when using data driven approaches, it is possible to identify surprising and novel results that would otherwise have gone unnoticed in

hypothesis-driven research. In addition, because data driven research examines the whole system of disease correlations, it is possible to investigate groups of correlations instead of single correlations.

The work on non-temporal co-morbidities and disease trajectories presented in this thesis contains examples of both benefits and challenges of data driven research. Diseases tended to correlate strongly in ways that can be expected from disease classifications, but less expected correlations were also identified. Identifying the new and unexpected correlations that were also interesting for further investigations was a challenge. It was also shown that analyzing correlations in a more stringent manner reduced the number of correlations, so the final set of correlation contains fewer trivial diagnoses.

1.4 Hypotheses in data-driven research

Even though data-driven analysis is often used as an antonym of hypothesis-driven analysis, there are always some fundamental hypotheses or assumptions in data-driven research that are not always tested explicitly. The main hypotheses for this thesis is that some biological processes cause diseases to occur in the same patients in a non-random fashion, and that this is reflected in medical registry data. There can be several explanations to why diseases co-occur non-randomly in the same patients. For example a common risk factor can predispose for multiple diseases, a disease mechanism can lead from one to another (such as cancer metastasizing or hepatitis C causing liver cancer which further can cause liver failure and related diseases) or a genetic mutation can drive both diseases. Irrespective of the cause, it is of interest to analyze disease correlations. From an epidemiological point of view, it is important in its own right to gain knowledge about the population prevalence of diseases and correlations with other diseases and risk factors. This can assist for prevention of diseases, by identifying risk groups who need to be monitored for a disease. From a molecular biological view, analysis of disease correlations is a source of understanding the underlying genetics.

The hypothesis that the registry data can uncover biological processes is not tested directly in this thesis. However, we found many known and logical correlations among the results, which can serve as a sanity test for the hypothesis. It has also been observed in many epidemiological studies. However, we also found examples where the diseases appeared in the reverse order of the known causal relation. An example is the significant correlation from Manuscript I of hyperplasia of prostate leading to prostate cancer. Here the prostate cancer is the likely cause of the hyperplasia, and the direction observed in the registry is a consequence of how the diseases were diagnosed. Another such example from a different study of pneumonia patients where the patients who died quickly after arriving to the ER were found to be the healthiest [21]. This bias was caused by the routine that the patients who died quickly were not diagnosed as thorough as those surviving. There are many such examples and it is important to consider this when interpreting the results of registry analyses.

1.5 Disease progression and disease trajectories

The study of disease progression over time is an important step towards being able to predict the disease outcome for a single patient. When doing data-driven analysis of disease progressions on a large scale, the number of possibly trajectories that can be studied is a main issue. The number of combinations of diseases grows exponentially for each step added. It is therefore necessary to limit the number of possibly trajectories up front. Descriptive statistics (as the one presented in manuscript I) can be used to identify diseases that can predict severity of the future outcome. After identifying the diseases of interest, there are several ways of analyzing the disease progression. The progressions can be modeled as a state machine using methods such as Hidden Markov Models. This will help to characterize the progressions, but can also be used to predict further outcome. Alternatively, methods from machine learning can be used for prediction. This can involve methods such as neural networks that given the current history of the patient can predict future outcome. Many other options are available, and the most accurate results are likely to come from methods that combine features from different fields such as epidemiology, machine learning, network analysis etc.

Chapter 2:
Data

The data used in all studies of this thesis is from Danish National Patient Registry (NPR), which is a population wide disease registry. This chapter describes this registry along with other population registries.

Recording of different population statistics has been performed for a long time. In Denmark the Bureau for Population Statistics (Statistics Denmark) was founded in 1850 with the task of counting the population for taxation purposes. Nation-wide disease registries started already in 1942 with the Danish Cancer Registry. Since the introduction of the Central Person Registry (CPR) in 1968, all Danish citizens have been assigned a unique ID commonly known as the CPR-number. Individuals who have moved to Denmark to work or to stay permanently also get a CPR-number. The CPR system has been a contributing factor for the creation of a growing number of high quality registries, where the patients can be tracked across registries. While most of the registries were started for administration purposes, they have also been used for surveillance and research.

Hospitals started reporting diagnoses and procedures to the NPR in 1977. It is one of the highest quality disease registries available for research in the world as it covers all encounters over a long period of time and allows researchers to obtain fully person-identifiable data [22]. The data offers a great potential for doing large-scale studies. The other Nordic countries have disease registries as well. However, none of them cover as long a period of time as the Danish. Some cover fewer areas and others only recently have person-identifiable data available [22]. Other disease registries such as the US Medicare claim registry offer a larger population. However, it is a challenge to track patients without a centralized person identifier. In addition, the Medicare data has more biases since it only covers patients aged 65 and above who cannot afford private healthcare and some young disabled patients.

2.1 Central Patient Registry

CPR contains information on date of birth, date of death, names, addresses and civil status and other administrative data [23]. Many of the variables such as addresses have historical data as well. The parents of each person are listed making it possible to track kinship. Siblings can be identified through common parents, cousins through common grand parents and so on. This is limited to the generations alive in 1968. For older generations, it is possible to track them through church books, which are available online [24]. The date of death available through CPR are generally more accurate than the date supplied in death certificates, and are therefore used in the Causes of Death Registry.

2.2 National Patient Registry

The National Patient Registry (NPR) contains data on diagnoses, surgical procedures, treatments and administrative information like waiting time for all hospital contacts in Denmark. Diagnoses are coded in ICD-10 (described in section 2.4.1). Initially the registry included only somatic inpatient encounters, but since 1987, outpatient and Emergency Room (ER) encounters have also been registered. There is also a registry for psychiatric hospitals, which has the same structure as NDR. Additional permissions are required to use this data.

2.2.1 Reporting procedure

The reporting to NPR has to follow guidelines for minimum registration, *Fællesindhold for basis registrering af sygehuspatienter*, published by Statens Seruminstitut. During a hospital encounter notes about the patient's state is made in the patient files. This includes physical examinations, confirmation and rejection of diagnoses, surgical procedures and treatment. When the patient is discharged, a doctor reads the patient files and reports diagnoses, procedures and treatment to the registry along with administrative information [22,25].

2.2.2 Data quality

The data from NPR has been used for many studies [17], however there have not been many reviews of the overall data quality. It has mainly been verified for special uses where discharge summaries and patient files have been obtained for a selection of diagnoses. Positive Prediction Value (PPV) and Negative Prediction Value (NPV) are common measurements of accuracy. The PPV is the percentage of rightfully assigned diagnoses (true positives) out of total number of reported diagnoses (sum of true and false positives), and NPV is the amount of true negatives for patients without the diagnosis. A study from 2011 shows that the data has high quality for calculating the Charlson comorbidity index [25]. Discharge Results showed an overall PPV of 98% and PPV between 82% and 100% for each of the 19 diagnoses covered by the Charlson index. A quality study of COPD diagnoses from 2011 reports a high PPV of 92%, but a lower NPV of 82% [26]. Other recent studies report a PPV of 75% for venous thromboembolism [27], 65.5% for acute coronary syndrome [28] and 59% for rheumatoid arthritis [29]. These studies demonstrate that quality of coding depends on the specialty and that conclusions should be drawn with care when relying solely on registry data. An older study from 1995 concludes that the less specific (3 digit ICD-8) codes have higher accuracy than the more specific (5 digit ICD-8) codes [30] which is still likely to be the case for the newer ICD-10. Using the less specific codes reduces errors where one specific code was wrongly chosen instead of another specific code.

In [29] it is found that patients with only one rheumatoid arthritis encounter had a PPV of 42%, whereas patients with three or more have a PPV of 91%. Taking the number of encounters into consideration when analyzing the data could make the results more reliable. However, it is difficult to quantify how many encounters are needed, and it will be heavily dependent on the specialty. Studying the correlation between the quality and the number of repetitions for a wide variety of specialties would be very interesting and of use to many registry studies. An unpublished in-house study of heart defects shows a very small PPV for patients given an unspecified heart failure code, but a higher PPV for patients with specific diagnoses. This might be a general trend, and the PPV of

unspecific codes could be a focus when reviewing the accuracy of the registry. Temporality can also be an important factor in studying accuracy: Some diseases require an extensive follow-up period. If no such period is observed after a discharge with the disease, the patient might to not have the disease.

A key issue for data quality is the fact that the data is used for billing purposes. The hospitals are reimbursed by the state according to Diagnosis Related Groups system for inpatients and Danish Ambulatory Grouping System for outpatients. These system takes diagnoses, age, the length of stay, and other factors into consideration when placing an encounter into a group with an associated reimbursement rate [31]. An important factor is which diagnosis is categorized as the main diagnosis and which is supplementary diagnosis. Changing which diagnosis is the main and which is the supplementary can change the reimbursement rate. Therefore, economic considerations also contribute in deciding the main diagnosis in addition to medical judgments. The reimbursement system also gives bias towards including additional diagnoses present in the patient, which are not relevant for the current encounter, and more severe diagnoses. As we have been ignoring the main and supplementary diagnosis categories, and have used codes at a less specific level where severity is not taken into account, the issues of main versus supplementary diagnosis and severity are not relevant to our studies. However, inclusion of diagnoses unrelated to the current encounter can influence our analysis.

Given the issues with data quality, interpretation of results from registry studies should be done carefully. Special care must be taken for rare diseases where a large error rate has a big effect. For this reason the studies presented in this thesis are focusing on the strong signals from large patient groups.

2.2.3 NPR dataset

The data set we have analyzed data contains all records from 1996 to 2010. It covers 6.2 million patients with 68 million hospital encounters and 54.9 million unique patient-diagnosis associations. It is further described in Manuscript I.

The NPR data set was cleaned by removing records with invalid personal identification numbers, records with only legacy ICD-8 diagnoses and records lacking primary diagnoses. Each record covers the time between admission to one ward until discharge from the hospital or to another ward. Records for inpatients were concatenated to cover an entire admission when there were discharges between wards. Readmissions of inpatients 1 day or less from a discharge were also concatenated to the first admission. 1.5 million inpatient records were concatenated with other records giving 66.5 million concatenated admissions, outpatient visits and ER visits in total.

2.3 Other registries

2.3.1 Statistics Denmark

Statistics Denmark releases yearly statistics reports on population census, income, employment, etc. These reports have been freely available online since 2001 through *statistikbanken.dk*. They also have data available for researchers in a person identifiable form. This includes data that can be used to describe the socioeconomic status of each

person, which is of interest as confounding factor in medical research. However, the data is only available through a closed data system. Researchers can only analyze the data with the programs made available by Statistics Denmark, making it tedious to use it. In addition, only small data sets can be exported from there to prevent users from exporting person-identifiable data. For medium and large-scale data analysis, this limits the usability.

2.3.2 Diabetes Registry

Diabetes is becoming an increasingly common disease [32]. Therefore, it is the target of many medical studies and of interest to epidemiologic studies. Many diabetes patients are treated by the general practitioner or at specialized clinics and only severe complications, such as diabetic foot ulcer, are treated at hospitals. Consequently, the prevalence of diabetes found in NPR is too low. In order to obtain an estimate of the diabetes prevalence in Denmark, the Danish Diabetes Registry, which integrates different data sources, was started. Unfortunately, the registry is based on uncertain indicators (for example repetitive blood samples), and there is no guarantee that people registered in it have diabetes [33,34]. As a result, it is not well suited for epidemiological studies.

2.3.3 Drug prescription registry

Statens Serum Institue maintains a registry of medicine prescriptions in Denmark. However, its aim is to survey the total medicine consumption and not all prescriptions are available on a person-identifiable level. Hospitals register medicine usage at department level. Since the data is on prescriptions and not sale or usage, it is not certain that patients used the medicine. Furthermore, the data is available through Statistics Denmark and is therefore prone to same difficulties as described in section 2.3.1.

2.4 Disease classifications and resources

2.4.1 ICD

The International Classification of Disease, ICD, is a classification that covers the full spectrum of pathology. It is updated regularly; version 8 (ICD-8) was used in NPR from 1977 to 1993, and version 10 (ICD-10) since 1994. The ICD-10 version used in NPR is a Danish adaptation with more codes. The US Medicare data still uses version 9, but will soon update to version 10. While the changes from version 8 to 9 are not large, there are many fundamental changes introduced in version 10. Therefore, there is no one to one mapping between them. This makes it almost impossible to compare correlations between the two versions if the mapping between versions is not carried out before calculating the correlations.

ICD-10 is structured in 21 chapters that cover different areas of pathology. For example, there are chapters for different classes of diseases (e.g. chapter I 'Infectious diseases' and chapter II 'Neoplasms') and different anatomical structures. Each chapter is subdivided into more specific block such as 'Intestinal infectious diseases' in chapter I. Each of these blocks covers a range of codes. The codes are alphanumerical and start with one letter followed by two or three numbers. Codes with three characters are called level-3 codes,

and with four characters are called level-4 codes. The Danish ICD-10 adaption allows for more alphanumerical levels.

The classification is hierarchical, so A000 'Cholera due to Vibrio cholerae 01, biovar cholerae' is a sub-diagnosis of A00 'Cholera'. This fact can be used when counting the diagnoses. Patients having A000 can also be counted as having the more general A00 (the diagnosis can be "rounded"). The same applies for further sub-levels, e.g. G229A (only in the Danish ICD-10) where the patients having G229A, also can be counted for G229 and G22.

2.4.2 SNOMED-CT

Systematized Nomenclature of Medicine and Clinical Terms, SNOMED-CT, is classification for diseases and healthcare with a strong focus on use in EHR systems. It is a combination of the two independent classifications SNOMED and CT. The terms are connected through relationships. The "is a" relationship is used to define the hierarchical structure where each term can have multiple parent terms. There are 17 base hierarchies (e.g. Clinical findings) and all terms can be traced back to one or more of these. Because of the possibility for multiple parents, the hierarchy cannot be used to "round" diagnoses to a more general diagnosis. Unlike ICD-10 the codes of SNOMED-CT are numeric identifiers that are used as a key in a database structure, and they make no sense by themselves.

Advantages of using SNOMED-CT include that the similarity of terms can be calculated for all terms against all. In this way, the classification contains knowledge that for example could be used for analyzing disease correlations. In ICD-10, codes can only be in the same block or chapter or not related at all. Another advantage is the possibility of making for compound terms where two or more terms can be pre- or post-coordinated. This enables users to extend the classification for their needs and still maintain compatibility with other users of SNOMED-CT.

2.4.3 UMLS

Unified Medical Language System, UMLS, is a meta-classification that combines more than hundred classifications, among others ICD and SNOMED-CT. It can be used to perform cross mapping between different classifications. Like SNOMED-CT, it consists of terms connected by relationships. Because it consists of multiple classifications, it has issues such as redundant terms, and cyclic "is-a" relationships, where a term can be both an ancestor and predecessor of another term [35]. Some of these errors can be avoided by limiting which classifications are used.

2.4.4 Mapping issues

There is no objective truth of how diseases should be hierarchically structured and how many details to include in a classification. Therefore, there are large issues when mapping between different classifications and even between different versions of the same classification. Despite efforts to create cross mapping between classifications, no simple

solution to this problem exists. Sometimes the mapping is complex so that a combination of two diseases is mapped to one disease. Therefore, the full medical record of a patient can be necessary to map successfully. There is no publicly available full mapping between ICD-8 and ICD-10. ICD-9 has been mapped to ICD-10 as the registries in United States are switching to ICD-10.

Chapter 3:
Disease correlations

Correlation is a statistical term for measures of how much two variables are related to or depend on each other. Correlations between diseases commonly measure how much presence of one disease increases or decreases the risk of presence of another disease. Disease correlations are sometimes called (clinical) associations in epidemiology. This chapter mainly addresses non-temporal correlations while temporal correlations are discussed in chapter 4 and 5.

Correlations between pairs of diseases have been quantified by counting the number of patients who have both the diseases in this thesis. A pair of diseases is said to be co-morbid if they co-occur in the same patient. The strength of correlation for pairs of diseases is quantified using Relative Risk (RR). Together with statistical tests, this measure can be used to identify co-morbidities that are unlikely to occur together by chance and can be candidates for further investigations.

Disease correlations can both be positive and negative. In the work in progress I on non-temporal correlations 8% of the correlations were negative and in [4] a third were negative. A possible explanation for the high number of positive correlations is positive feedback: most diseases cause the patients to become weaker, and weakened patients get more diseases. It is interesting to analyze both positive and negative correlations. If the mechanism causing negative correlation can be identified and replicated, it can be used for treatment or prevention. While manuscript I and the work in progress I have only addressed positive correlations, the work in progress II have investigated negative correlation as well.

3.1 Definitions of co-morbidity

It is common to study one given index disease and name all diseases co-occurring with it co-morbidities [36]. The index diseases could also be replaced by a risk factor (e.g. being a blood donor [19]), a surgical procedure (like in the work in progress II) or other medical events. Presence of co-morbidities often gives rise to complications to treatment of diseases and generally affect the quality of life [37,38]. This is especially a problem for chronic diseases. For example, presence of COPD can cause complications among patients with myocardial infarction [39–41]. Therefore, some studies require that the co-morbidities are chronic diseases. The Charlson index is used to assess the increased mortality for patients with chronic co-morbid diseases [42]. Given the presence of one or more of 19 primarily chronic diseases, the Charlson index is used to assess short-term mortality.

Temporality is a key factor to consider in the definition of comorbidity. When the aim is to predict future disease, it is of obvious necessity to resolve the order of diseases. This is discussed in details in chapter 4 and 5. For some diseases, it takes several years for patients to develop complications, for example, retinopathy in diabetes [43]. In these

cases, a long time span is needed to detect the complications. In our study of colectomy described in the work in progress II, it was of interest to remove the initial co-morbidities to be able to identify the long-term effects. Colectomy patients will be weakened by extensive surgery, and there is in an initial increased mortality risk caused by infections in addition to complications [44].

3.2 Large-scale correlation analysis

3.2.1 Previous large-scale studies

Doing large-scale disease correlation analysis implies that all diseases are tested against all diseases for significant correlations. Only few large-scale studies that use registry data exist until now. The largest study so far is the study of Mendelian diseases in manuscript II, which covers more than 100 million patients from the US and Denmark. The second largest study uses 3 years of US Medicare data covering 13 million patients [4]. The study focuses on analyzing the correlations as a network, proving network properties (such as importance of "hub" diseases that have many significant correlations to others) and in general demonstrating that methods from network analysis are applicable for the analysis of disease correlations.

3.2.2 Correction for multiple testing

Data-driven analysis usually result in several thousands or hundreds of thousands significant pairs of diagnoses. Therefore, it is important to perform correction for multiple testing. If correction for multiple testing is not performed, the effective α (probability of no false positives) decreases exponentially with the number of tests. For 100 tests with an α of 5%, the effective α is above 99%, i.e. it is almost certain that false positives are among the hypotheses that are deemed significant. Two methods often used are Bonferroni correction and the Benjamini-Hochberg procedure. Bonferroni correction is the stricter of these two. While the Bonferroni corrections guarantees that each test has an effective probability for being falsely positive of less than α, the Benjamini-Hochberg procedure guarantees that the percentage of false positives in the significant set is maximum α.

Both methods can be used, and it depends on the data which method is preferable. A drawback of correcting for multiple testing is that only the strongest correlations, which are often the obvious, will remain significant. Since p-value is not a measure of "interestingness", it is undesirable to only investigate correlations with very small p-values. In cases where usage of Bonferroni correction would remove all significant pairs in the data or reduce the significant pairs to a few strong ones, the Benjamini-Hochberg method is preferable.

3.3 Work in progress I: Analysis of non-temporal co-morbidities

This section presents the work in progress on analysis of non-temporal co-morbidities. The main part of the data analysis is finished, but the results still need to be interpreted. The aim of the study was to use the registry data to identify novel disease correlations that could be explained by either genetics or disease mechanisms. All pairs of diseases co-occurring in a patient were analyzed irrespective of the order in which they occurred or the time between their occurrences.

3.3.1 Materials and methods

Data

We analyzed data from NPR, which is described in section 2.2 and manuscript I. We excluded codes related to pregnancy (ICD-10 chapters XV and XVI), general symptoms and signs not linked to a disease (chapter XVIII), external causes (chapters XIX and XX), and administration (chapter XXI). We subsequently rounded ICD-10 codes to level-3. This rounding reduces the dataset to 20.7 million unique patient–diagnosis associations covering 1,135 ICD-10 codes. This also limits the number of statistical tests, which need to be performed, from 1,022,534 to 460,657.

To address biases from age and gender, the population was split into age and gender bins to stratify the correlation measure. The age bins were defined by birth decade (see supplementary figure S1 in Manuscript I). Most of the patients from the pre-1900 bins are deceased by 1996 and these bins were therefore removed. The bin for people born in 2010 only cover one year and was also removed.

Correlation measure

RR is a measure commonly used in medical and epidemiological research, and was used for this work as well as in manuscript I. RR is defined as the probability for some event in one group of subjects divided by the probability for this event in another group of subjects. The groups are often case group versus control group, or treated versus untreated. In this work, the case group was the observed patients with a pair of diseases, and the control group was what would be expected by random from the total population given no correlation between the diseases. Thus, the full population was used to calculate the frequency for both "groups".

Statistical tests are used to identify if a RR is large by random or not. This can be combined with a threshold on RR. For example, it can be required that the RR value is 2 (corresponding to 100% increased risk) for the correlation to be subject to further studies. Any such threshold is of course arbitrary.

An issue with RR is that it overestimates the correlation of rare diseases. If both diseases in a pair are rare, the expected value will be small which can result in high RR values based on few patients. For example, a pair of diseases where both occur in 200 out of one million patients will have a RR of 25 if they co-occur in one patient. An alternative to RR is Pearson's correlation, which correctly estimates pairs of rare or frequent diseases, but underestimates the correlation when one disease is frequent and the other rare. The two measures were used and compared against each other in [4], where they

resulted in different sets of significant pairs. However, it is hard to say if one is more correct than the other. We choose to use RR as it is widely used and accepted within the medical community.

For a pair of disease A and B where there is N patients in the full population, N_A patient have disease A, N_B have B, C_{AB} have both disease, RR is given by the ratio between the observed frequency of A and B together and the randomly expected frequency,

$$RR = \frac{\frac{C_{AB}}{N}}{\frac{N_A}{N}\frac{N_B}{N}} = \frac{C_{AB}}{N\frac{N_A}{N}\frac{N_B}{N}}$$

It can be noted that when using the same population for both the observed and expected frequency, the definition becomes equal to ratio between the number of observed patients and the number of patients expected by random. Correction for age and gender was performed by changing calculating the number of patients expected by random for each age and gender bin and sum them,

$$RR = \frac{C_{AB}}{\sum_{bin}\frac{N_{A,bin}N_{B,bin}}{N_{bin}}}$$

Bins were removed if one or both diseases were assigned to less than five patients.

To address the issue that RR is overestimated for pairs of rare diseases, we added a pseudo count to the observed and expected value. For a pseudo-count of s, the formula for the pseudo count corrected relative risk (RRP) is,

$$RRP = \frac{C_{AB} + s}{\sum_{bin}\frac{N_{A,bin}N_{B,bin}}{N_{bin}} + s}$$

RRP has a maximum value of $(C_{AB} + s)/s$. A pseudo count of $s = 1$ will limit to RRP is maximum $C_{AB} + 1$. Thus, any pseudo count equal to or larger than 1 will limit the RRP of pairs with few co-occurrences. To select the pseudo count, we calculated the RR for integer pseudo-counts between 1 and 10. We evaluated average number number of patients with both diseases for the pairs with the 1000 largest RR values (figure 3.1) and found 3 to be an appropriate pseudo count.

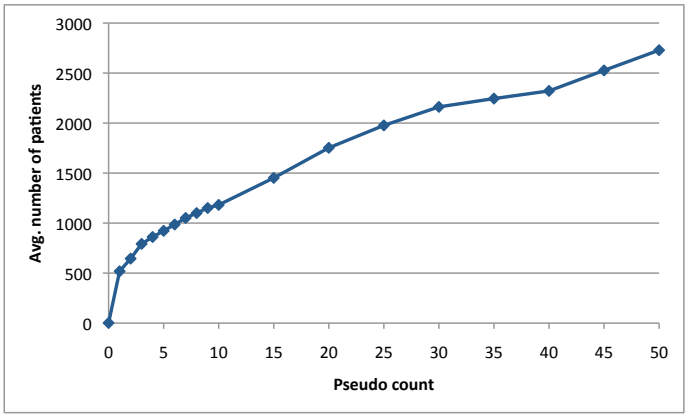


Figure 3.1 Selection of pseudo count. The 1,000 pairs of diagnoses with the largest RR were identified for a range of pseudo count values. For each value, the average number of patients who have both the diseases from a pair was subsequently found. The plot shows the average number of patients plotted as a function of pseudo count. When not using pseudo count, the top 1,000 pairs have 1.9 patients on average, while for a pseudo count of 1 the top 1,000 have 518 patients. The slope of the function decreases for pseudo counts larger than 3. Therefore, we chose to use a pseudo count of 3. However, any value between 1 and 3 would also have been acceptable.

Fisher's exact test and the Mantel-Haenszel tests were used to calculate p-values for the estimates. Fisher's exact test was applied to cases with patients from only one bin, and Mantel-Haenszel was applied for cases with patients from more bins. We used a $p > 0.05$ significance cutoff and applied Bonferroni correction to address the issue of multiple testing.

Clustering

We performed clustering on the statistical significant of pairs diagnoses (Bonferoni corrected $p<0.05$, $RRP > 1$) to obtain an overview of the comorbidities and identify groups of strongly correlated diagnoses. The MCL [45] clustering algorithm was used for this purpose. In order to focus on positive correlations and have a zero-point in the data, pairs with RR values less than 1 were removed, and 1 was subtracted from the remaining,

$$RRP_{cluster} = RRP - 1 \text{ , for all pairs having } RRP > 1$$

MCL has one key parameter, the inflation factor, which influences the size and number of clusters. It can be chosen arbitrarily. As the aim of clustering was to find groups of strongly correlated diagnoses, we quantified the clustering result by the ratio of RR within the clusters to RR between the clusters. If there are N_C diagnoses in the cluster and $N_{diag} - N_C$ diagnoses outside the cluster then the ratio will be,

$$\frac{\frac{2}{N_C\left(N_C-1\right)}\sum_{A,B\in within} RRP_{A,B}}{\frac{1}{N_C\left(N_{diag}-N_C\right)}\sum_{A,B\in between} RRP_{A,B}}$$

The inflation factor was selected by running the MCL algorithm for a range of different inflation factor values and calculating the ratio (figure 3.2).

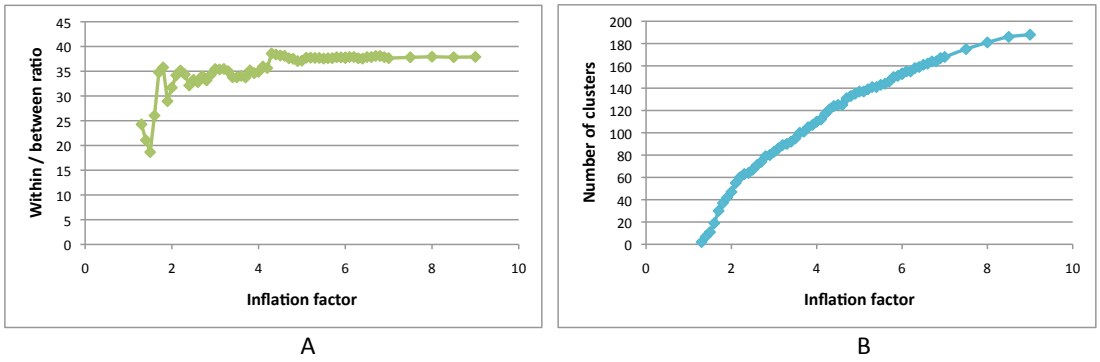


Figure 3.1 Selection of inflation factor for diagnosis clustering. (A) The ratio of average RR within and average RR between clusters as a function of inflation factors. The highest ratio is obtained for the inflation factor of 4.3. (B) The number of clusters as a function of inflation factors. There are 121 clusters for the inflation factor of 4.3.

A second round of clustering was performed, where the diagnoses from each cluster were concatenated into a meta-node, and the diagnoses of each patient in the original dataset were mapped to these meta-nodes. The patient-meta-node associations were used to calculate a RR between meta-nodes. Finally, MCL was used to cluster the meta-nodes. The inflation factor was selected in the same manner as with the initial clustering (figure 3.3).

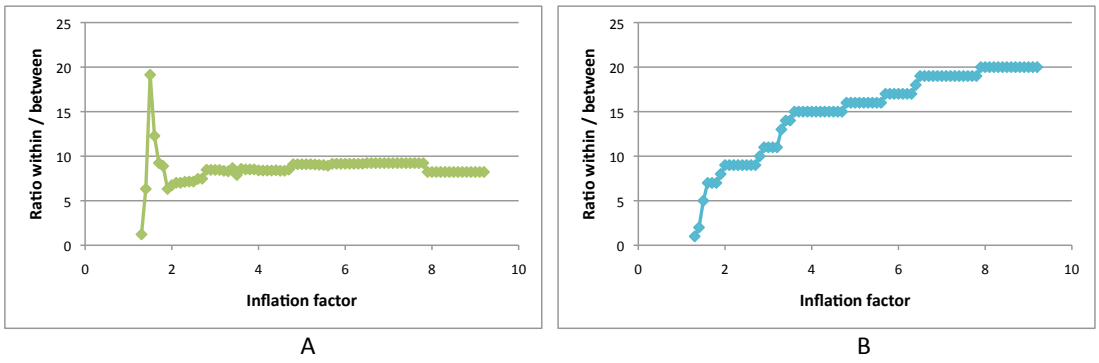


Figure 3.3 Selection of inflation factor for meta-node clustering. (A) The ratio of average RR within and average RR between clusters as a function of inflation factors. The highest ratio is obtained for the inflation factor of 1.5. However, this results in only five clusters, and the high ratio can be explained by a low number of links between clusters. For inflation factors above 1.9, the highest ratio is obtained for 6.5-7.3. (B) The number of meta-clusters as a function of inflation factors. There are 19 meta-clusters for the inflation factor in the interval 6.5-7.3.

3.3.2 Results

We identified 460,657 pairs of diagnoses that were co-occurring in five or more patients from at least one of the age and gender bins. Of those 148,081 were found to be significantly overrepresented ($p<0.05$, RRP>1).

In order to comprehend and visualize the full set of population wide comorbidities and to reveal strongly inter-connected groups of diagnoses, we performed MCL clustering and found 121 clusters. To display higher-level structure in the comorbidities we performed a second round of clustering, where all diagnoses within clusters were concatenated into 19 meta-clusters (clusters of meta nodes), see figure 3.3 and

supplementary table 1 (p. 92). The clusters tend to contain diagnoses from the same ICD-10 chapter. For example, one cluster in meta-cluster 5 only contains psychiatric and behavioral diagnoses (green) and five out of ten clusters in meta-cluster 4 are mainly neoplasm diagnoses (brown). Other clusters contain diagnoses that correlate because they affect the same anatomy: In meta-cluster 6, the central cluster mainly consists of cancer of the lip and oral cavity together with diseases of oral cavity, salivary glands and jaws.

We considered two approaches to infer knowledge using the clusters. One is to study the significant correlations within each cluster. It can be speculated that the diseases in the same clusters share etiology, which is especially interesting for diagnoses from different chapters. Another is to study the diseases that do not cluster together. For example, two clusters contain the majority of psychiatric and behavioral diagnoses (in meta-cluster 1 and 5). They are likely to represent two patient groups, and it can be interesting to do further studies that explain why they are not clustered together. This can be examined by doing patient stratification based on the clusters: The patient can be clustered based on their associations to meta-nodes. Patient groups that share diagnoses from the same meta-nodes can be identified, and it can be seen if the meta-nodes correspond to different patient groups. These patient groups can be further characterized by age, gender and other shared diseases. This would be similar to an earlier study where patient stratification was based on diagnosis data from a psychiatric hospital [5].

3.3.4 Further analyses

It is important to assess the robustness of the clustering. Some clusters can be connected by few diagnoses that correlate strongly with many diseases. If they are removed from the dataset before clustering, it is possible that the clustering will look completely different. This effect can be tested by doing cross-validation: The set of diagnoses (or correlations) is randomly split into N parts, and the clustering is performed on the data set with one of the N parts left out. The stability of the clusters can be assessed across the N different clustering results. The pseudo count also has a significant effect on the clustering. We also applied the clustering on the full data (i.e. also non-significant correlations) without the pseudo count correction, where we found 141 clusters. These clusters consisted of a very heterogeneous combination of ICD-10 chapters, which is opposite to what we found in the pseudo count corrected data. To test the effect of the pseudo-count, the clustering algorithm can be applied to data sets with different pseudo-counts. An important thing to keep in mind when doing these two analyses is to select the inflation factor in a stringent manner, so the difference is not due to difference in inflation factor.

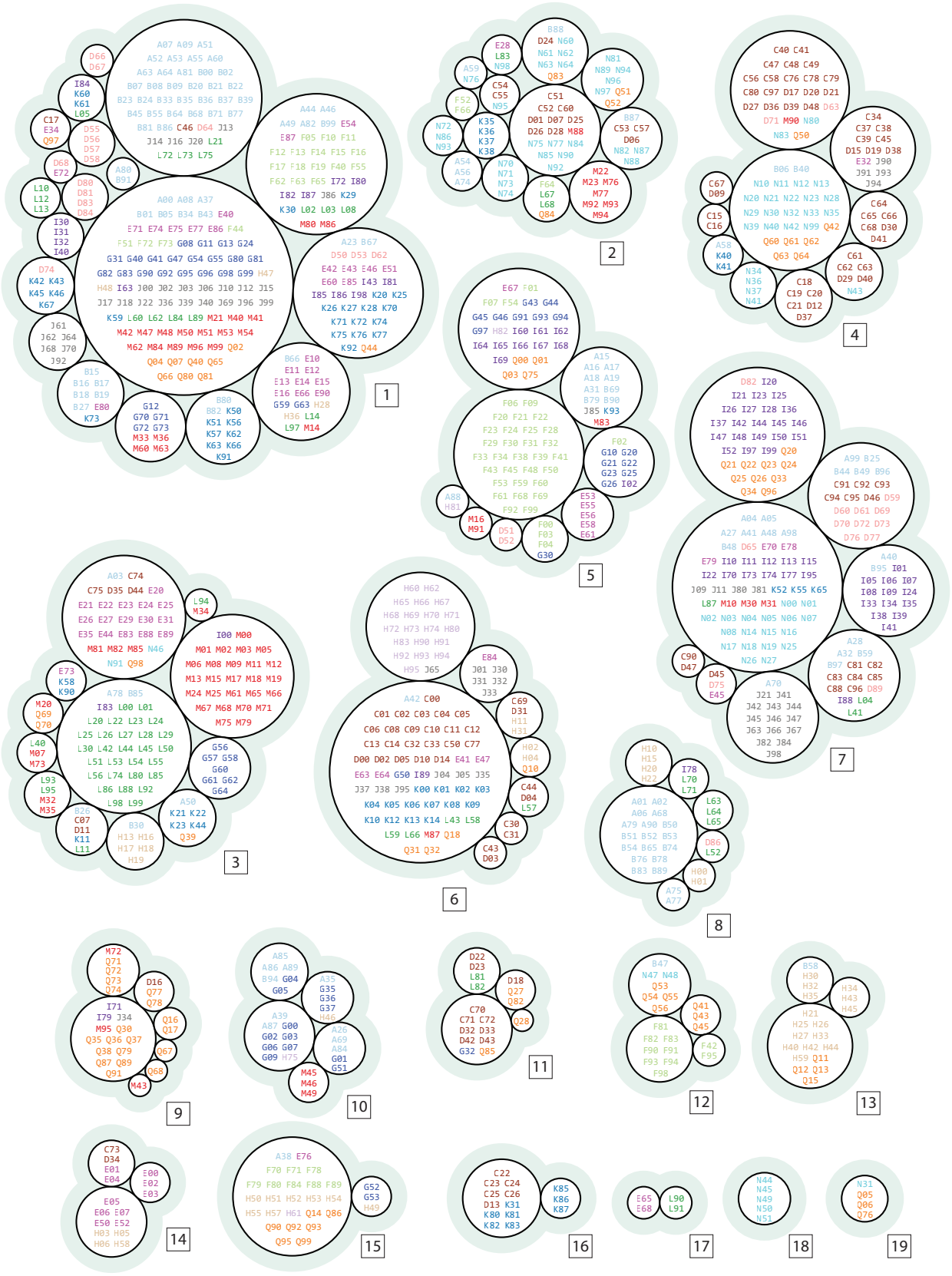


Figure 3.3 Result of the diagnosis clustering and meta-node clusteringEach circle is one of the 121 diagnosis clusters (or meta-node). The diagnosis codes are listed alphabetically within each cluster. The diagnosis codes and clusters are listed in supplementary table 1 (p. 92). The gray background fill around the clusters shows the meta-clustering. The numbers (1-19) refer to the meta-clusters. The layout of the clusters and meta-clusters is arbitrary.

3.3.4 Other results

Initially, we attempted to do reduce the set of correlations for manual curration, by grouping the correlations based on chapter and other properties. For example, codes from several chapters, such as injuries, were assigned to one group. Another group was disease pairs from the same ICD-10 chapter, which are likely to be similar. The result was nine groups where one group contained the most interesting correlations. Given more time, the groups can be evaluated further by picking random pairs from each group, and manually rating if they are obvious or interesting. However, the results were abandoned in favor of using clustering to group the diagnoses.

3.3.5 Discussion

In this work, it is shown how data from millions of hospitalizations can be compacted into clusters of co-occurring diseases that may even be reclustered into forming an overview of comorbidities as a macro-pattern. At the same time, these clusters illustrate fine-grained correlations between individual diseases that may have different phenotypic manifestations and belong to different ICD-10 chapters. This view complements the traditional diseaseome visualizations.

The major challenge in this work is the large amounts of significant co-morbidities identified. This is a challenge that applies to many fields that uses data-driven methods, and no general solution has been found for disease correlations yet. In [4], it is argued that methods from network analysis can be used to explore the correlations. However, it is not discussed how this will help to identify translational useful cases. The clustering approach presented here is one way of reducing the set of correlations to a smaller set of strong correlations. It is possible that even better results could be obtained by using clustering that allows one disease to be in more than one cluster, i.e. fuzzy clustering. This is likely to reflect the complex structure of diseases correlations.

Chapter 4: Temporal correlations and disease trajectories

The important aspect of temporality in disease correlations has been analyzed in manuscript I and the work in progress II (further described in chapter 5). The two studies use different methodology: In manuscript I, temporal correlations are analyzed in a data-driven manner. The correlation analysis is used to study disease progression over time. In the work in progress II, methods from survival analysis are used to assess the risk of cardiovascular diseases following a total colectomy procedure.

4.1 Temporal correlations

The temporal order in which diseases manifest themselves can be either random or determined by some known or unknown mechanism. In some cases, the temporality can be explained trivially by the age distributions of the diseases. Diagnoses given to newborns will obviously precede a wide range of diagnoses; cancer prevalence is higher among older people etc. HIV and AIDS is an example where there is a clear directionality. However, this does not imply that HIV is always diagnosed before AIDS. In some cases, the first HIV diagnosis can even be wrongly assigned to the patient after the outbreak of AIDS. Furthermore, the "symptom" diagnoses are in many cases diagnosed before the underlying disease causing it, for example, pneumonia in AIDS patients. Therefore, it is generally not possible to deduce causality based only on registry data.

Two methods were used to infer the temporality of correlations. The results from the non-temporal analysis were disregarded for both methods. We examined pairs of diseases (1 and 2). First, we calculated the RR of getting a disease 2 within a time frame given an occurrence of disease 1 (and the other way around). Then we used a binomial test to infer the temporal directionality of the significant pairs of diseases (i.e. testing if one disease occurs before the other). A correction for age was considered for the binomial direction test to correct for cases where the directionality was due to large difference in the age distributions of the diseases, as discussed above. However, this was found to be too conservative: the correction was punishing for the same temporal signal we were trying to identify. Given a maximum time frame of only 14.9 years, patients roughly belong to the same age group during the entire period. Therefore, this correction is not important.

Initially, we estimated RR and tested the significance of different time frames to be able to study the temporal pattern of each disease pair. However, with the focus on disease trajectories, we needed only to know if one disease increased the risk of the other. Therefore, only one time frame was used. Instead of manually looking through lists of correlations, the trajectories were used to identify the interesting temporal correlations.

4.2 Disease trajectories

The progression of a disease over time is also called a disease trajectory or illness trajectory. There is no formal definition of what constitutes a disease trajectory. Studies of trajectories focus on topics such as the age and gender's influence on the trajectory [46], conditions such as pain related to the disease [47] or how to give a patient the best nursing treatment [48,49]. An area that needs to be analyzed to realize the goals of P4 medicine is how one disease progresses into other diseases. This knowledge is of key importance to be able to make early diagnostics and infer the effect that different genotypes have on disease progression.

Ideally, a disease trajectory should be described using all confounding data that affect the outcome of diseases: Medical history, treatment, genetic data and environmental data. The NPR contains data that is crucial to make inference on disease trajectories: Diagnoses, treatment and surgery history along with age and gender, which are two of the most important confounding factors. Manuscript I uses the diagnosis history with the aim of describing the disease trajectories for a full nation in a data driven manner.

4.3 Modeling disease trajectories

Since there are no prior attempts at studying disease trajectories in a large scale for all diseases, there are no existing models for this purpose. The appropriate choice of model depends on the aim. Descriptive statistics of the trajectories require no model selection and can serve as a starting point for further analysis. If the aim is to study the effect of genotype or treatment on the disease progression, the models need to be able to handle stratifying factors. One possibility to address this need is to use state machine models, where a treatment switches to another state. Another possibility is usage of predictors such as neural networks that given the current medical and treatment history and genetics predicts future outcome.

Hidden Markov Models (HMMs) are state machine models that can be used for pattern recognition in various kinds of data, for example DNA sequences and protein [50,51]. In a study of cancer progression HMMs were used to predict metastases [52]. The input for HMMs is a sequence of tokens, which for example could be diseases. The states for a given string is not known, but can be estimated using the HMM. Each state is associated with a set of output tokens it can emit with some probability and a set of probabilities for transition to other states or remaining in the same state. For modeling disease trajectories, the tokens can be diagnosis codes and states can correspond to different stages of the disease progression. Given a sequence of tokens and a HMM, the likelihood that the sequence is a result of this HMM can be calculated. By estimating HMMs for a variety of disease progression, a patient can be matched to a certain progression, and the HMMs can be used as a base for predicting future outcome.

One concern with using HMMs for disease development is the Markov property of no memory: the next token in the sequence and its associated state only depends on the current token and state. This is not likely to be true for disease progression, where the full history for the patient is relevant for future states. Therefore, this issue has to be investigated further. Another main issue is to define the states and transition in a data driven setting where all diseases are considered. The questions of how many individual HMMs would be needed, which diseases each HMM should cover and which states and

transition patterns would be suitable are difficult to answer without further knowledge about disease trajectories. Thus, HMM did not fit our initial needs.

In order to avoid a long and time-consuming analysis of an appropriate model, we decided to use a descriptive approach, where the number of patients with a given disease trajectory was simply counted. We defined a disease trajectory as a set of diseases (ICD-10 codes) that must be observed in the patients in a given order. The first challenge with this approach is a combinatorial explosion: Given one thousand ICD-10 codes, there will be one million pairs to count and one billion triplets. To limit the number of trajectories to count, we restricted the trajectories to consist of pair-wise significant codes. Requiring the pairs to have a significant direction, so the first code is observed significantly more before the second code, restricted the diagnosis pairs to a set of 4,014, from which the trajectories were constructed.

4.4 Manuscript I: Patient-specific disease trajectories condensed from population-wide registry data

Anders Boeck Jensen, Pope L. Moseley, Tudor I. Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, Søren Brunak.

This manuscript presents the data driven work on identifying disease trajectories of disease development in the NPR data.

Patient-specific disease trajectories condensed from population-wide registry data

Anders Boeck Jensen[1,2], Pope Moseley[1,3], Tudor I. Oprea[1,3], Sabrina Gade Ellesøe[2], Robert Eriksson[1,2], Henriette Schmock[4], Peter Bjødstrup Jensen[2], Lars Juhl Jensen[2], Søren Brunak[1,2*]

[1]Center for Sequence Analysis, Department of System Biology, Technical University of Denmark, Kgs. Lyngby, Denmark.
[2]NNF Center for Protein Research, University of Copenhagen, Copenhagen, Denmark.
[3]Department of Internal Medicine, University of New Mexico, Albuquerque, New Mexico, USA.
[4]Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Copenhagen University. Hospital, Roskilde, Denmark

A key prerequisite for precision medicine is to estimate future disease progression from the history and current state of the patient. So far, disease correlations and disease trajectories have mainly been analyzed focusing on co-morbidities to a few index diseases or the progression of a single disease. Previous non-hypothesis driven analyses have been limited to time-unresolved diagnosis correlations, either without investigating the order in which diseases manifest themselves or analyzing data from a period of only a few years, which makes it difficult to resolve the order. These studies demonstrated the potential of using patient registry data for exploring co-morbidity networks, but the focus was on the properties of the networks rather than on defining a comprehensive set of disease or diagnosis trajectories. Here we present a non-hypothesis driven analysis of temporal disease progression patterns based on an electronic health registry covering the whole population of Denmark. We make use of the entire spectrum of disease and convert a 6.2 million patient registry into 1,171 statistically significant disease trajectories obtained from data collected over 14.9 years. These trajectories can be grouped into patterns that centre on a smaller number of keystone diagnoses, such as Chronic Obstructive Pulmonary Disease (COPD) and gout, which are central to disease progression and hence important to diagnose early to prevent adverse outcome. Our analyses also highlight the importance of stratifying patients both by age and gender and into inpatients, outpatients and emergency department visits.

Introduction

Population-wide analyses of disease correlations, co-morbidities and disease trajectories have so far mainly been carried out in a hypothesis driven manner with focus on a few diseases (Camilo and Goldstein, 2004; Finkelstein et al., 2009; Teno et al., 2001) or with focus on co-morbidities to index diseases (Petri et al., 2010). In contrast, this paper focuses on frequently observed temporal patterns over the entire spectrum of pathologies. Earlier data driven studies have used a network approach to analyze data covering three years of Medicare claims (mainly Americans 65 years or older) (Chen et al., 2009; Hidalgo et al., 2009). In this older population, investigators primarily identified diagnoses that were proximate to death (Hidalgo et al., 2009).

We aim for defining a comprehensive set of disease trajectories observed across an entire population, rather than mainly understanding the mathematical properties of co-morbidity networks as they have been constructed previously (Chen et al., 2009; Hidalgo et al., 2009). Both types of studies highlight the potential of using patient registry data for exploring co-morbidities and the temporal and non-temporal patterns they display.

The data foundation for the analysis is the Danish National Patient Registry (NPR), which covers all hospital encounters (inpatient admissions, outpatient visits, and emergency room visits) of the entire Danish population from 1996 to 2010. Mandatory reporting from all Danish hospitals to the NPR limits population bias. This data set covers 6.2 million patients with a total of 66.5 million hospital encounters and 101 million unique assignments of diagnoses coded in the International Classification of Diseases (ICD-10) terminology. The long time span and the size of the data set allowed us to analyze disease progression in the form of diagnosis trajectories. From the data we find 1,171 trajectories to have strong directional, statistical significance and thereby obtain a global picture of the most populated, directional co-morbidities observed in the clinic.

Results

As disease occurrences correlate strongly with age and gender, it is an obvious necessity to correct for these underlying baseline biases. Fig. 1 shows distributions of diseases across all 21 ICD-10 chapters in NPR stratified by age, gender and encounter type. Gender-specific differences in trends are clearly observable; most notably pregnancy-related diagnoses assigned to women primarily in the age interval 15–45 (green). For males, injury codes standout among young men (red), but Fig. 1 clearly shows that these diagnoses segregate much stronger by encounter type than by gender or age. These examples stand out as some of the strongest correlations between diagnoses and encounter type, but our analysis demonstrated that this trend holds true for most ICD-10 chapters. In this type of work we thus found that it is equally important to stratify diagnoses by the type of hospital encounter. An important consideration in the subsequent analyses was therefore to make use of this aspect to stratify diagnosis assignments into more precisely matched case–control groups. In this way we enable both the discovery of statistically significant correlations that would otherwise have been masked and the removal of statistically significant correlations that are trivially explained by encounter types.

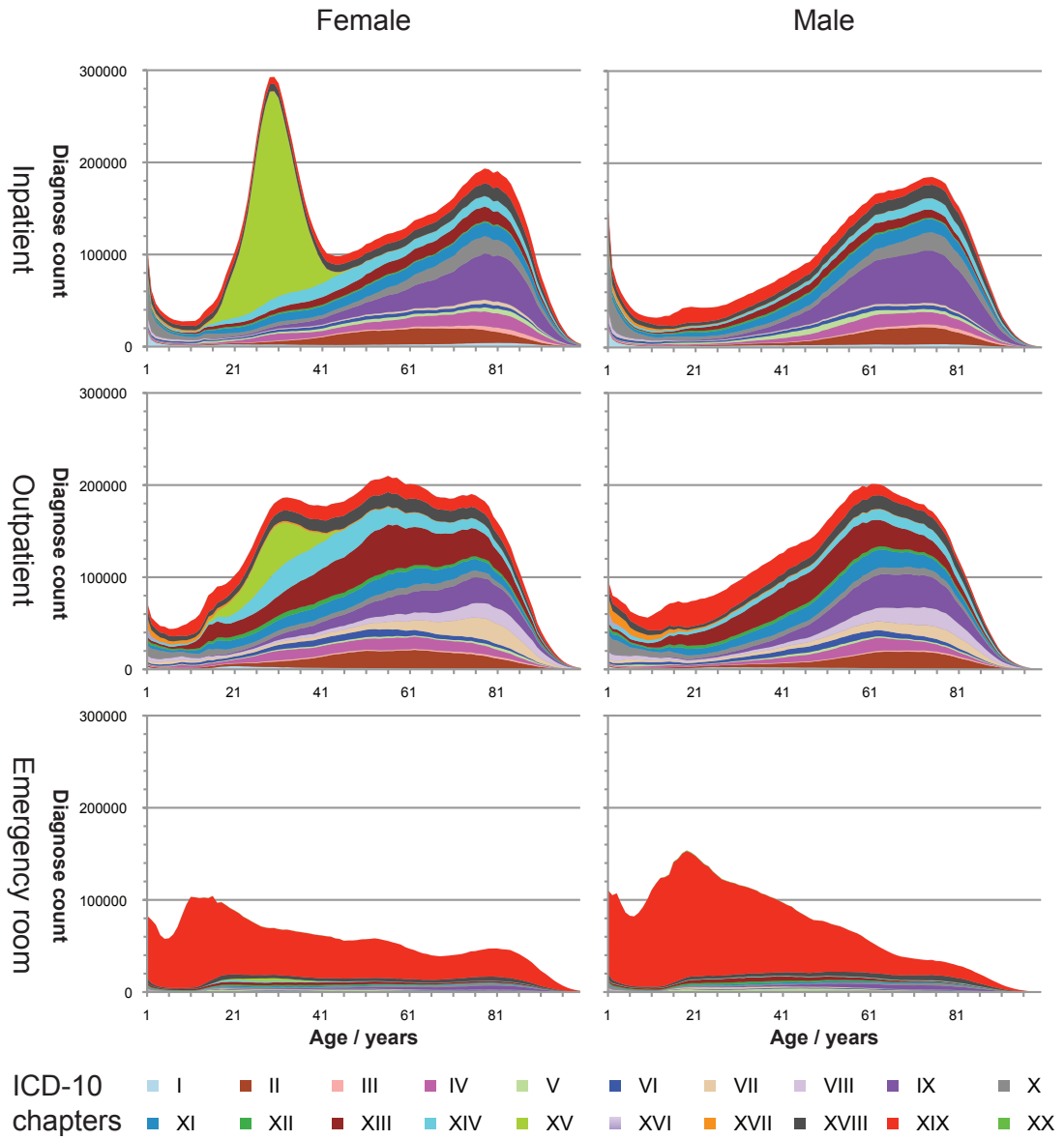


Fig. 1. ICD-10 diagnoses from the National Danish Patient Registry covering the entire Danish population in the period 1996-2010. The data panels show females (left), males (right), inpatients (top), outpatients (middle), and emergency room patients (bottom). The color-coding corresponds to ICD-10 chapter structure starting from the bottom by "Certain infectious and parasitic diseases", (chapter I, hardly visible), "Neoplasms" (chapter II, brown), 21 chapters in all.

To identify statistically significant, temporal correlations between diagnoses we used a case–control scheme where cases and controls were matched by age, gender and type of hospital encounter. The scheme consists of a pre-filtering step where initial p-values are estimated initially using a binomial test, and then confirmed using the full case-control matching.

From the full data set, we identified 1,194,343 pairs (D1 → D2) of diagnoses where D2 occurs within a five-year time frame of D1. From them we excluded in total 370,737 codes related to pregnancy (chapters XV and XVI of ICD-10), general symptoms and signs not linked to a disease (chapter XVIII), external causes (chapters XIX and XX), and administration (chapter XXI). Among the 823,606 tested pairs, we identified 62,821 that were observed in at least 10 encounters, had RR > 1 and were significant in the pre-

filtering step with $p < 0.001$ (Bonferroni corrected for multiple testing). This number increases by ~10,000 if patients are stratified only by age and gender or only by type of hospital encounter, and by an additional ~8,000 when stratified by neither (see Table S1). Thus, stratification by type of hospital encounter is complimentary to and just as important as stratification by age and gender.

In addition to identifying pairs with increased risk, we tested for significant directionality: Of the pairs D1 → D2 with significant $RR > 1$ we identified those where significantly higher number of patients had D1 occurring before D2 compared to the opposite direction or in the same admission. In all, 4,014 pairs were found to have a significant direction. All pairs were validated using the sampling model to be significant with $p < 0.01$. Table S3 lists all the pairs and their corresponding RR and p-values.

The 4,014 directional pairs were then combined into longer trajectories consisting of three, four or more diagnoses. We identified a set of 5,784 trajectories with three diagnoses that covers between 1 to 16,197 patients in the last step (2,939 of them cover more than 100 patients). These were extended to 1,171 trajectories of four diagnoses and further to 52 with five diagnoses all covered by at least 20 patients in the last step. The entire set of recurrent trajectories of four diagnoses each is shown in Table S4.

To produce a more comprehensive overview we further clustered the trajectories based on which diagnoses they shared. As similarity measure between diagnosis pairs, we used the Jaccard Index. The clustering identified 15 clusters where the five largest respectively covered 46, 25, 12, 9 and 8 diagnoses. Below we focus on five of these trajectory clusters: diseases of the prostate, chronic obstructive pulmonary disease (COPD), cerebrovascular disease, cardiovascular disease, and diabetes mellitus. Below we describe their properties in detail.

The prostate disease cluster is the simplest, progressing from prostate hypertrophy (ICD-10 code: N40) through prostate cancer (C61) and obstructive uropathy (N13) to metastatic cancer (C79) and cancer-associated anemia (D63) (Fig. 2). Except for the expected prostate-specific complications, this nearly linear trajectory cluster is representative of general cancer progression to metastasis and anemia.

The COPD cluster has a characteristic bowtie structure (Fig. 3). It starts by a variety of diagnoses from multiple ICD-10 chapters, including cardiovascular, skin, endocrine, and behavioral disorders. These all converge on COPD (J44) and proceed to respiratory failure (J96), pneumonia (J15), septicemia (A41) and other diagnoses. We tested that COPD was a central diagnosis in the cluster by calculating the RR of COPD occurring between all diagnoses preceding and succeeding COPD. We found that COPD had a RR of 5.1 ($p<10^{-5}$).

There is wide acceptance that cardiovascular disease-related morbidities are worsened in patients with COPD (Curkendall et al., 2006; Finkelstein et al., 2009; Salisbury et al., 2007; Sidney et al., 2005). This association, however, generally considers the impact of cardiovascular events on pre-existing COPD or the co-existence of the two diagnoses. In contrast, our analysis demonstrates that a subsequent diagnosis of COPD has a profound impact on a number of cardiovascular diagnoses, whether angina pectoris (I20) or atherosclerosis (I70). In fact, all trajectories starting with atherosclerosis are followed by a subsequent COPD diagnosis, which supports the temporal pattern of diagnosis as well as pathophysiologic link. Once the diagnosis of COPD occurs, the disease trajectories tell a

story of rapid progression (typically 1.8–2.5 years) to a variety of subsequent diagnoses. However, the most common outcome after COPD appears to be death. Using a Kaplan-Meyer estimate, we found that 49.7% of patients following a trajectory containing COPD die within 5 years compared to 21.3% in a matched control cohort. Over the full data period (14.9 years) 86.9% of COPD trajectory patients die while 36.2% of patients in the matched cohort die. The high-mortality rate is confirmed in another study (Suissa et al., 2012): a 50% mortality at 3.6 years and 75% at 7.7 years from initial hospitalization.

Similar to the COPD cluster, the cerebrovascular and diabetes clusters are characterized by convergence on keystone diagnoses, namely epilepsy (G40) and retinal disease (H36), respectively (Fig. 4 and Fig. 5). Epilepsy (with an RR of 6.6, $p<10^{-5}$) is likely a marker of significant cerebrovascular compromise (Camilo and Goldstein, 2004) reflecting the severity of the underlying disease. Similarly, retinopathy (RR=20.1, $p<10^{-5}$) is a marker of the degree of system-wide diabetic vasculopathy (Moss et al., 2003); population studies suggest that diabetic retinopathy is present in more than half of all diabetics (Kohner, 1989).

Gout (M10), like COPD and retinal disease, is a keystone diagnosis (RR=6.8, $p<10^{-5}$) within the cardiovascular cluster and serves as central disease in a diabetes independent cardiovascular diseases cluster (Fig. 6). Associations between gout and cardiovascular disease have long been considered (Freedman et al., 1995), and allopurinol has recently been suggested for management of cardiovascular disease (Kelkar et al., 2011). In contrast, the recent CHARGE study failed to show a link between serum uric acid and cardiovascular risk (Yang et al., 2010). Our population-wide trajectory data support the epidemiologic relationship between gout and cardiovascular diseases.

In addition to providing an important analysis of temporal disease associations across an entire population, the trajectories and their associated networks may be quite useful in refining study groups for comparative effectiveness research (CER). For example, the 4 diagnosis trajectories beginning with angina (I20) ending with cardiac arrest (I46) include several combinations of disease diagnoses in the second and third positions. Chronic ischemic heart disease (I25) significantly increases the risk of a cardiac arrest to 1.13 (1.12 to 1.42). Gout further increases the risk 1.99 (1.3-3.05). While numerous studies demonstrate the impact of ischemic heart disease in patients with renal failure (N18), renal failure as a subsequent diagnosis does not further increase the risk for cardiac arrest in the angina patient who develops chronic ischemic heart disease. Thus, one might use this data to design interventions based upon angina populations with and without gout, chronic ischemic heart disease, and so on.

Discussion

Systematically adding the temporal dimension to population-wide co-morbidity data has not been attempted previously using a non-hypothesis driven approach at this scale. We show for the first time that hospitalizations across an entire population of significant size can be used to extract and group trajectories as a novel way of describing biological disease progression and subsequently identifying keystone diagnoses.

In this work we defined a diagnosis trajectory as an ordered series of diagnoses where the diagnoses were assigned to patients in the specific order. The order had to be followed strictly for a patient to be considered following it. Thus, for a trajectory starting with the

diagnoses D1 → D2 patients who had the diagnoses assigned in order D2 → D1 → D2 again were not considered as following the trajectory. In cases where multiple diagnoses from the trajectory were assigned in the same discharge, they were considered to be in the correct order.

This strict definition puts limits on how much variability each single trajectory is able to cover. It means that bifurcating diagnoses which later converge on other diagnoses are not included in a single trajectory. Nonetheless, using our trajectory clustering approach we were able to cover this type of disease progression as well. Interestingly, using the clustering approach we were further able to reduce the patterns of disease progression from 1,171 individual trajectories down to less than ten major clusters covering most of the populated paths through the disease terminology space. As the underlying data foundation here is considerable and unbiased, this condensation is remarkable. However, it does also obviously reflect the numerical constraints set of the requirement for directionality in terms of relative risk.

In terms of trajectory interpretation it is essential eventually to establish to what extent the directionality reflect underlying causal patterns or not. For example, it is interesting to speculate whether the disease state associated with the COPD diagnosis is the cause, or whether COPD is an ICD-10 surrogate for a variety of factors associated with increased morbidity, such as smoking, adverse effects of medications, or poor general health. The expected high degree of association between COPD and atherosclerosis supports the obvious smoking linkage. Hospitalization for pneumonia in the setting of underlying chronic disease is common, with odds ratios reported at 4.4 for COPD and 3.24 for heart failure in one large study (Farr et al., 2000). Our model suggests an even more profound effect of these chronic conditions on pneumonia within a relatively short time span. COPD as a subsequent diagnosis of many trajectories may also demonstrate a medical systems issue, that of undiagnosed COPD, which becomes manifest only after another serious diagnosis is made. It is likely that COPD is coexistent with the initial diagnosis of most trajectories, yet it occurs as a second data point in multiple trajectories.

Our findings demonstrate that the population-wide disease trajectory approach uncovers diagnosis linkages that have had unclear or conflicting relationships through epidemiologic or smaller sample case control approaches. We further demonstrate the importance of patient stratification and that stratification by type of hospital encounter is as important as stratification by age and gender.

The trajectories have also a predictive potential where preceding steps can be used as a basis for predicting the most likely next step in the disease progression. A major additional perspective in using the catalog of disease trajectories established here is obviously to use them in the context of precision medicine and combine them with detailed molecular level characterization of each patient for better disease management of individual patients along the course each patient will take.

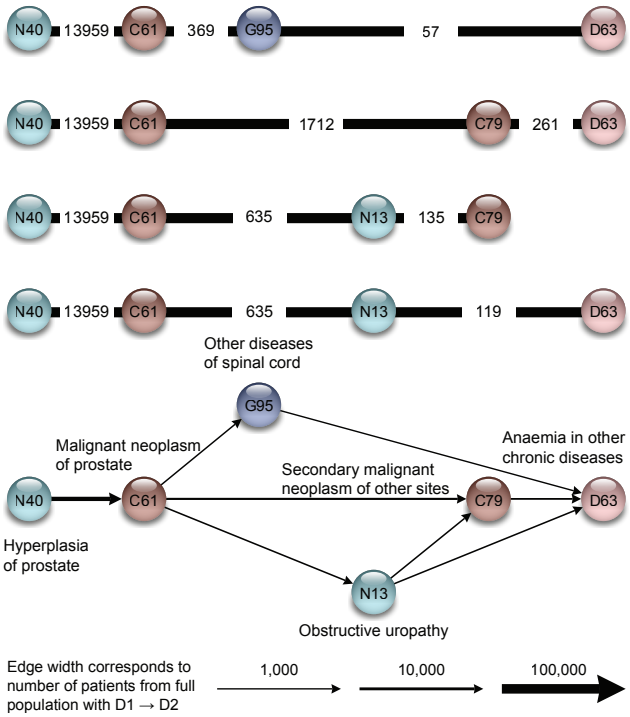


Fig. 2. Disease trajectories and trajectory-cluster for prostate cancer. The figure illustrates the transition from trajectories to a trajectory cluster. Each circle represents a diagnosis and is labeled with the corresponding ICD-10 code. The colors represent different ICD-10 chapters. The temporal diagnosis progression goes from left to right. **A**) All trajectories that contribute to the prostate-cancer cluster. The number of patients, who follow the trajectory until a given diagnosis, is given in the edges. **B**) The prostate cancer trajectory-cluster that represents all the trajectories. The width of the edges corresponds to the number of patients with the directed diagnosis pair from the full population. The cluster describes a normal progression from having hyperplesia of prostate diagnosed to having prostate cancer, cancer metastasis and anemia.

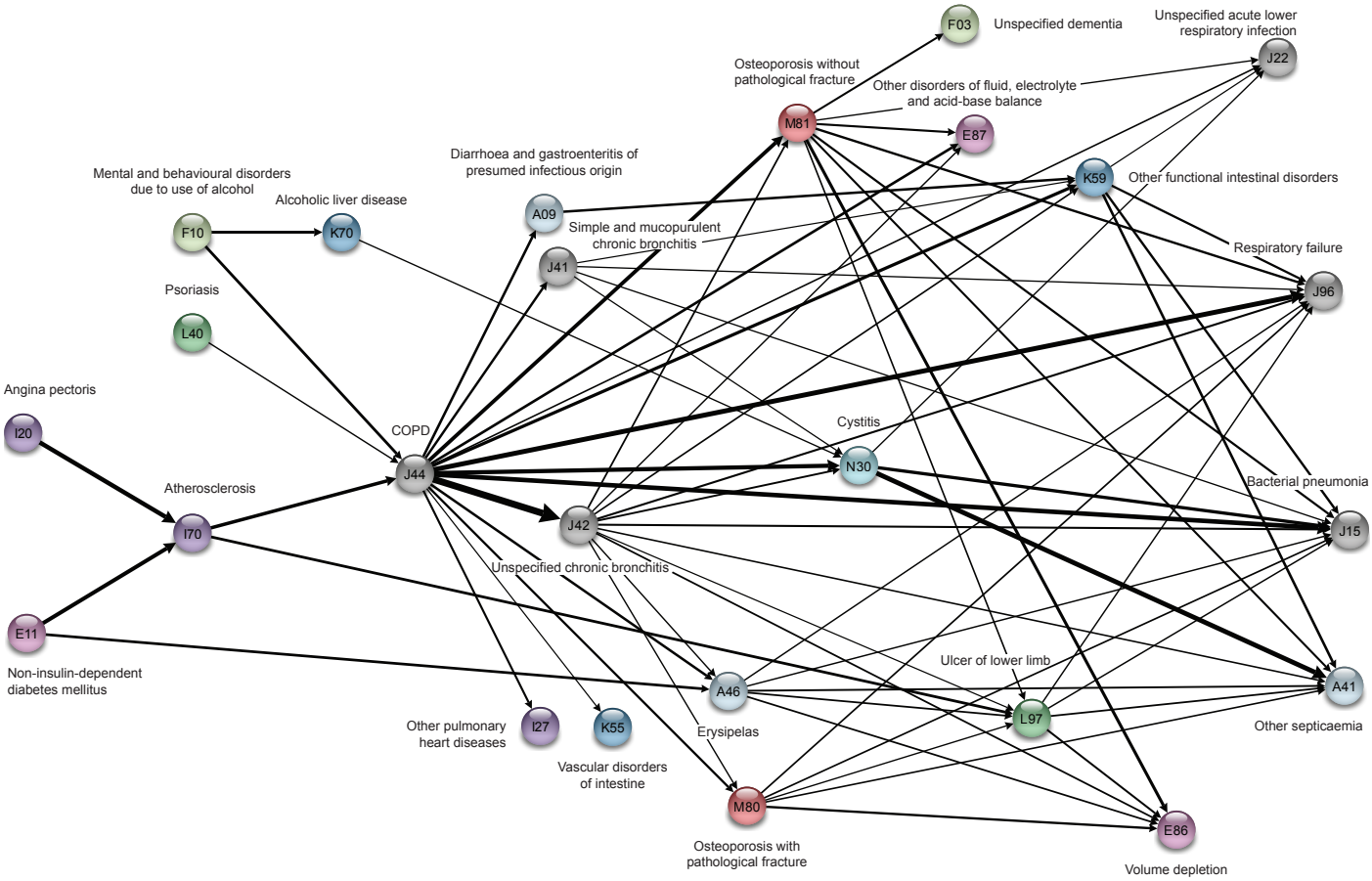


Fig. 3 COPD disease trajectory clusters. **A**) The COPD cluster showing five preceding diagnoses leading to COPD and some of the possible outcomes.

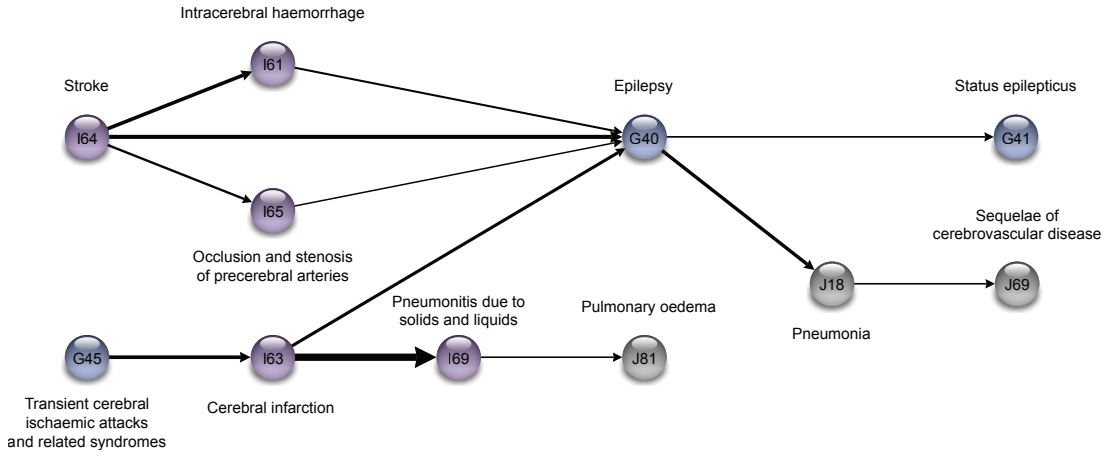


Fig. 4 Cerebrovascular Cerebrovascular cluster with epilepsy as key diagnosis.

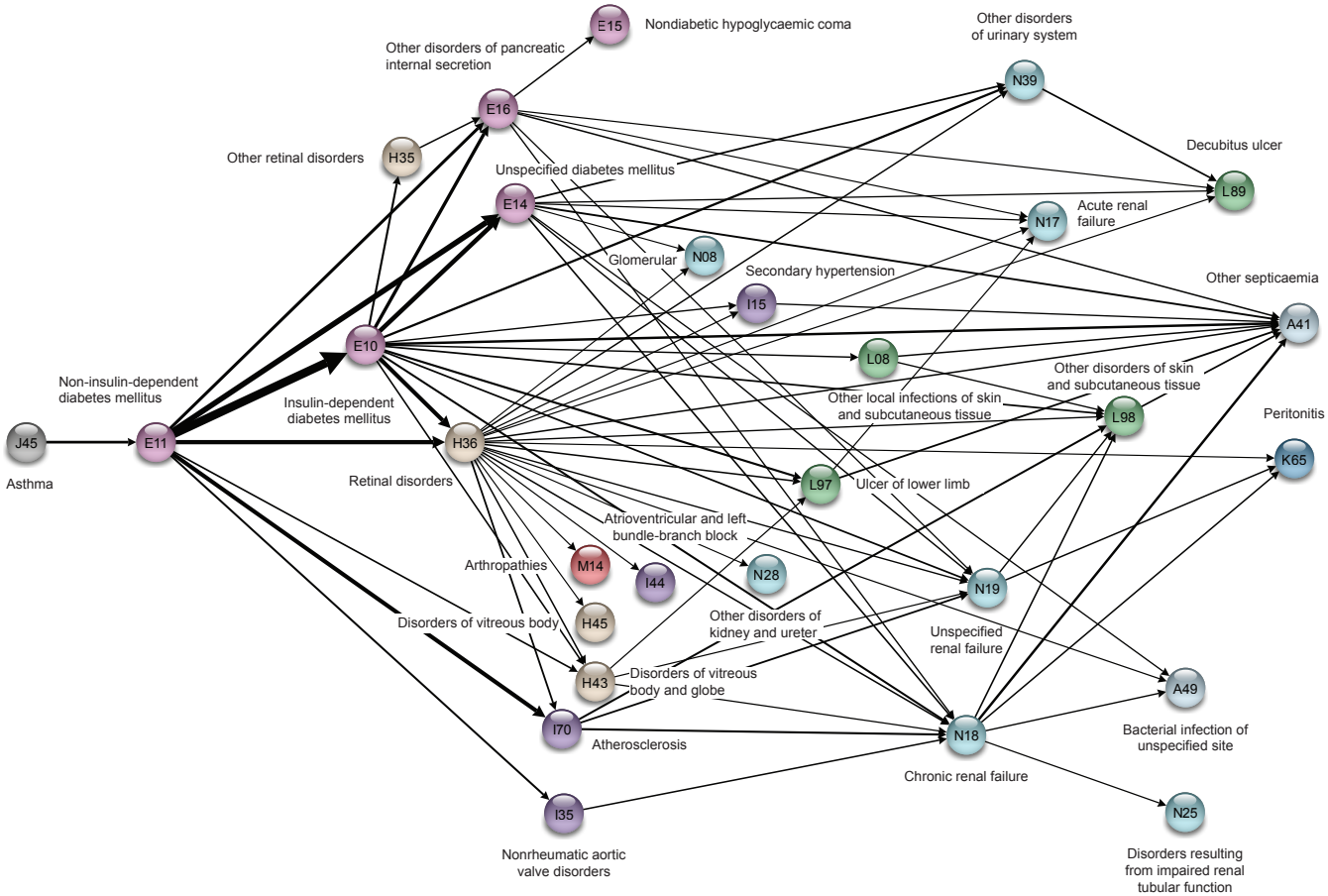


Fig. 5 Diabetes disease trajectory clusters. Diabetes cluster showing progression from non-insulin dependent to insulin dependent diabetes. Retinal disorders are key diagnoses marking progression to worse conditions.

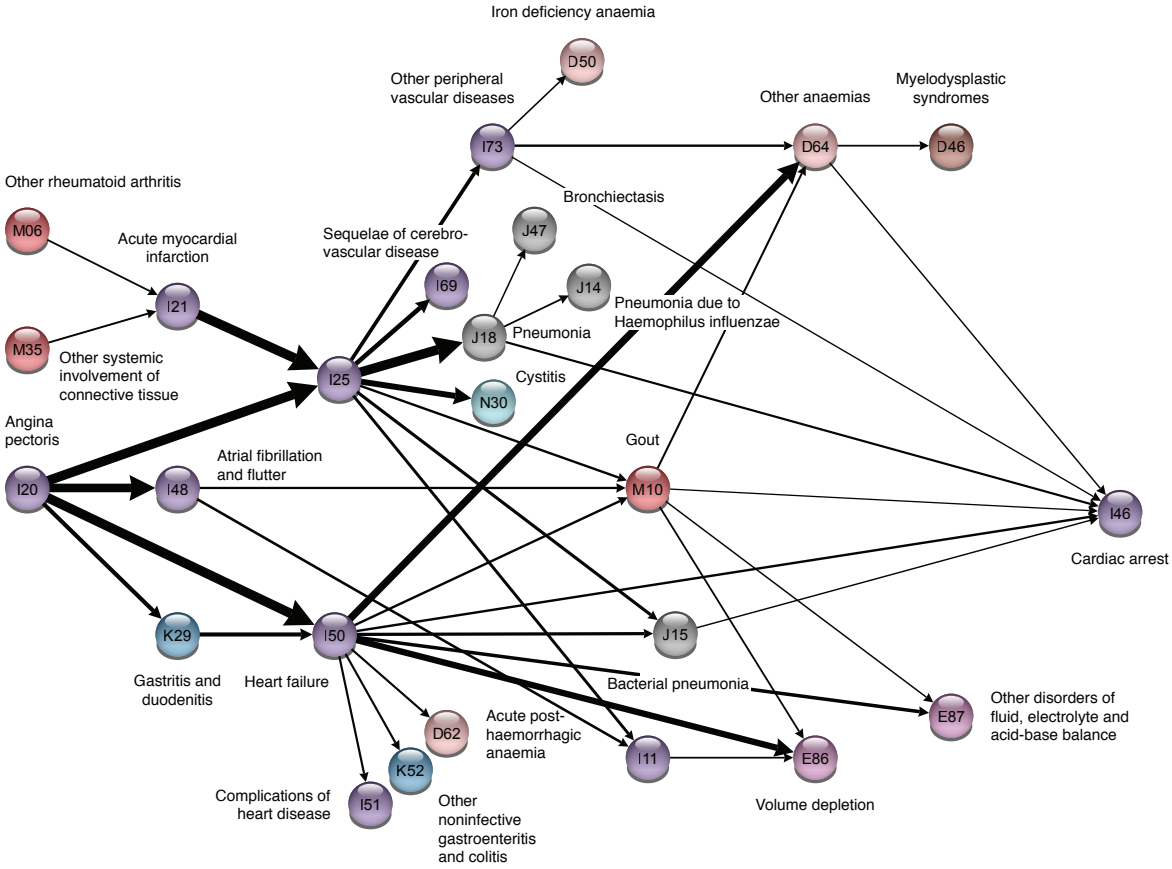


Fig 6 Cardiovascular disease trajectory clusters. A key finding is that gout is a central diagnosis in the cardiovascular cluster supporting evidence that gout is important to progression of cardiovascular diseases in a keystone manner.

Materials and Methods

Study design

The objective of this retrospective study was to identify and characterize disease trajectories using population-wide disease registry data using a data driven approach. The trajectories were derived using pairs of significant time-dependent diagnose correlations. A case-control scheme was used estimate the correlations' strength and p-value.
The data used in the analysis is from the Danish National Patient Registry (NPR), which contains administrative information and primary and secondary diagnoses coded in the International Classification of diseases (ICD-10) covering every hospital contact in Denmark. It includes public and private hospital visits and covers all types of encounters: inpatient (admitted to the hospital with overnight stay), outpatient (visit without overnight stay) and emergency department contacts.

The data set covers the period January 1996 to November 2010, and includes 68 million records for 6.2 million individuals. For inpatients the records cover the time between admission to a ward until discharge to either another ward or out of the hospital. Records covering two or more discharges between wards were combined into one covering the entire admission. In cases of re-admissions the same or the following day after a discharge the records were also combined. Doing this 1.5 million inpatient records were combined with other records giving 66.5 million encounters in total. Because private hospitals have only reported contacts from 2002, all of these, in total 1 million, were removed to maintain an unbiased data set. Private hospitals in Denmark approximately handle less than 1.6% of all admissions (routine treatment) and add less than 1% unique patient-diagnosis associations as 38.4% of the patient-diagnosis associations from private hospital are already covered by the public hospitals.

The ICD-10 system has a hierarchical structure, where codes can be rounded to a less specific parent diagnosis code, block or chapter. We used this structure to round all codes to level 3.

Diagnosis correlation measure

We used relative risk (RR) to measure the strength of a correlation between a pair of diagnoses (D1 → D2). RR is defined as the ratio of observed number of D1 discharges (discharges with D1 assigned) followed by a D2 discharge within a 5-year time frame and the number expected at random with no correlation between the two diseases. The expected number was assessed using a case-control scheme sampling random control cohorts. P-values for RR were obtained using a binomial test that models the sampling procedure and corrected for multiple testing using Bonferroni correction. Due to correction for multiple testing for 823,606 pairs, we needed more than 82 million samples for each pair. The binomial model was used as a pre-filtering step in order to avoid a total running time of several thousands of computing years if performing the full procedure. Pairs included in the trajectories were validated using the full random sampling procedure.

To account for confounding factors, control patients were matched to come from the same age and gender group as the case patient (as shown in Fig. S1). The type of encounter was matched for the D1 discharge. The season of the year and changes in diagnostic methods and focus over years are also confounding factors. We controlled for these by sampling the control discharge from the same week as the case D1 discharge.

Temporal correlation analysis

Given a pair of diseases (D1 → D2) the case cohort was formed by identifying all discharges with D1 assigned. Control cohorts were formed by matching each case discharge with a random discharge that fulfills the matching criteria (same age and gender of the patients and same type of encounter and week of the discharge), resulting in a number (N) of control cohorts. D1 discharges that have one or more subsequent D2 discharges within a five-year time frame were counted (Fig. S2). We denote the number for the case cohort as C_{cases} and as C_i, $i = 1 .. N$; for the control cohorts. RR is given by,

$$RR = \frac{C_{cases}}{\frac{1}{N}\sum_i C_i}$$

The p-value was calculated as the percentage of the control cohort co-occurrence counts that are larger than the observed co-occurrence count,

$$p = \frac{1}{N}\left|\left\{i \mid C_i \geq C_{case}\right\}\right|$$

RR was estimated using 10,000 sampled control cohorts.

In the binomial test modeling the sampling procedure, we considered each sampling of a single control discharge as a Bernoulli trial. Given the matching criteria there is a set of n_{match} patients to select the random control from. A number, c, of these discharges will have D2 discharge within the time frame. This number can be pre-calculated without any sampling. The probability of sampling a control with a D2 discharge within the time frame is,

$$P(\mathrm{D2}) = \frac{c}{n_{match}}$$

The probability distribution for the total number of sampled D2 discharges is the sum of all single Bernoulli trials. We approximated the distribution by with the average of the probabilities for all D1 discharges,

$$P(\mathrm{D2})_{test} = \frac{1}{n_{\mathrm{discharges}}}\sum_{i=0}^{n_{\mathrm{discharges}}}\frac{c_i}{n_{match,i}},$$

where $n_{\mathrm{discharges}}$ is the number of discharges with D1.

To make sure the binomial model is a valid substitution for the sampling giving p-values that are at least as conservative as the sampling procedure, we ran full sampling for 1,500 pairs and compared with the simplification. We expect the simplification to perform worst where the variance of the probabilities contributing to the average probability is high. Therefore, we tested the 1,000 pairs with the largest variance, while 500 others were chosen at random. Fig. S3 shows the true p-values plotted against estimated p-values. The binomial model was found to be more conservative than the sampled p-values for small p-value. Thus the simplification is a valid substitute for the sampling procedure. To further guard against false positives due to the binomial model, the significance cutoff level was set to 0.001.

Testing for directionality
The diagnosis pairs (D1, D2) that had RR > 1 and a significant p-value for one or both directions (D1 → D2 and/or D2 → D1) were tested for directionality. Binomial tests were used to identify pairs where significantly more patient had D1 assigned before D2 or the other way around. For this, the first D1 and D2 discharges for patients with both diagnoses were identified and the order for each patient established. The number of patients with each order of the diagnoses was counted: N_{D1} with D1 assigned first, N_{D2} with D2 assigned first and N_{same} with D1 and D2 in the same discharge. Using two binomial tests we tested if N_{D1} or N_{D2} were significantly larger compared to a binomial distribution (N_{D1} + N_{same} + N_{D2} samples with probability 50%). The p-values were Bonferroni corrected. If one of the tests showed a p-value less than 0.05 the pair was considered having a significant direction (only one of the tests can have significant p-value).

Diagnosis trajectories
We counted the patients as following a disease trajectory only if the patient had the diagnoses assigned strictly in the order specified by the trajectory. Each step of the trajectory corresponded to a single diagnosis.

We used the pairs of diagnoses from the temporal correlation analysis with significant direction to identify the trajectories to count. Trajectories with three diagnoses were obtained by combining pairs with overlapping diagnoses (D1 → D2 and D2 → D3 combined to D1 → D2 → D3). They were subsequently extended with more overlapping pairs to obtain even longer trajectories. A greedy approach was used to find the three long trajectories covering the most patients. The pairs were sorted in descending order according to their discharge count. Pairs with an overlapping diagnosis were found starting from the top of the list and the number of patients following the full trajectory was counted. We stopped when the three long trajectories had no patients following them. All four and five long trajectories were then counted. Using this approach, all the trajectories covering the largest number of patients can be identified without counting every possible trajectory.

Diagnosis trajectory clustering
In the 1,171 four long diagnosis trajectories we identified groups of trajectories having large diagnosis overlap and representing variants of general patterns of disease progression. To identify these patterns systematically, we performed MCL clustering (Schaeffer, 2007) that assigned each of the 140 codes that make up the 1,171 trajectories to a cluster. The Jaccard index was used as similarity measure (counting how many trajectories both diagnoses are part of and normalizing by the total number of trajectories either is part of). Trajectories with all diagnoses within the same cluster were combined into directed trajectory-clusters in which the patterns could be examined (Figs. 2-6).

Since the clustering was based on diagnoses, some trajectories had diagnoses from multiple clusters. Of the original 1,171 trajectories there were 378 that had all diagnoses within the same cluster. We increased this number to 608 by merging one smaller cluster into the largest and by including particular diagnoses to clusters if they contributed to complete trajectories with three diagnoses already within the cluster. In this way, some diagnoses appear in multiple clusters. Of the 608 trajectories 466 were within the largest cluster, 129 within the second largest cluster, 6 within the third largest, 5 within the

fourth largest and 2 in each their cluster. The second largest through the fourth largest cluster each revealed a distinct pattern of disease progression (the COPD, cerebrovascular and prostate cancer patterns) while the largest cluster had two major patterns in it: one focusing on diabetes mellitus and another focusing on cardio vascular diagnoses.

To divide the largest cluster into the two patterns, the diagnoses within it were once again clustered using MCL. We used the same similarity measure as before, but using larger inflation factor. This resulted in four new sub clusters, where the largest covered diabetes mellitus diagnoses. We merged the second and third largest sub clusters which together covered cardio vascular diagnoses. Finally, the five clusters were visualized by representing diagnoses as nodes and making directed edges between consecutive diagnoses for all the trajectories within the same cluster.

Verifying central diagnoses

In most of the trajectory clusters we identified a keystone diagnosis. In order to verify that they are important to the disease progression in the clusters, we for each key diagnosis counted how often it occurred between diagnoses preceding it and diagnoses succeeding it in the full population. We identified two sets of diagnoses: all diagnoses that could lead to the key diagnoses (the preceding set), and all diagnoses that could be reached from the key diagnoses (the succeeding set). Next we identified all patients who had one diagnosis from the preceding set follow by one from the succeeding set. Like when counting the trajectories, we discarded patients who have a diagnosis from the succeeding set before the first from the preceding set, and the diagnoses were allowed to occur in same admission. We then counted the patients having their first occurrences of the key diagnosis in the time from the first occurrence of a preceding diagnosis to the first occurrence of a succeeding diagnosis.

In order to evaluate the count of the key diagnosis, we calculated RR and assigned p-values by matching control patients using the same criteria as for the temporal correlation analysis. For each case patient we identified the number of days between the occurrence of the preceding diagnosis to the occurrence of the succeeding diagnosis. We counted the number of occurrences of the key diagnosis in the same period among the matched controls. The findings are summarized in Table S2.

References

Camilo, O., and Goldstein, L.B. (2004). Seizures and epilepsy after ischemic stroke. Stroke *35*, 1769–1775.

Chen, L.L., Blumm, N., Christakis, N.A., Barabási, A.-L., and Deisboeck, T.S. (2009). Cancer metastasis networks and the prediction of progression patterns. British Journal of Cancer *101*, 749–758.

Curkendall, S.M., DeLuise, C., Jones, J.K., Lanes, S., Stang, M.R., Goehring, E., and She, D. (2006). Cardiovascular disease in patients with chronic obstructive pulmonary disease, Saskatchewan Canada cardiovascular disease in COPD patients. Annals of Epidemiology *16*, 63–70.

Farr, B.M., Bartlett, C.L., Wadsworth, J., and Miller, D.L. (2000). Risk factors for community-acquired pneumonia diagnosed upon hospital admission. British Thoracic Society Pneumonia Study Group. Respiratory Medicine *94*, 954–963.

Finkelstein, J., Cha, E., and Scharf, S.M. (2009). Chronic obstructive pulmonary disease as an independent risk factor for cardiovascular morbidity. International Journal of Chronic Obstructive Pulmonary Disease *4*, 337–349.

Freedman, D.S., Williamson, D.F., Gunter, E.W., and Byers, T. (1995). Relation of Serum Uric Acid to Mortality and Ischemic Heart Disease: The NHANES I Epidemiologic Follow-up Study. Am. J. Epidemiol. *141*, 637–644.

Hidalgo, C.A., Blumm, N., Barabási, A.-L., and Christakis, N.A. (2009). A dynamic network approach for the study of human phenotypes. PLoS Computational Biology *5*, e1000353.

Kelkar, A., Kuo, A., and Frishman, W.H. (2011). Allopurinol as a Cardiovascular Drug. Cardiology in Review *19*, 265–271.

Kohner, E.M. (1989). Diabetic retinopathy. British Medical Bulletin *45*, 148–173.

Moss, S.E., Klein, R., Klein, B.E.K., and Wong, T.Y. (2003). Retinal vascular changes and 20-year incidence of lower extremity amputations in a cohort with diabetes. Archives of Internal Medicine *163*, 2505–2510.

Petri, H., Maldonato, D., and Robinson, N.J. (2010). Data-driven identification of co-morbidities associated with rheumatoid arthritis in a large US health plan claims database. BMC Musculoskeletal Disorders *11*, 247.

Salisbury, A.C., Reid, K.J., and Spertus, J.A. (2007). Impact of chronic obstructive pulmonary disease on post-myocardial infarction outcomes. The American Journal of Cardiology *99*, 636–641.

Schaeffer, S. (2007). Graph clustering by flow simulation. Computer Science Review *1*, 27–64.

Sidney, S., Sorel, M., Quesenberry, C.P., DeLuise, C., Lanes, S., and Eisner, M.D. (2005). COPD and incident cardiovascular disease hospitalizations and mortality: Kaiser Permanente Medical Care Program. Chest *128*, 2068–2075.

Suissa, S., Dell'aniello, S., and Ernst, P. (2012). Long-term natural history of chronic obstructive pulmonary disease: severe exacerbations and mortality. Thorax *67*, 957–963.

Teno, J.M., Weitzen, S., Fenell, M.L., and Mor, V. (2001). Dying trajectory in the last year of life: Does cancer trajectory fit other diseases? Journal of Palliative Medicine *4*, 457–464.

Yang, Q., Köttgen, A., Dehghan, A., Smith, A. V., Glazer, N.L., Chen, H., Chasman, D.I., Aspelund, T., Eiriksdottir, G., Harris, T.B., et al. (2010). Multiple Genetic Loci Influence Serum Urate and Their Relationship with Gout and Cardiovascular Disease Risk Factors. Circ Cardiovasc Genet *3*, 523–530.

Supplementary Materials

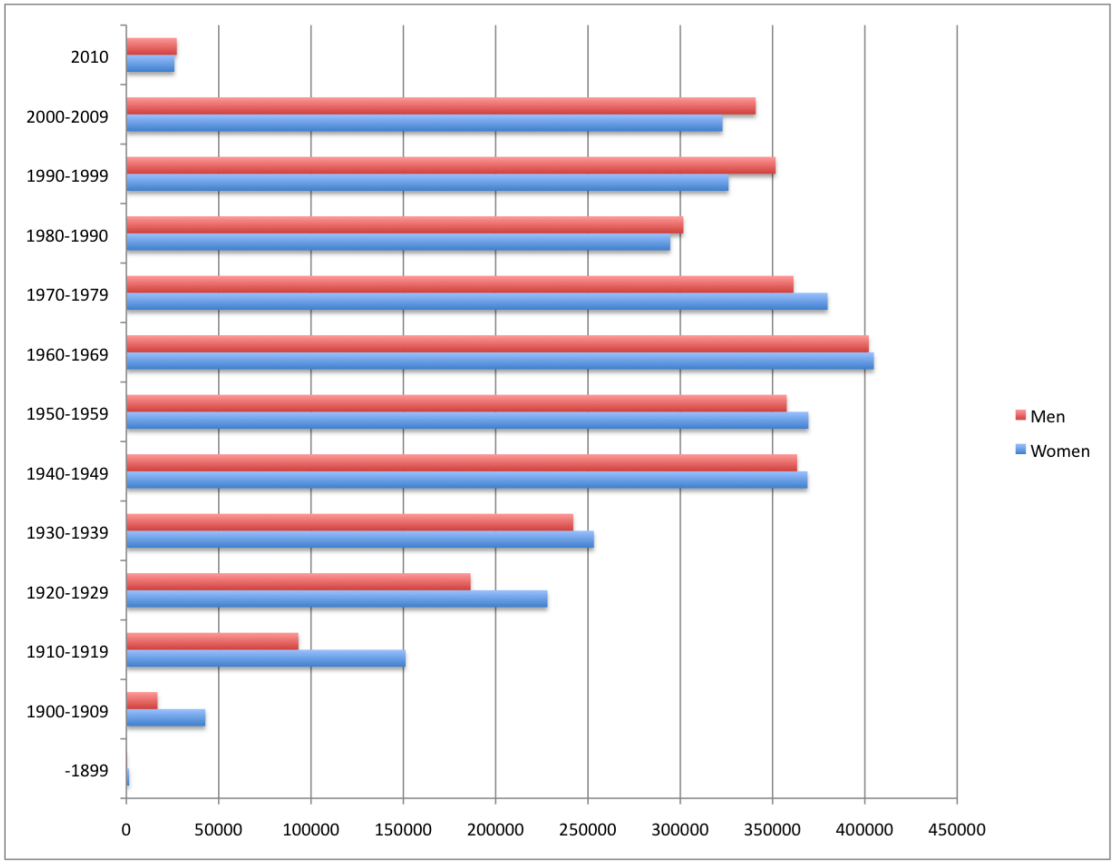
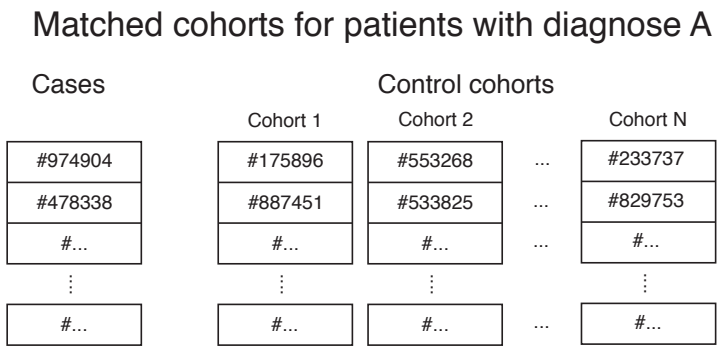
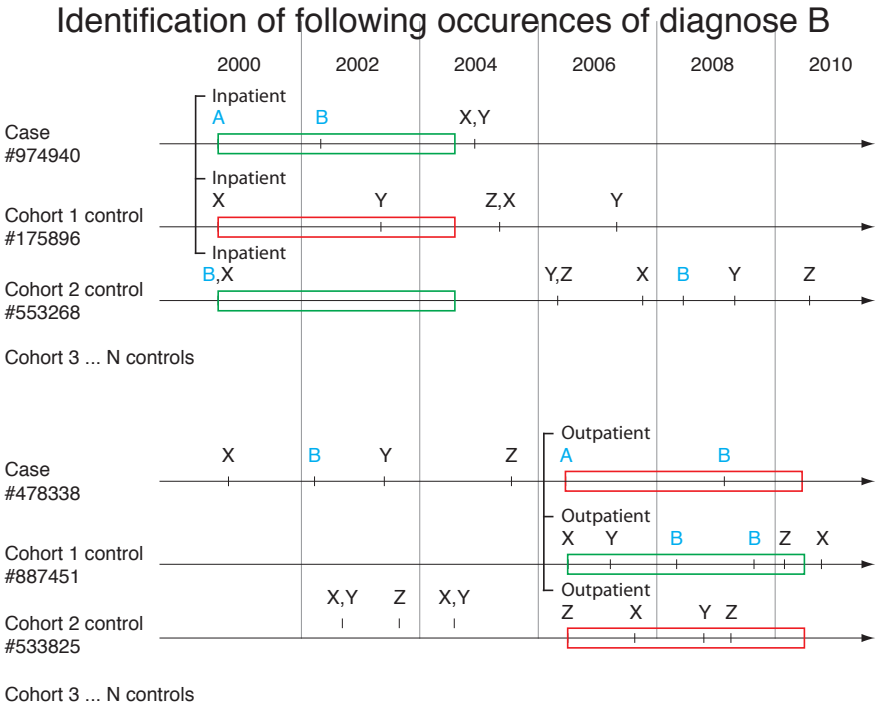


Fig. S1. The NPR population distributed over gender and age (birth decade). The bins shown were used for stratification when matching control patients. For the temporal correlation analysis patients born before 1900 and in 2010 are removed from analyzes using the binning, as the subpopulations are too small to give adequate statistical results. However, the full population is used when counting the trajectories.

Suppl. Fig. S2.A



Suppl. Fig. S2.B



Suppl. Fig. S2.C

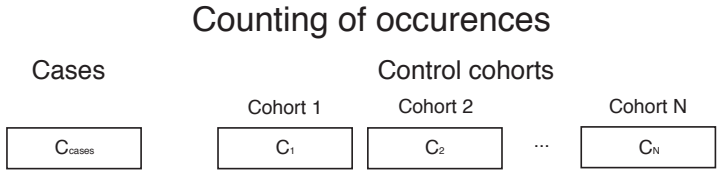


Fig. S2. Illustration of the random sampling procedure with N samplings of the co-morbidity of diagnose A followed by diagnose B within 1 year. A) Each discharge with diagnose A assigned are identified for all patients in order to make a cohort of case patients/discharges. Each case discharge is matched with a set of N randomly chosen control patients. The control patients are chosen among all patients with a discharge the same week as the case having the same gender and age group as the case patient. Each line in A) shows the case and its matched controls. B) The diagnose history of the cases and controls is examined to see if diagnose B occurred within 5 year of the matched week in which the case had diagnose A. X, Y and Z represents arbitrary other diagnoses. C) The number of these occurrences is counted for each cohort giving a number of overlaps. The count for the cases is the observed overlap while the control cohorts are used to estimate the p-values.

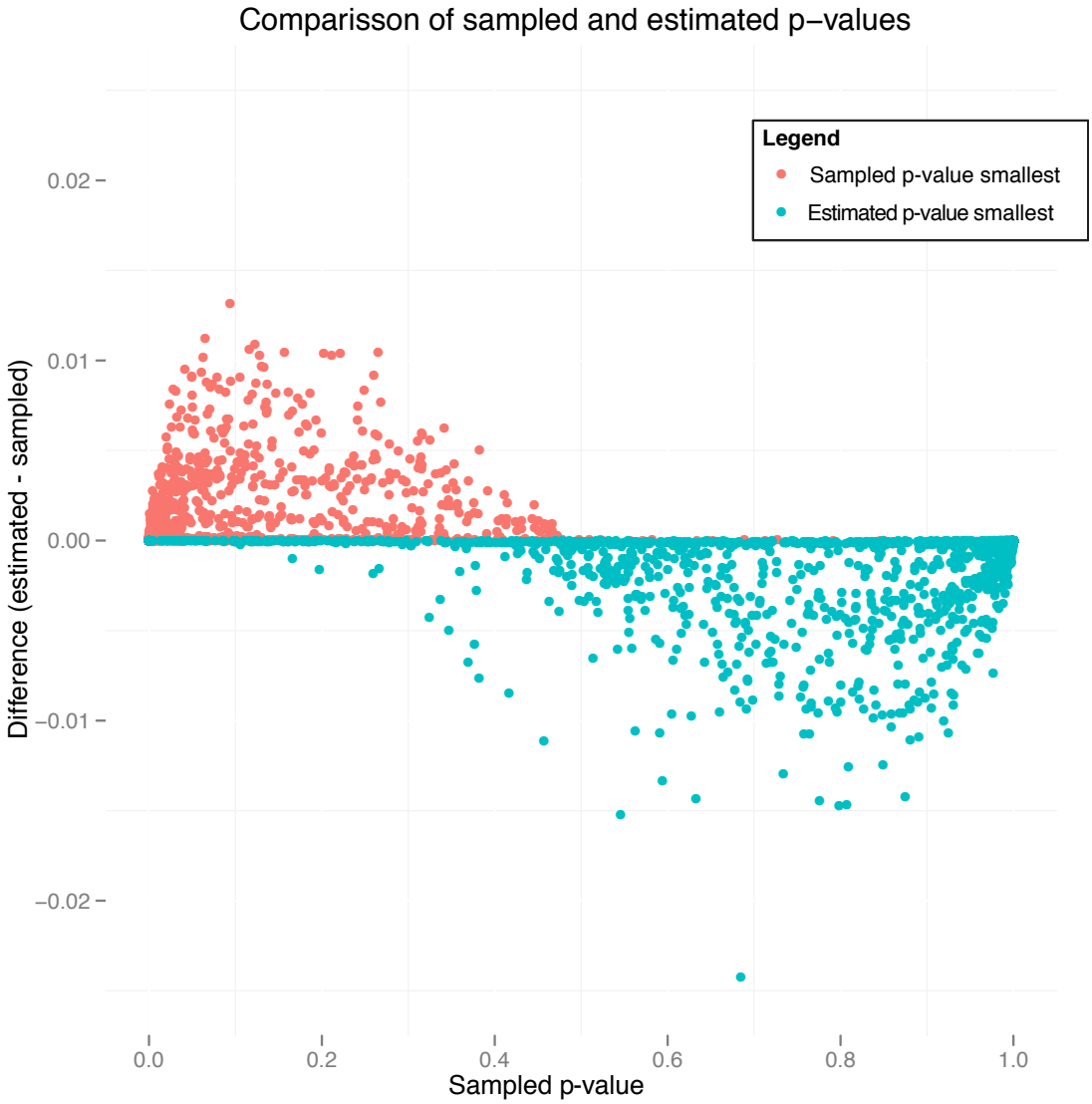


Fig. S3. Benchmarking of p-values estimated with binomial testing. Each point represents the sampled p-value and the difference between estimated and sampled for a pair of diagnoses for some time limit. The estimated p-values are from the model with one population average. Red points are where the sampled model has smallest p-value, implying that the estimated model is more conservative. The fact that the estimated model is more conservative for small p-values reduced the likelihood that using the estimate will cause a false positive. As an extra precaution against false positives we used a p-value of 0.001 when using the estimated p-values.

Stratifications	**Age and gender**	**Not age and gender**
Type of encounter	62,821	72,832
Not type of encounter	72,937	81,058

Table S1. Number of significant pairs in the temporal analysis given different combinations of patient stratification (for type of hospital encounter and for age and gender).

Key diagnose	Have pattern	Have key diagnosis		Controls	RR	p-value
G40 Epelepsy	41,681	2,902	(6.96%)	441.9	6.6	$< 10^{-5}$
H36 Retina disorder	66,758	5,255	(7.87%)	261.9	20.1	$< 10^{-5}$
J44 COPD	156,403	14,460	(9.25%)	2813.4	5.1	$< 10^{-5}$
M10 Gout	105,878	2,470	(2.33%)	363.5	6.8	$< 10^{-5}$

Table S2. Statistics on key diagnoses in trajectory clusters. The table shows counts of how many patients had the pattern (preceding diagnosis followed by succeeding diagnosis), how many of those had the key diagnosis in the period between the preceding and succeeding diagnosis, how many occurrences of the key diagnosis were found on average among the matched controls and RR and p-value for this. 10,000 sampled control cohorts were used to assess RR and p-values.

Table S3 is 245 pages long and was left out.

Table S3. Directional diagnosis pairs. All 4,014 pairs from the temporal analysis that were significant with RR > 1 and had significant directionality are given in the table. Diagnosis codes for the pair (D1 → D2) and statistics are given: Number of patients D1 assigned before D2, RR, associated p-value for the binomial model and the sampled p-value with the number of samples used to obtain it, and the p-value for directionality. The reported p-values are after Bonferroni correction.

Table S4 is 37 pages long and was left out.

Table S4. Trajectory counting. The table reports all 1171 length 4 trajectories along with patient counts, median age at onset of first diagnosis and median duration from onset of first diagnosis to each subsequent step.

4.5 Extending the trajectory model

Our trajectory approach is quite rigid: The requirements for codes to participate are so conservative that only 4,014 out of more than 1 million co-occurring pairs or 62,821 pairs with RR > 1 passed the cut-off. The approach can be made less conservative by using a more relaxed cut-off by using less strict correction for multiple testing or by skipping the testing for directionality. This will give a bigger base to construct more trajectories, but will also allow for more noise in form of disease pairs that have no true temporal correlation with each other. Some initial work with the trajectories indicated that the strict definition gave trajectories that were more interesting than without the requirement for directionality. However, because multiple changes were made in the same revision, the contrast between with and without this requirement was never fully made.

The definition that each step is exactly one ICD-10 code puts a limit to how much the trajectory can cover. A consequence of this is that many trajectories are variants of each other with one code changed. Using clustering we were able to group these variants and illustrate their structure in one figure. However, patients can get diagnoses from different branches after a given step and there is no fixed starting point. Thus, patients can be in multiple diagnosis steps in the trajectory cluster at a given time. Without constraints on branching and a fixed starting point, it is impossible to unambiguously count how many patients followed a given transitions in the bundled trajectory. The HMM models discussed in section 4.3 can be used to resolve the counting issue. The information about selection of disease and transitions between diseases that are required to make HMMs are given in the descriptive trajectory counting. Each of the trajectory clusters can be the basis for modeling a HMM. The transitions between states can be limited to follow statistical significant paths like the trajectories. Figure 4.1 illustrates an example of how a HMM can be designed given the trajectory clusters. A method for obtaining more trajectories to base the HMMs on is to use the three-long trajectories rather than the four-long. Furthermore, patients can be clustered based on which three-long trajectories they share with other patients instead of clustering the trajectories based on diagnoses. In this way, a patient data point will be the set of trajectories he or she follows. A cluster in this space is a group of patient following similar trajectories. HMMs can be used to make consensus trajectory for each group of patients.

4.5.1 Further modeling time

Although time is modeled in the correlations and by the order of the diseases, time is not a direct variable in the trajectories. The time that passes between two diagnoses can contain crucial information about the outcome of the treatment. A short time can be indicative of fast progressions towards worse condition for the patient. Alternatively, it can be an indication of fast diagnosing and a fast treatment leading to better condition. Further development of the trajectories should model the time between diagnoses directly.

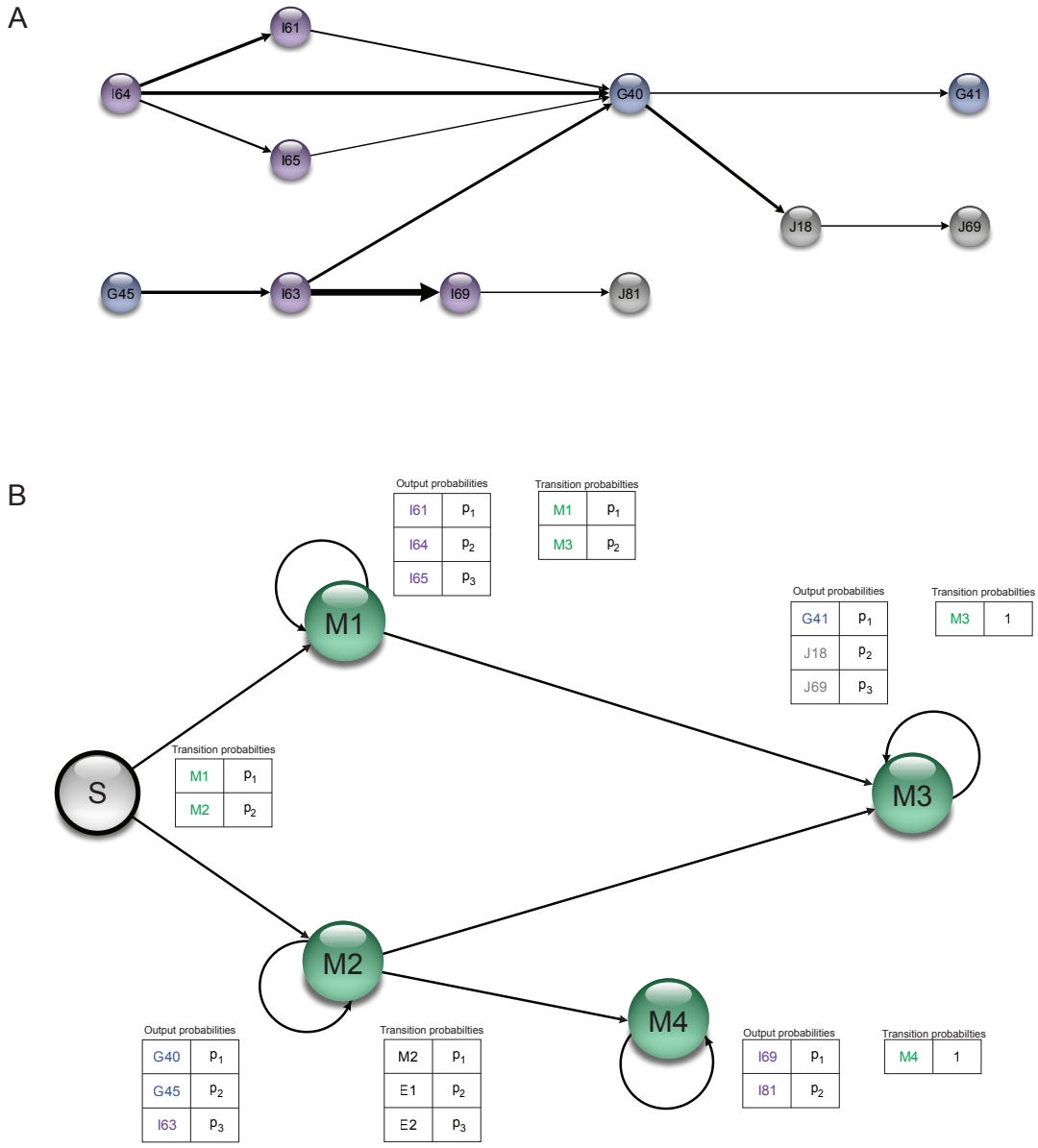


Figure 4.1 Illustration of one HMM for the cerebrovascular trajectory group (from manuscript I). (A) The trajectory cluster. (B) One way of converting the trajectory into a HMM. S is the starting state and the M1-M4 are states reflecting different disease progression. The transition probabilities and output probabilities have to be estimated based on the data. The sketched model contains only diagnoses from the trajectory in the sequence for simplicity of the figure. An implementation of the model should contain probabilities of all diseases. The presence of other diseases is likely to assist in defining the states.

With the current data, the effect of time can already be observed. The time of a trajectory can be reported in absolute time from first diagnosis or delta time between diagnoses. When using absolute times some trajectories reveal patients that discontinue the trajectory have longer time to their last diagnosis step than the patients who "finishes" the full trajectory. This might be an artifact of censoring: patients with long time to a diagnosis step have less time covered by the data to develop the next disease. However, it might also reveal important details about the trajectories.

The simplest method to model the time between the diseases in the descriptive model is to use a waiting state between disease steps, and to bin the time into large bins like within 30 days, 1 year, more than one year. A problem with this is that this will increase the number of trajectories, and thus reduce the number of patients following each branch of a trajectory. Therefore, only cases where time has been shown to be important should be included in the current model. If HMMs are used to model the trajectories, the waiting time can be an output symbol. Each state can be associated with some waiting times or every second state could be waiting time to next disease occurrence. Alternatively, the time can be modeled using an exponential distributions for each step of the HMM. Since statistical inference on the HMMs is based on multiplying probabilities for a sequence to obtain a likelihood, it should be possible to integrate this with the waiting time probabilities.

4.5.2 Including data on treatment and symptoms

The current trajectories do not include treatment and surgical procedures. These data are available in NPR. While the current counting of disease trajectories is descriptive and only models outcome, adding the treatment adds a variable that doctors can actively affect. If the outcome of different kinds of treatment is successfully modeled, this can be of great importance to decision support. Treatments can be a step in the disease trajectories or an output symbol in a HMM.

The classification of treatments and surgical procedures does not contain information on which treatments or procedures are used for a given disease. The association between treatment and diseases can be inferred from the data, possibly using a similar approach as the disease correlations.

Codes from the symptom chapter of ICD-10 were left out from the correlation analysis. However, they can be used to describe the disease state. It is possible that some symptoms can be used to assess the severity of the current state and help predict future outcome.

4.5.3 Including genetic information

An aim of the trajectory analysis was also to investigate the effect of genetics on disease development. The genetic information would be input for a predictive model, so future outcome would be predicted based on the patient's medical history and genome. As discussed in chapter 6, it is possible to infer some associations between diagnosis codes and genes. However, to properly test if this can be used to predict future outcome, we need the genetic data on patients.

Chapter 5:
Colectomy & the gut bacteria

It has long been known that the humane cells are outnumbered by 1 to 10 by microbial cells, which influence physiological processes in their host [53]. The bacteria residing in the gut tract are collectively called gut bacterium. The different gut bacteria can have a malignant, benign or beneficial effect on diseases depending on the personal gut genome. Among the hypothesized effects are influences on obesity [54], diabetes [55] and cardiovascular diseases [56].

The direct way of assessing the effect of specific gut bacteria on diseases requires the gut bacteria genome to be sequenced. However, a simple way of determining the total effect is to study patients who undergo a total colectomy where their entire colon is removed. Since this operation removes all the colonic bacteria, neither the malignant nor the beneficial effects will be present in the patient after the surgery. Thus, analyzing the increased or decreased risk of diseases following colectomy is a way of gaining insight about the effect of the gut bacterium. In the work in progress II the hypothesis that removing the gut bacterium through total colectomy reduces the risk of cardiovascular diseases is tested.

5.1 Survival analysis

In manuscript I, temporal correlations were assessed using RR for a single time interval where occurrence of the disease was counted. Another approach for quantifying temporal correlations in a time dependent fashion is using methods from survival analysis. In survival analysis, the time it takes for some event to occur is analyzed. Any type of event can be studied, for example, occurrence of co-morbidities in patients. The main goal is to estimate the number of occurrences over time and to compare the outcome among different groups of subjects. The hazard function measures the percentage of survivors (subjects for whom the event has not occurred) at a given time point. It can be estimated with the non-parametric Kaplan-Meyer estimator and two or more groups can be compared with parametric estimators such as proportional hazards models. The latter results in odds-ratio that similarly to RR quantifies the difference between the groups.

5.1.1 Kaplan-Meyer estimator and proportional hazards models

Incomplete data due to censoring is a main issue that is addressed in survival analysis. A subject is censored if the event has not taken place by the end of the period covered by the data for the subject. The most common type of censoring is right-censoring, which applies to the colectomy study. This happens when patients are still alive at the end of the data period or are lost to follow-up (e.g. if the patient moves out of the country). In the case of the colectomy study, this happens as the data covers a fixed period of 14.9 years. Patients having their colectomy in the start of the period have many years to

develop co-morbidities, while patients in the end of the period have only few years or none at all.

The issue was addressed in 1958 where Kaplan and Meier presented their estimator where the number of events among subjects at risk is counted [57]. Subjects contribute as positive survivors as long as the event has not occurred and they are covered by the data. Patient alive at the endpoint of their data are not counted as being at risk here after, and they no longer contribute to the percentage of survivors.

Special statistical tests for comparing different groups are needed due to the censoring issue [58]. The proportional hazards models can be used for this purpose. They are parametric models that assume that the effect of some explanatory variables on the hazard function can be explained by a base hazard function multiplied with some parameter φ. The base hazard function corresponds to the neutral hazard, and the parameter φ is a function of the explanatory variables, which can be different groups (such as case-control, man-woman). A partial log-likelihood method due to [59] can be used to estimate the constants for the φ function. The value of the φ function for a group can be interpreted as the log-odds for the group.

5.2 Work in progress II: Reduced risk of cardio-vascular disease following total colectomy

This work has been done in collaboration with Teresa A. Ajslev, Thorkild I. A. Sørensen. The introduction, method and the main part of discussion are part of a manuscript draft that was written by Teresa (and corrected by Thorkild, Søren Brunak and me).

5.2.1 Introduction

The gut bacteria may play a role in atherosclerotic cardio-vascular disease (CVD) [60]. The host-microbial symbiosis has developed through primate evolution and may benefit both bacteria and the human host [61–64]. However, the intestinal bacterial symbiosis and the intestinal barrier function can be disrupted. Overgrowth of 'harmful' bacterial strains may take place, which through nutritional fermentation may influence CVD risk [60,61,64,65]. In addition, a disrupted intestinal barrier, may lead to increased diffusion of lipopolysaccharide to the blood stream, which may induce systemic inflammation through endotoxemia, possibly contributing to the development of atherosclerotic diseases [61,66]. Lower gut bacterial richness has been associated with obesity as well as with an altered metabolic profile [67]. In view of the high frequency of CVD in the Western world, it may be that a disrupted gut microbiome is prevailing in this population. Based on these considerations, our hypothesis was that patients from a Western population, who have had a total colectomy, will have a reduced long-term risk of CVD. Using access to a nation-wide patient register in Denmark, we compared the long-term occurrence of CVD in patients who have had a total colectomy with all other patients as well as with four groups of patients who have undergone other types of surgery.

5.2.2 Methods

For this study, we used the nation-wide Danish National Patient Registry (NDR) covering all hospital contacts for patients from year 1996 through 2010 (14.9 years). Total colectomy is coded as JFH in the Scandinavian NOMESCO Classification of Surgical Procedures, and covers total removal of the large intestine with or without ileostomy as well as with or without ileo-rectal pouch. Patients who had undergone this procedure were compared with five groups of patients from the register. Group 1 consists of inpatients discharges from the entire NDR population. The four other groups included patients who had undergone various types of surgery; Group 2 - appendectomy, Group 3 - any orthopedic surgery, Group 4 - any abdominal surgery leaving the colon intact, and Group 5 - major surgeries, unrelated to the gastrointestinal tract or the cardio-vascular system (supplementary table 5.5). For each colectomy patient, five patients within each comparison group were matched by gender, year of birth (within same year), year and month of surgery (+/-24 months) and otherwise randomly selected (in a few cases for Group 2, the year of birth was we extended to +/- 1 year, because of too few matching patients). The cause for colectomy may influence the subsequent risk of CVD and therefore the investigation was supplemented by analyses stratified in two subgroups. One consisted of colectomy patients having either co-occurrence of a cancer diagnosis: any C-diagnoses in ICD-10; the other consisted colectomy patients with inflammatory

bowel diseases (IBD): Crohn's Disease (K50), Ulcerative Colitis (K51), other non-infective gastroenteritis and colitis (K52) and irritable bowel syndrome (K58) at the time of surgery.

The time period at risk of CVD

Since the focus of the study is the long-term CVD following colectomy, we restricted the analyses to occurrence of CVD from the age of 45 years. To avoid the possible influence of morbidity and mortality related to the surgery as such, the time period at risk began 1,000 days after colectomy and the corresponding time points in the comparison groups. The matched patients in the comparison group were required to be alive 1000 days after their surgery or discharge, so that all included colectomy patients have five controls alive after 1000. The choice of 1,000 days was based on the cumulative hazard plots for the colectomy group and the five comparison groups; from that point in time, the slopes of the cumulative hazards were the same in all groups indicating comparable risks of fatal morbidity and mortality over time (Supplementary Figure S1).

CVD events

Seven groups of diagnoses, likely encompassing CVD, were selected from the International Classification of Disease (ICD-10) covering hypertensive diseases (I10–13), acute ischemic heart diseases (I20, I21, I24, I46), chronic ischemic heart disease (I23, I25, I51), other heart diseases (I44–46, I48–51), cerebrovascular diseases (I61-I69), other arterial diseases (I35, I70-I74) and heart failure (I50). A composite end-point combining all the mentioned diseases was also included. The first occurrence of any of the diagnosis from a group or the composite endpoint was used as the occurrence date. Supplementary table 5.6 shows the names of ICD-10 diagnoses as well as the number of observed events in the colectomy patients and in the subgroups with co-occurring cancer and irritable or inflammatory bowel diseases.

Statistical analyses

The risk of CVD in the colectomy patients were compared to the risk in the other groups of patients by hazard ratios with p-values estimated by analyses based on the Cox' proportional hazard models. Gender and age were included as covariates in the model as well. Each of the five comparison groups was compared separately with the groups of colectomy patients, the total group and the subgroups with the defined co-occurring diagnoses at time of surgery. The underlying time variable of the models was the number of days from the 1,000 days after colectomy. All CVD diagnoses were individually analyzed, irrespective of the occurrence of the other CVD diagnoses. Patients were censored from the analyses either at time of death, but not at the occurrence of other CVD diagnoses. Level of significance set to 5% in all analyses adjusted according to Bonferoni correction (giving an effective level of 0.11%).

5.2.3 Results

Colectomy was performed in 4,057 patients age 45 and above from year 1996 to 2010 with a median age of 67 years. A high post-surgical death rate was observed, with 18.6%

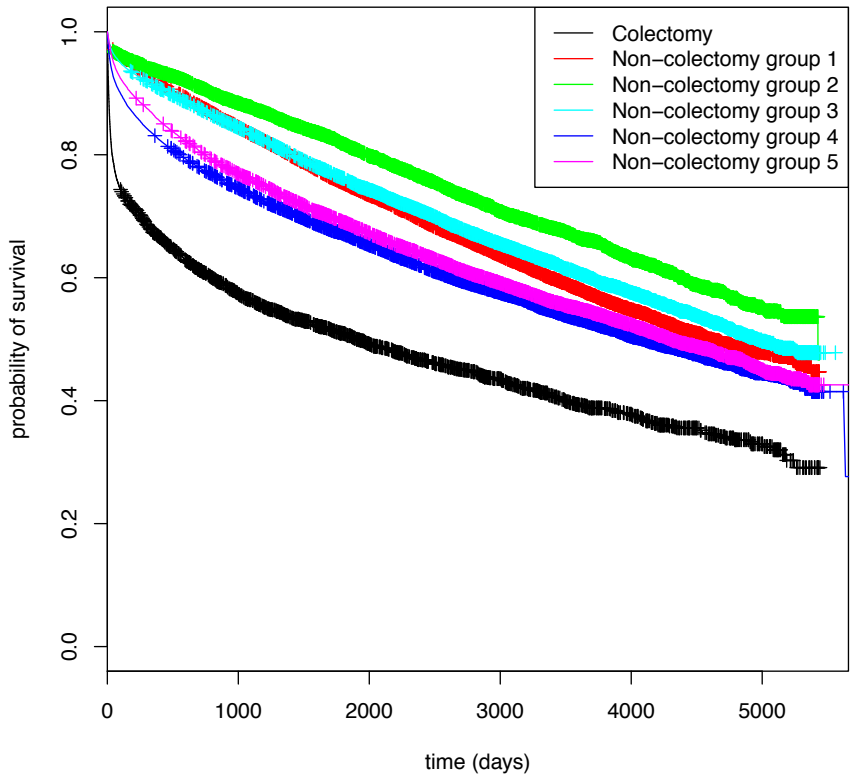


Figure 5.1 Kaplan-Meyer survival plot of the colectomy patients and all non-colectomy groups. The plot demonstrates the high mortality in the first 1,000 days following colectomy surgery.

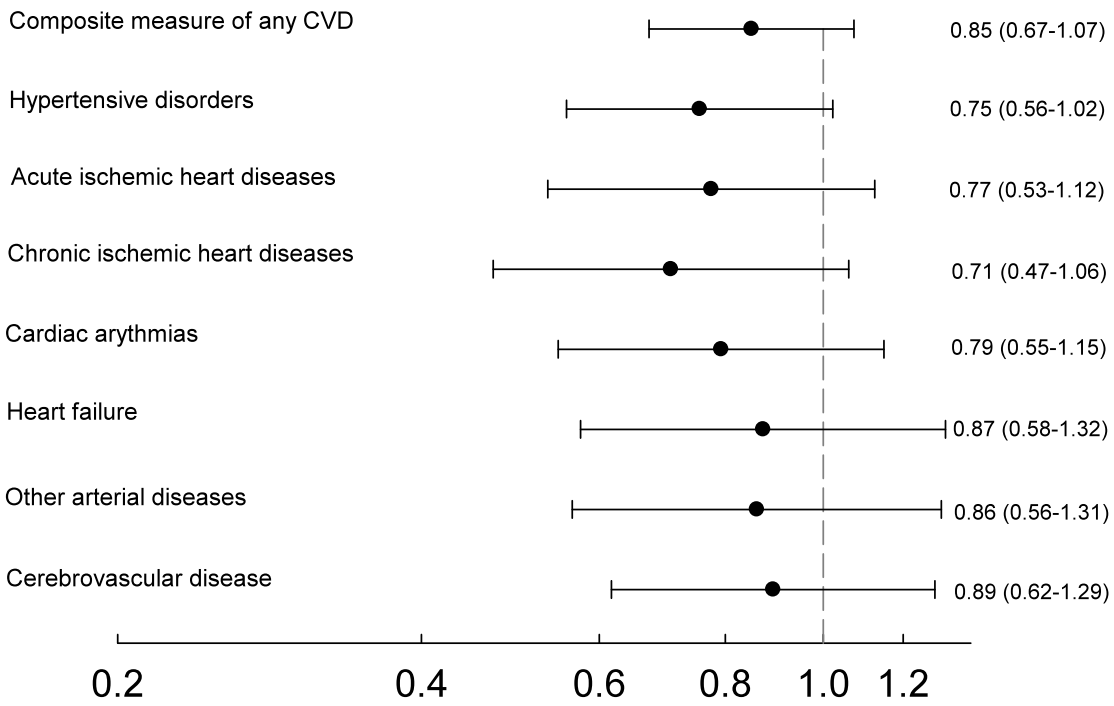


Figure 5.2 The hazard ratios and confidence intervals for the non-colectomy group 1 (all patients).

of colectomy patients deceased within the first 30 days and 41% (n=1,662) deceased within 1000 days. (figure 5.1). Thus, the group studied at risk of CVD included the 1,946 colectomy patients alive 1000 days after surgery and five non-colectomy groups with 9,730 individuals in each. Table 5.1 shows the number of CVD events in the colectomy and non-colectomy groups. The hazard ratio are shown with confidence intervals for the non-colectomy group 1 in figure 5.2, and for all groups and stratifications in figure 5.3.

After correction for multiple testing, the risk of hypertensive disorders was significantly reduced in the colectomy patients compared to group 3, any orthopedic surgery, (p=0.0010, hazard ratio 0.74, 99.9% CI: 0.55-0.996), but not significant compared to the non-colectomy groups. We also saw a trend towards reduced risk for acute ischemic heart disease (0.53, CI: 0.77-1.12 for all patients) and chronic ischemic heart disease (0.47, CI: 0.71-1.06 for all patients) compared to the non-colectomy, but these results were not significant. We investigated the risk of developing heart failure, cerebrovascular disease and other arterial diseases and found no significant changes in hazard ratios (p = 0.28, 0.31, 0.24 respectively compared to all patients). Furthermore, we grouped the

All patients	**Colectomy**		**Group 1**		**Group 2**	
Diagnosis group	Non-censored pats.	Pats. w/ diagnosis group	Non-censored pats.	Pats. w/ diagnosis group	Non-censored pats.	Pats. w/ diagnosis group
Composite	1159	231	6408	1240	7352	1357
Hypertensive disorders	1464	133	8967	876	9199	905
Acute ischemic heart diseases	1621	89	9358	546	9627	569
Chronic ischemic heart disease	1643	74	9784	515	9896	449
Other arterial diseases	1666	70	10326	423	10245	348
Cardiac arrhymias	1614	89	9847	574	9932	510
Heart failure	1724	72	10455	419	10461	397
Cerebrovascular diseases	1680	92	10044	520	10180	439

All patients	**Group 3**		**Group 4**		**Group 5**	
Diagnosis group	Non-censored pats.	Pats. w/ diagnosis group	Non-censored pats.	Pats. w/ diagnosis group	Non-censored pats.	Pats. w/ diagnosis group
Composite	7741	1493	7424	1458	7434	1378
Hypertensive disorders	9531	966	9255	921	9259	933
Acute ischemic heart diseases	10001	576	9883	591	9856	526
Chronic ischemic heart disease	10300	466	10155	451	10138	449
Other arterial diseases	10405	425	10427	394	10442	407
Cardiac arrhymias	10301	590	10138	553	10141	535
Heart failure	10715	403	10661	362	10647	375
Cerebrovascular diseases	10399	475	10438	472	10482	478

Table 5.1 All patient diagnosis count. Number of non-censored colectomy and non-colectomy patients for each diagnosis group and the number out of these who was diagnosed with one disease from the group.

	Cancer	Non-cancer
IBD	53	847
Non-IBD	672	374

Table 5.2 Number of colectomy patients stratified according to presence and non-presence of cancer and IBD.

colectomy patients based on the likely cause of colectomy into a cancer and inflammatory bowel disease group (table 5.2). Table 5.3 and 5.4 shows the diagnosis counts for the two stratifications. We were not able to find any significant risk changes in any of the groups.

5.2.4 Discussion

This study investigated long term risk of CVD following total colectomy across a panel of diagnoses likely related to atherosclerosis. The postsurgical mortality rate after colectomy is high. By limiting the risk period to 1000 days following surgery and onwards the over-all mortality was similar in the colectomy group and the non-colectomy groups. We found patients with colectomy who survived 1000 days from the surgery had a significantly decreased risk of hypertensive disorders and a trend towards reduced risk of acute and chronic ischemic heart diseases.

In the literature there is increasing evidence suggesting that the gut bacteria – whether through a "harmful" bacterial composition or by a disrupted intestinal barrier – is involved in CVD such as atherosclerosis. The direction of the causal relationship is however still unknown [66,68,69]. One theory is that bacterial fermentation in the gut leads to higher levels of trimethylamine-N-oxide (TMAO) which has been suggested to promote atherogenesis [60,62,65,70].

Another theory is that increased lipopolysaccharide (LPS) diffusion to the circulation through a disrupted intestinal barrier leads to metabolic endotoxemia [61,62,71]. For all of the investigated CVD diagnoses, a low-grade inflammatory origin of disease is possible, with previously observed elevated inflammatory markers among diseased individuals [66,69,72]. We encourage others to perform studies investigating the systemic metabolic and inflammatory states among the colectomy patients compared to relevant controls, before the pathway through the intestine can be confirmed.

A potential limitation is the possibility of inaccuracy in CVD diagnosis and classifications, which may have introduced bias. However, since the specific diagnoses were grouped with similar diagnoses, we do not suspect that systematic or random misclassification should have influenced results. On the other hand, depending on severity of the underlying disease there may be differences in number of hospital contacts and thus in the possibility of obtaining a CVD diagnosis. To take care of such possible bias, five different comparison groups covering patients from different disease severity groups were generated. In addition, bias through confounding by indication - meaning that having surgery performed depends on disease severity and presence of co-morbidities before surgery - should also have been revealed through the use of several comparison groups representing different severity of diseases.

Patients were censored if they had prior history of the disease group. For the composite endpoint, 1,236 of 2,395 patients were censored. For the other groups between 671 and 931 patients were censored. It is likely that having CVD from one group gives increased risk of getting a CVD from another group. This bias could be addressed by censoring the 12,36 patients with prior history of the composite group.

Another possible bias could be due to lifestyle changes following surgery, such as smoking cessation. The initiation to such changes may be higher in colectomy patients having inflammatory bowel disease than their comparison groups because smoking may relieve the symptoms in these patients. We cannot assess this properly by information from the registers.

The stratification into cancer and IBD patients showed no significant results. A possibility to improve the stratification is to narrow the cancer diagnoses to colorectal cancer. Likewise, the IBD patients could be limited to Crohns' disease and ulcerative colitis. To compare the two groups, they also need to be matched for size and age and gender composition.

Cancer patients	Colectomy		Group 1		Group 2	
Diagnosis group	Non-censored patients	Patients with diagnosis group	Non-censored patients	Patients with diagnosis group	Non-censored patients	Patients with diagnosis group
Composite	380	80	2189	435	2504	491
Hypertensive disorders	512	46	3255	313	3341	326
Acute ischemic heart diseases	576	25	3473	194	3565	200
Chronic ischemic heart disease	587	25	3626	204	3617	163
Other arterial diseases	597	23	3851	161	3787	154
Cardiac arrhymias	561	33	3635	232	3654	199
Heart failure	625	21	3896	158	3897	163
Cerebrovascular diseases	596	37	3717	193	3749	182

Cancer patients	Group 3		Group 4		Group 5	
Diagnosis group	Non-censored patients	Patients with diagnosis group	Non-censored patients	Patients with diagnosis group	Non-censored patients	Patients with diagnosis group
Composite	2662	520	2569	543	2553	486
Hypertensive disorders	3453	352	3377	336	3360	338
Acute ischemic heart diseases	3733	225	3662	215	3629	181
Chronic ischemic heart disease	3814	174	3773	181	3739	165
Other arterial diseases	3856	162	3910	165	3874	146
Cardiac arrhymias	3811	224	3716	204	3767	229
Heart failure	4009	164	3987	157	4002	164
Cerebrovascular diseases	3882	189	3890	197	3918	192

Table 5.3 Cancer patient diagnosis count. Number of non-censored colectomy and non-colectomy patients for each diagnosis group and the number out of these who was diagnosed with one disease from the group.

IBD patietnts	Colectomy		Group 1		Group 2	
Diagnosis group	Non-censored patients	Patients with diagnosis group	Non-censored patients	Patients with diagnosis group	Non-censored patients	Patients with diagnosis group
Composite	615	115	3154	614	3636	629
Hypertensive disorders	732	60	4229	434	4364	431
Acute ischemic heart diseases	785	44	4348	275	4472	284
Chronic ischemic heart disease	781	36	4539	238	4656	219
Other arterial diseases	810	33	4773	187	4774	149
Cardiac arrhymias	791	40	4593	267	4645	216
Heart failure	819	34	4831	200	4857	172
Cerebrovascular diseases	806	37	4663	232	4760	190

IBD patietnts	Group 3		Group 4		Group 5	
Diagnosis group	Non-censored patients	Patients with diagnosis group	Non-censored patients	Patients with diagnosis group	Non-censored patients	Patients with diagnosis group
Composite	3802	714	3651	673	3621	671
Hypertensive disorders	4487	456	4353	415	4365	454
Acute ischemic heart diseases	4613	255	4567	278	4584	259
Chronic ischemic heart disease	4775	209	4705	201	4706	208
Other arterial diseases	4827	184	4809	173	4836	198
Cardiac arrhymias	4802	257	4740	264	4702	231
Heart failure	4947	171	4924	143	4889	160
Cerebrovascular diseases	4820	211	4831	200	4830	203

Table 5.4 Cancer patient diagnosis count. Number of non-censored colectomy and non-colectomy patients for each diagnosis group and the number out of these who was diagnosed with one disease from the group.

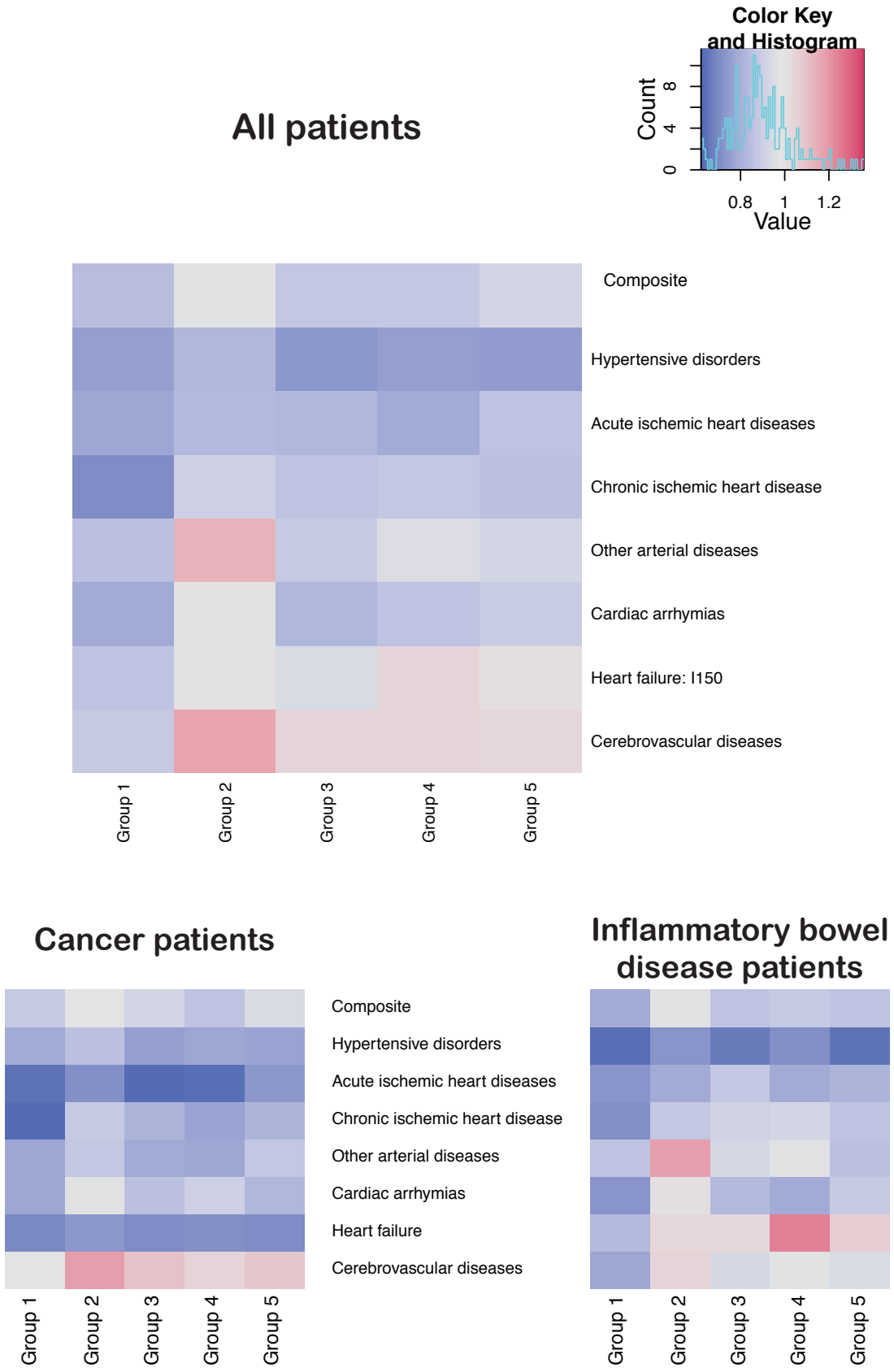


Figure 5.3 The hazard ratios for all non-colectomy groups shown as a heatmap for the full group of patients and the cancer and IBD stratifications. Blue colors correspond to a reduced risk, while red colors correspond to an increased risk. The significance is not illustrated.

5.2.5 Supplementary material

	Non-colectomy groups	NCSP codes	
1	All hospital in-patients		Hospital admission, not necessarily surgery
2	Appendectomy	JEA (combined with ICD-10 K35.3)	
3	Orthopaedic surgery	N	Musculoskeletal system
4	Surgeries in the gastrointestinal tract leaving the intestine intact	JFA	Local operation intestine
		JK	Biliary tract
		JA	Abdominal wall, mesentery peritoneum, and greater omentum
		JB	Diaphragm and gastro-oesophageal reflux
		JC	Oesophagus
		JD	Stomach and duodenum
		JF	Intestine
		JG	Rectum
		JH	Anus and perianal tissue
		JFB0	Small bowel resection
		JFB1	Reversal of small bowel segment
		JFB20	Ileocaecal resection
		JFB21	Laparoscopic ileocaecal resection
5	Large surgeries not in the gastrointestinal-tract	BA	Thyroid glands
		BB	Parathyroid glands
		BC	Adrenal glands
		G	Chest wall, pleura, mediastinum, diaphragm, trachea, bronchus and lung
		JM	Spleen
		K	Urinary system, male genital organs and retroperitoneal space
		L	Female genital organs

Suplementary table 5.5 Specification of non-colectomy patient groups.

Diagnose group	ICD code	
Hypertensive disorders	I10	Essential (primary) hypertension
	I11	Hypertensive heart disease
	I12	Hypertensive renal disease
	I13	Hypertensive heart and renal disease
Acute ischemic heart diseases	I20	Angina pectoris
	I21	Acute myocardial infarction
	I24	Other acute ischaemic heart diseases
	I46	Cardiac arrest
Chronic ischemic heart disease	I23	Certain current complications following acute myocardial infarction
	I25	Chronic ischaemic heart disease
	I51	Complications and ill-defined descriptions of heart disease
Cardiac arrhymias	I44	Atrioventricular and left bundle-branch block
	I45	Other conduction disorders
	I47	Paroxysmal tachycardia
	I48	Atrial fibrillation and flutter
	I49	Other cardiac arrhythmias
Heart failure	I50	Heart failure
Cerebrovascular diseases	I61	Intracerebral haemorrhage
	I62	Other nontraumatic intracranial haemorrhage
	I63	Cerebral infarction
	I64	Stroke, not specified as haemorrhage or infarction
	I65	Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction
	I66	Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction
	I67	Other cerebrovascular diseases
	I68	Cerebrovascular disorders in diseases classified elsewhere
	I69	Sequelae of cerebrovascular disease
Other arterial diseases	I35	Nonrheumatic aortic valve disorders
	I70	Atherosclerosis
	I71	Aortic aneurysm and dissection
	I72	Other aneurysm
	I73	Other peripheral vascular diseases
	I74	Arterial embolism and thrombosis

Suplementary table 5.5 Specification diagnosis groups.

Chapter 6:
Digging for information on genetic diseases in health data

Interactions between proteins and links between diseases and proteins have extensively been characterized the last decades [4]. Protein-protein interactions (PPI) can be used to identify links between diseases at the level of functional protein modules, networks or pathways [13,73], and there are various papers and databases that gather these interactions such as STRING [74], "INWEB" [13,75] and BioGRID [76]. This can be supplemented with data from disease registries and medical patient records, which constitutes a valuable piece of information for identifying diseases that correlate through molecular pathways [4,5,7,77]. To elucidate if there is a molecular link between the diseases that correlate, we need to know what genes have been related to the diseases. We can among others use the information gathered in databases such as OMIM [78] or Uniprot [79] for this purpose. Given a curated set of disease-gene links and a reliable protein network, the genetic etiology of diseases, as well as the molecular link between correlated diseases can be studied.

6.1 Linking genes and diseases

The Online Mendelian Inheritance in Man, OMIM, is a resource for linking Mendelian and complex diseases to genes [78]. Each genetic disease and gene listed in it has a structured text that describes the gene or disease with references to the literature. Text mining can be applied to obtain structured data out of the text records [13]. OMIM can be used to obtain curated link between diseases and disease genes [5]. Given a pair of highly correlated diseases, we can study the shared disease proteins and the protein networks that connect these proteins to uncover the molecular link between diseases.

The ICD-10 diagnosis codes in the registry data have to be mapped to OMIM disease identifiers. In doing this, one has to take into account that the diseases in OMIM become highly subdivided and that the database has no hierarchical structure, which complicates the mapping. It should be considered that ICD-10 is disease centric while OMIM is gene centric. There is no direct mapping between ICD-10 and OMIM publicly available currently, but other disease centric classifications like Disease Ontology are easier to map to OMIM. Other resources such as UniProt, which is a protein database that provides curated information that can be used to obtain disease-gene links. Unlike OMIM, Uniprot is more disease oriented, which would make the mapping between ICD10 and Uniprot disease-protein links much easier, faster and reliable.

The main challenge in using registry data and health records is the lack of genetic data of the patients. Without this, it is unknown which kind of genetic disposition the patient has for the diseases. The Mendelian diseases are mainly associated to a specific gene mutation. On the contrary, the complex diseases caused by mutations in multiple genes and other factors, and even summing the effects of all these mutations and factors, we often can explain very little of that disease. Another challenge is to remove the effect of

other confounding factors. With correction for only age and gender, many important factors like drugs and lifestyle are contributing to the correlation signal. It is naïve to think that the correlations identified in the registry data are mainly due to genetics.

6.2 Mendelian code

An approach to uncover the underlying genetic etiology of complex diseases is presented in manuscript II. Some rare Mendelian diseases can predispose for common complex diseases. In the manuscript the example of Huntington's disease predisposing for type II diabetes is cited. A possible explanation is that type 2 diabetes and Huntington's disease in part share genetic causes. The same phenomenon may be seen for other pairs of Mendelian and complex diseases. If this is the case, such pairs can be identified by strong correlation between the two diseases. The study presented in the manuscript II tests this hypothesis using registry data.

6.3 Manuscript II: A Non-Degenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk

David R. Blair, Christopher S. Lyttle, Jonathan M. Mortensen, Charles F. Bearden, Anders Boeck Jensen, Hossein Khiabanian, Rachel Melamed, Raul Rabadan, Elmer V. Bernstam, Søren Brunak, Lars Juhl Jensen, Dan Nicolae, Nigam H. Shah, Robert L. Grossman, Nancy J. Cox, Kevin P. White, and Andrey Rzhetsky

Published in Cell 2013, 155:1 p. 70-80

This is a large-scale cross-nation study, where my contribution to the work was to run the method on the NPR data and provide the results.

Cell

A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk

David R. Blair,[1] Christopher S. Lyttle,[2] Jonathan M. Mortensen,[7] Charles F. Bearden,[8] Anders Boeck Jensen,[9] Hossein Khiabanian,[10] Rachel Melamed,[10] Raul Rabadan,[10] Elmer V. Bernstam,[8] Søren Brunak,[9,11] Lars Juhl Jensen,[9,11] Dan Nicolae,[3,4,5] Nigam H. Shah,[7] Robert L. Grossman,[4,6] Nancy J. Cox,[4,5] Kevin P. White,[4,5,6,*] and Andrey Rzhetsky[4,5,6,*]
[1]Committee on Genetics, Genomics, and Systems Biology
[2]The Center for Health and the Social Sciences
[3]Department of Statistics
[4]Department of Medicine
[5]Department of Human Genetics
[6]Computation Institute, Institute for Genomics and Systems Biology
University of Chicago, Chicago, IL 60637, USA
[7]Stanford Center for Biomedical Informatics Research, Stanford, CA 94305, USA
[8]School of Biomedical Informatics, Department of Internal Medicine, the University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[9]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Copenhagen, Denmark
[10]Department of Biomedical Informatics, Center for Computational Biology and Bioinformatics, Columbia University, New York, NY, 10032, USA
[11]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen, Denmark
*Correspondence: kpwhite@uchicago.edu (K.P.W.), arzhetsky@uchicago.edu (A.R.)
http://dx.doi.org/10.1016/j.cell.2013.08.030

SUMMARY

Although countless highly penetrant variants have been associated with Mendelian disorders, the genetic etiologies underlying complex diseases remain largely unresolved. By mining the medical records of over 110 million patients, we examine the extent to which Mendelian variation contributes to complex disease risk. We detect thousands of associations between Mendelian and complex diseases, revealing a nondegenerate, phenotypic code that links each complex disorder to a unique collection of Mendelian loci. Using genome-wide association results, we demonstrate that common variants associated with complex diseases are enriched in the genes indicated by this "Mendelian code." Finally, we detect hundreds of comorbidity associations among Mendelian disorders, and we use probabilistic genetic modeling to demonstrate that Mendelian variants likely contribute nonadditively to the risk for a subset of complex diseases. Overall, this study illustrates a complementary approach for mapping complex disease loci and provides unique predictions concerning the etiologies of specific diseases.

INTRODUCTION

Clinicians and geneticists have previously observed that rare, Mendelian disorders, such as thalassemia and cystic fibrosis, certain chromosomal abnormalities (such as Down and Kleinfelter syndromes), and severely deleterious copy-number variants (CNV) often predispose patients to more common, apparently nonMendelian diseases. For example, patients with beta-thalassemia, Huntington disease and Friederichs ataxia often develop type 2 diabetes mellitus (De Sanctis et al., 1988; Podolsky et al., 1972; Ristow, 2004), and carriers of the genetic variants associated with Lujan-Fryns and DiGeorge (velo-cardio-facial) syndromes display an increased risk for schizophrenia (De Hert et al., 1996; Sinibaldi et al., 2004). Additionally, bearers of the 16p11.2 microdeletions and microduplications often develop autism (Kumar et al., 2008; Tabet et al., 2012). In such cases, the simple and complex diseases have been long suspected of sharing genetic architecture; whether there is a broader pattern of such associations, however, remains unclear.

A large and growing number of Mendelian and chromosomal diseases have been precisely assigned to particular causal genetic events. Although Mendelian disorders often manifest many of the same complexities that are associated with multigenic diseases, such as incomplete penetrance and genetic modification (Badano et al., 2006), they remain the best understood in terms of their underlying genetic etiologies. This is because the variants underlying Mendelian diseases are generally highly penetrant and nearly unaffected by the environment. Furthermore, their physiologic effects are often severe, allowing for very early diagnosis, sometimes even prenatally. Therefore, in contrast to more complex human disorders, the clinical diagnosis of a Mendelian disease reveals unique insight into the genotype of the affected patient. Consequently, we hypothesize that statistically significant comorbidities between complex and Mendelian illnesses represent a type of genetic association, in

which a non-Mendelian phenotype is mapped to the genetic loci that cause the Mendelian disease.

By analyzing millions of electronic clinical records obtained from distinct regions of the United States and Denmark, we demonstrate that such "transitive" genetic associations are consistent and ubiquitous, yielding insight into the etiology of complex diseases. Furthermore, we observe that each complex disease possesses a unique Mendelian disease allelic architecture, creating a nondegenerate code that identifies each illness by its associated Mendelian loci. In support of our transitive association hypothesis, we demonstrate that complex disease genome-wide association signals are specifically enriched within the genetic loci indicated by this code. Finally, we use mathematical modeling to demonstrate that the variants underlying Mendelian disorders likely interact with one another to contribute to complex disease risk, highlighting the potential of clinical data for uncovering complicated genetic architectures.

RESULTS

Clinical Record Analysis

We mined the administrative data associated with millions of clinical records for evidence of comorbidity among Mendelian and complex diseases. As a rule, such records are maintained in order to facilitate patient billing rather than academic research, and therefore, they may be incomplete and variably biased (van Walraven and Austin, 2012). However, this does not diminish their overall utility for making accurate inferences about clinical phenotypes in large populations. The key to such analyses is to carefully consider how missing data and biases may affect the conclusions of the intended research and, if required, introduce appropriate corrections. Because we conditioned our inferences on the observed disease incidence counts, our comorbidity estimates did not depend on the accurate estimation of marginal disease prevalence. Therefore, we assumed a "missing at random" model for undocumented records that is common practice for epidemiological studies with uninformatively missing data (Lyles and Allen, 2002). Finally, we took great care to focus our data analysis on clearly identifiable phenotypes (see Experimental Procedures), and we detected disease comorbidity using a sophisticated statistical pipeline that accounted for a large set of potentially confounding demographic, socioeconomic, and environmental factors (for details, see Extended Experimental Procedures and Figure S1 available online).

We judged the quality of our statistical inferences by comparing the results generated from multiple, distinct clinical data sets. In the present study, we examined eight data sets, with the smallest and largest describing approximately 150,000 and 100 million unique patients, respectively (see Table 1; Figure 1A). We found that our estimates of the comorbidity risks for the complex-Mendelian disease pairs were remarkably consistent (see Figures 1F and 1G, all correlation p values $< 5 \times 10^{-8}$), which is reassuring considering that the data sets represent populations in different geographic regions with variable ethnic structure and disease prevalence (Figures 1B and 1C). Although the US data set may possibly partially overlap with the smaller, North American ones (CU, NYPH, SU, TX, and UC), the smaller data sets should be nearly completely disjoint from one another and from DK, indicating that duplicate records do not drive this result (see Extended Experimental Procedures for a more detailed treatment of potentially confounding factors). Although other groups have mined clinical record data sets for disease comorbidities in the past (Hidalgo et al., 2009; Lee et al., 2008), the vast majority of the relationships detected in this study are likely to be novel, as associations among complex and Mendelian diseases have never been analyzed at this scale (over 100 million unique patients) (see Figures 1D and 1E for a comparison to previously published results).

Table 1. The Clinical Record Data Sets Utilized in This Study

Data Set	Description	Encoding Type	Number of Unique Patients
CU	Columbia University, 1985–2003	ICD9	1,505,822
DK	Denmark; database covering most of the country's population	ICD10	6,214,312
NYPH	New York Presbyterian Hospital and Columbia University; 2004–present	ICD9	767,978
SU	Stanford University	ICD9	806,369
TX	University of Texas at Houston	ICD9	1,599,528
UC	University of Chicago	ICD9	146,989
USA	MarketScan insurance claims data set	ICD9	99,143,849
MED	Medicare database	ICD9	13,039,018
	Total:		123,223,865

This table provides a brief description, the ICD encoding type, and the size of each data set. The MED data set was used for comparison and was not included in the full meta-analysis.

A Nondegenerate Mendelian Phenotypic Code for Complex Diseases

Figure 2 summarizes all of the significant comorbidities that were detected among the complex and Mendelian disorders within our compendium of clinical records (see Table S4 for detailed results). Each colored cell in the matrix indicates the logarithm of the relative risk associated with a significant clinical signal, and the complex diseases are grouped according to our current understanding of their pathophysiology. Reassuringly, many of the known comorbidities are replicated within our data set. For example, we detected significant comorbidity between lipoprotein deficiencies and myocardial infarction (Strong and Rader, 2012) and ataxia telangiectasia and breast cancer (Sellers, 1997). However, the majority of the 2,909 associations shown in Figure 2 have not been previously reported. For example, our analysis uncovered significant clinical comorbidities between Marfan syndrome and several neuropsychiatric diseases (autism, bipolar disorder, and depression), and it determined that fragile X is significantly associated with asthma, psoriasis, and viral infection, highlighting a potential immune system dysfunction in these patients (Ashwood et al., 2010).

In Figure 3A, the rows and columns of the comorbidity matrix have been rearranged such that disorders with similar

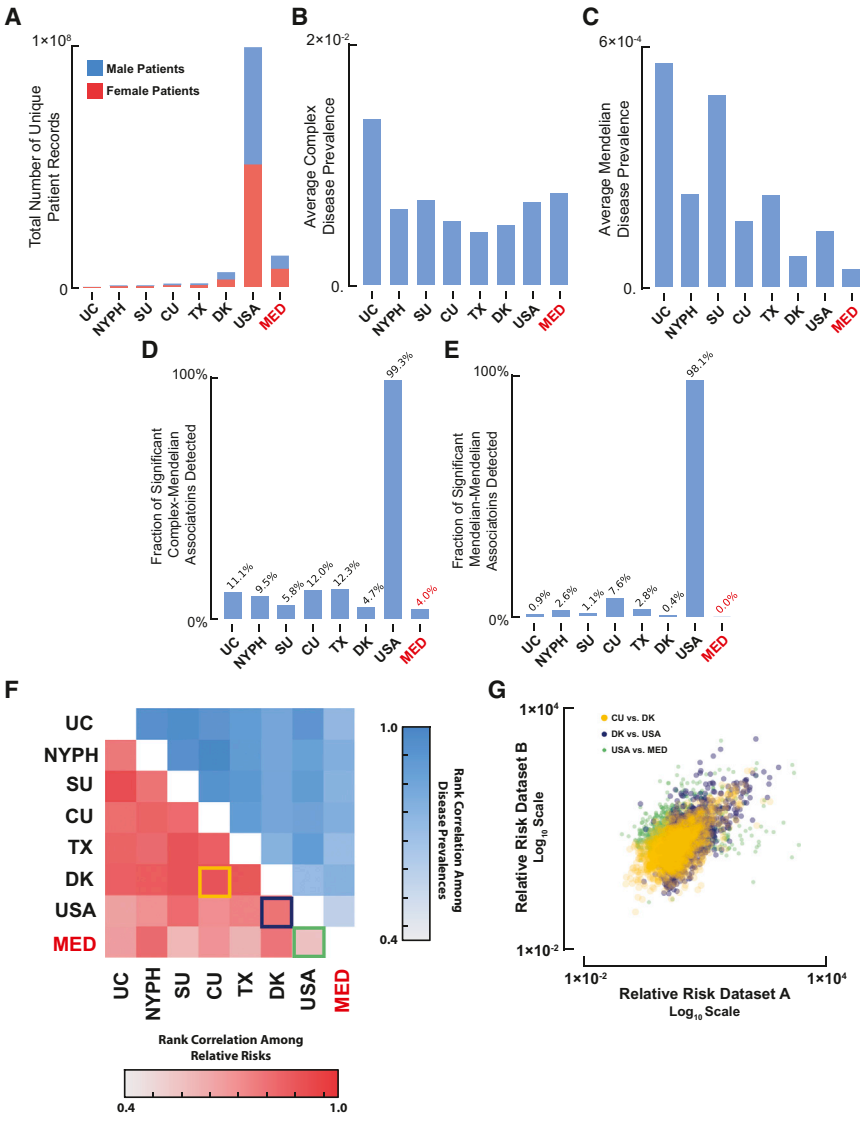


Figure 1. A Systematic Comparison of the Eight Clinical Record Data Sets Analyzed in This Study

(A) The total number of records in each data set, broken down by gender.

(B and C) The average prevalence for the complex and Mendelian diseases across the eight data sets.

(D and E) Using the superset of the discovered associations (based on the original seven data sets; see Extended Experimental Procedures for details), we compared the number of association signals that were detected in each data set independently, depicted as the percentage of all associations discovered in the union of the seven data sets (excluding *MED*): (D) Mendelian-complex and (E) Mendelian-Mendelian associations.

(F) The rank correlation among relative risk estimates (lower diagonal) and disease prevalence (upper diagonal) for each significantly comorbid complex-Mendelian disease pair across the eight distinct data sets.

(G) Scatter plots depicting the relative risk correlations for three pairs of data sets, indicated using the colored boxes in (F).

See also Tables S2 and S3.

comorbidity structure are placed adjacent to one another. Importantly, this rearrangement demonstrates that each complex disease was comorbid with a diverse and unique combination of Mendelian phenotypes. Despite extensive variation within this "Mendelian code," much of our current understanding of the pathophysiology of complex diseases is nonetheless recapitulated (see Figure S2). To illustrate, we computed the Euclidean distance between every pair of shared risk profiles and produced the neighbor-joining tree (Saitou and Nei, 1987) that best approximates this set of statistics (Figure 3B). Not surprisingly, the resulting tree contained many groupings that are highly consistent with our current knowledge of disease etiology. For example, autism, intellectual disability, and epilepsy form a tight cluster in the tree (replicated in 96% of bootstrap pseudosamples), consistent with previous genetic studies that have uncovered variants underlying the risk for all three neuropsychiatric traits (Shinawi et al., 2010).

Complex Disease GWA Signals Are Enriched within the Genetic Loci Implicated by the Mendelian Code

We conjectured that the significant complex-Mendelian comorbidities displayed in Figure 2 indicate that the genes and pathways perturbed in the Mendelian disorders also play a role in the etiology of the corresponding complex diseases. Thus, we hypothesized that the "Mendelian code" could be used to pinpoint loci that harbor complex disease-predisposing genetic variants. To test this prediction, we probed legacy genome-wide association (GWA) results (NIH, 2012) and asked whether common variants associated with the complex diseases were enriched within the loci implicated by the Mendelian comorbidities. Overall, we observed that complex disease GWA signals were globally enriched in Mendelian loci (106 observed, 55.3 expected, 1.92-fold enrichment, $p = 4.0 \times 10^{-10}$), an observation that has been previously highlighted by others (Lupski et al., 2011). Furthermore, when we restricted our analysis to unique signals only (i.e., removed duplicate signals that were replicated in subsequent studies), the enrichment fell to 1.6-fold but remained highly significant (63 observed, 40.4 expected, $p = 4.6 \times 10^{-5}$). Importantly, complex disease-specific GWA signals were specifically enriched in the precise loci indicated by the Mendelian phenotypic code (1.97-fold enrichment, 40 observed, 20.1 expected, $p = 5.7 \times 10^{-5}$, see Table S1 for detailed results), suggesting that the comorbidities highlighted in Figure 2 reflect a shared complex-Mendelian genetic architecture. Moreover, the GWA signals enriched in comorbid

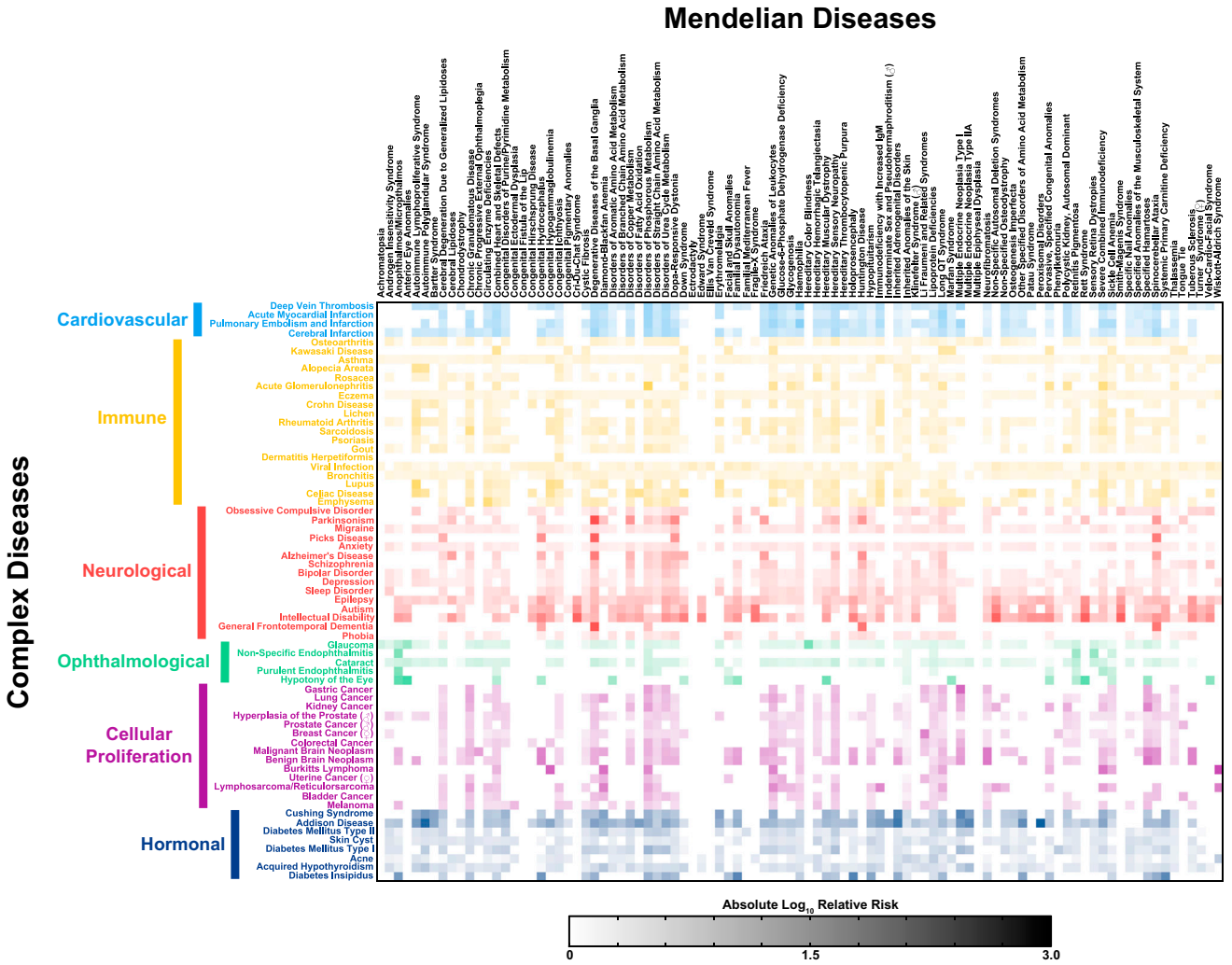


Figure 2. The Significant Comorbidity Relationships among the Complex and Mendelian Disease Pairs

Entries in the matrix indicate the $\log_{10}$-transformed relative risk associated with each significantly comorbid complex-Mendelian disease pair. The complex phenotypes are grouped by our current understanding of their pathophysiology. The symbols ♂ and ♀ indicate male- and female-specific diseases, respectively. The numerical values underlying each association are provided in Table S4. The statistical procedure for generating these values is outlined in Figure S1. See also Tables S1, S2, and S3.

Mendelian loci were more likely to be detected in multiple studies than those in other genic SNPs, including those that lie within noncomorbid Mendelian loci (replication rates: 0.8 versus 0.36, p = 0.026, Mann-Whitney-U test). Overall, these results suggest that the loci implicated by the Mendelian code are likely to contain a spectrum of complex disease predisposing variants, providing testable hypotheses for future gene resequencing and exome analyses (see Discussion for details).

Mendelian Disorders Share Significant Clinical Comorbidity

Our analysis generated a surprisingly large number of statistically significant clinical associations between pairs of Mendelian disorders (462 after conservative statistical filtering; see Extended Experimental Procedures; Figure 4, Figures S3 and S4; Table S5). We propose that these associations represent interactions among genetic variants in distinct Mendelian loci, and we found that it was possible to map individual interactions to specific biological hypotheses. As an example, we observed significant shared risk between fragile X and glycogenosis (odds ratio = 859.09), and this effect remained highly significant after controlling for a wide variety of potentially confounding factors, including disease similarity, age, gender, ethnicity, and environment (see Extended Experimental Procedures). A link between fragile X and glycogenosis has been previously proposed in the molecular genetics literature (De Boulle et al., 1993; Zang et al., 2009), and glycogen metabolism has been suggested to play an important role in fragile X pathophysiology and treatment (Min et al., 2009). A few anecdotal cases aside, however, most of the relationships in Figure 4 represent totally undocumented interactions among rare and highly deleterious genetic variants.

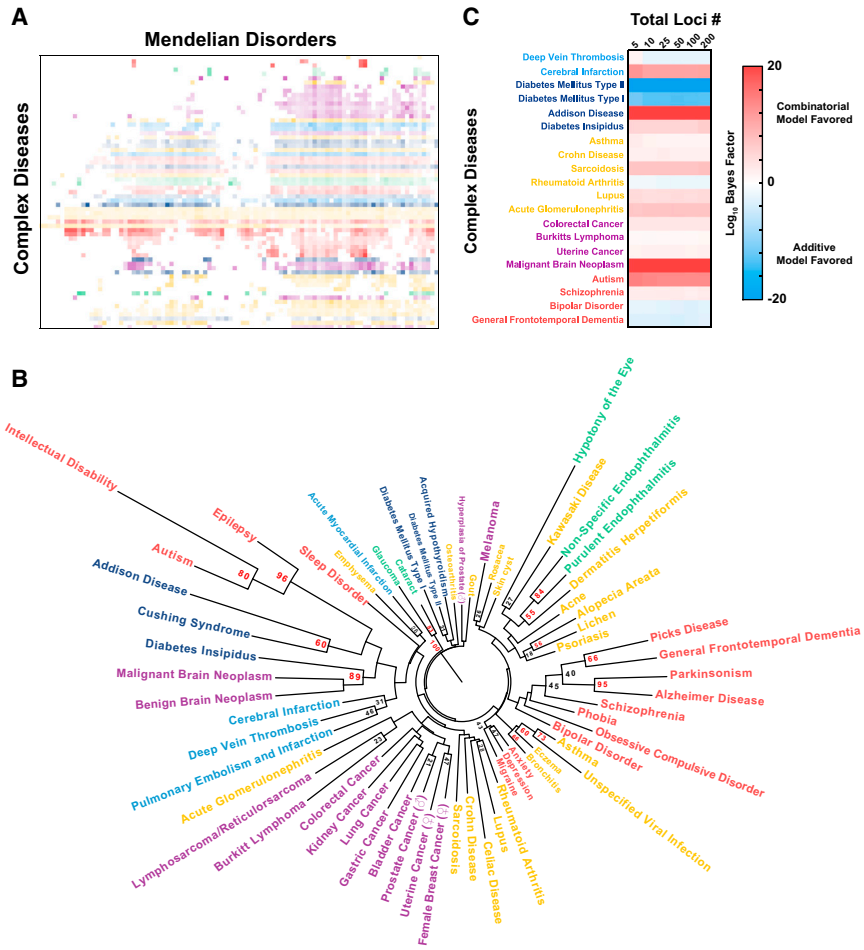


Figure 3. Complex-Mendelian Comorbidities Provide Unique Insight into the Etiology of Complex Diseases

(A) The data matrix from Figure 2 is reordered such that similar rows and columns are adjacent to one another (accomplished using greedy clustering).

(B) The neighbor-joining tree for the complex phenotypes was constructed from the Euclidean distances among the relative risks displayed in Figure 2 and (A). The bootstrap numbers (10,000 replicates) over tree arcs indicate the reliability of the corresponding partitions, with 100 being the most reliable and zero the least. The color of the tree labels is preserved with regard to the groupings of the phenotypes depicted in Figure 2.

(C) Heatmap comparing the qualities of fit for the two multilocus genetic models discussed in the main text over a range of loci numbers. The value of the $\log_{10}$-Bayes factor indicates the support for the combinatorial model in comparison to the additive model. A $\log_{10}$-Bayes factor of one indicates that, given the data, the combinatorial model is ten times more likely than is the additive model. See Figure S5 for a graphical comparison of the model fits to the complex disease risk data. See also Tables S1, S2, and S3 and Figure S2.

We do acknowledge that some of the apparently significant comorbidities could be due to confounding factors. First, miscoding errors during medical billing could create false signals of comorbidity. This could happen, for example, if two distinct physicians examined the same patient but erroneously entered different billing codes because of the clinical ambiguity of the Mendelian disease. Second, the co-occurrence of Mendelian phenotypes could be an artifact of a cryptic population structure. As a result of assortative mating, some subpopulations could be enriched with multiple Mendelian diseases, increasing the apparent rate of rare disease co-occurrence. Although these biases seem plausible, we do not believe that they contribute significantly to the comorbidities depicted in Figure 4 for the following reasons. First, although medical billing errors were likely present in the data sets, we went great lengths to estimate and remove their effects (see Extended Experimental Procedures). Second, our statistical analysis procedure included a variety of demographic and environmental covariates, and we found that these potential confounders contributed only marginally to the shared risk among Mendelian disorders, casting doubt on the cryptic population structure hypothesis.

Perhaps more importantly, there are additional, orthogonal pieces of evidence that indicate that the previous two confounders are unlikely to contribute pervasively to Mendelian-Mendelian comorbidity. For example, we found that comorbid Mendelian disorders, even after removing all clinically similar disease pairs, tended to map to genetic loci that are significantly more functionally alike than is expected by chance, as measured by their distances within a large human gene network (Lee et al., 2011) (see Extended Experimental Procedures, p value < 0.00001). This result fits naturally with the theory of widespread epistasis among Mendelian variants, but it cannot be easily explained using either of the other two hypotheses. Additionally, cryptic population structure, billing code errors, and genetic interactions make very different predictions with respect to complex disease risk in patients diagnosed with multiple comorbid Mendelian disorders (see Experimental Procedures). In the next section, we use probabilistic modeling to provide direct statistical evidence that the risk for several complex diseases is highly consistent with the genetic modifier hypothesis described above.

Mendelian Loci Contribute to Complex Disease Risk in a Nonadditive Manner

Examining the complex disease risk in patients with compound Mendelian phenotypes offered an additional avenue for assessing the likelihood of the three mechanisms proposed in the previous section. As a simple example, assume that the relationships in Figure 4 were dominated by miscoding errors. If this were true, then an individual diagnosed with one comorbid Mendelian disorder should have the same average risk for the complex disease as an individual diagnosed with two. Instead, we

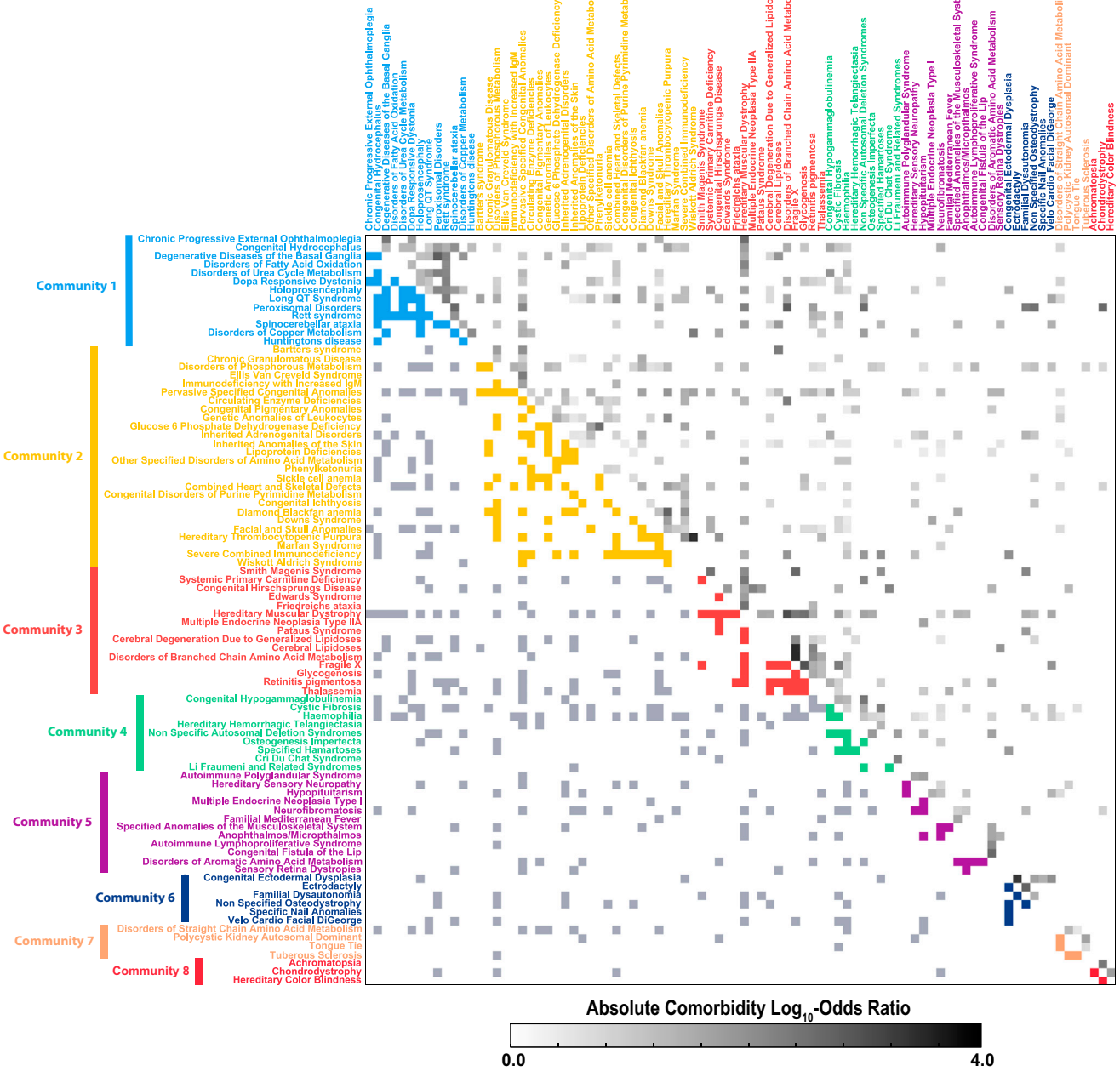


Figure 4. The Significant Comorbidity Relationships Detected among All Pairs of Mendelian Diseases

The upper diagonal of the matrix displays the $\log_{10}$-transformed odds ratios for the significant associations, with grayscale intensity indicating the effect size of the association. The lower diagonal displays the community structure determined using a network-clustering algorithm (Blondel et al., 2008), with each community corresponding to a unique color and associations between diseases within the same community colored accordingly. The numerical values underlying each association are provided in Table S5. The statistical procedure for generating these values is depicted in Figure S3. An unfiltered version of the matrix is displayed in Figure S4.

See also Tables S2, S3, and S5.

observed that individuals diagnosed with two comorbid Mendelian phenotypes had a higher average risk for the complex disease in 62 out of the 65 of the illnesses considered in this study (p value = 6.2×10^{-12}, Wilcoxon signed-rank test). Such analyses provide only indirect evidence for the genetic modifier hypothesis. To provide direct statistical evidence, we formulated two probabilistic genetic models for complex disease risk in patients diagnosed with compound Mendelian phenotypes. The first, termed the additive model (Risch, 1990), is consistent with cryptic population structure and assumes that the

Mendelian variants contribute independently to complex disease risk. The second, called the combinatorial model, invokes a simple mechanism for genetic epistasis among the Mendelian variants. By fitting each model to the clinical data sets, we formally tested whether the genetic modifier hypothesis was supported by the observed risk profiles of the complex diseases.

The two genetic models that we considered share several assumptions in common. First, both assume that each complex disease is associated with a set of genetic loci, some of which are linked to Mendelian phenotypes as well. This assumption ensures that each model is capable of accounting for the comorbidity structure that was observed within the clinical data. Second, the models assume that the genetic loci under consideration possess only dominant, recessive, or X-linked (haploid) variants, although the frequency and penetrance of such variants can vary freely. Third, they assume that the penetrance values for the complex diseases, at both Mendelian and other loci, are sampled from some population-level distribution. Similarly, both models assume that the frequencies of the deleterious genotypes are sampled from a population-level distribution as well. Finally, the models assume that the total number of loci associated with any complex disease is finite and fixed.

The two models differed in one important assumption only: the additive genetic model assumes that the effects of the deleterious genotypes contributed independently (additively) to complex disease risk (Risch, 1990), whereas our nonadditive model breaks this assumption by introducing ''communities'' of loci. Essentially, such communities represented loci that normally function in a coordinated manner, and our nonadditive model assumes that at least one adverse genetic event must be present within multiple communities in order to generate significant complex disease risk. Thus, this community-based genetic model requires combinations of particular deleterious genotypes, so we refer to it as the combinatorial model to differentiate it from other nonadditive genetic mechanisms. In the present study, the combinatorial model was constructed to be as simple as possible and included only two communities of loci.

Although the assumptions outlined above are simple, they generate two models that make distinctly different predictions in terms of the average complex disease risk in patients with multiple comorbid Mendelian phenotypes (see the Extended Experimental Procedures for details). Specifically, the additive model predicts that the average complex disease risk should increase linearly as function of the number of comorbid Mendelian phenotypes, whereas the combinatorial model predicts a *superlinear* (polynomial) increase. Furthermore, if billing record miscoding errors were included into the additive model, the increase in complex disease risk would become *sublinear*. All three signatures were visually apparent in the risk profiles for the complex diseases (see Figure S5), although sublinear increases were rare (approximately 5 out of 65 illnesses). To formally quantify the evidence in favor of each model, we took a Bayesian approach and computed their posterior probabilities conditioned on the clinical data (see Extended Experimental Procedures).

Because of the computational burden associated with fitting genetic models to over 100 million patients, we selected a representative sample of 20 complex diseases for analysis. In practice, the population-level mean of the genotype frequencies and the total number of complex disease predisposing loci were not jointly identifiable, so we repeated the model selection procedure for a range of potential loci numbers (see Experimental Procedures). Each model was clearly favored for a subset of diseases, but the combinatorial model had stronger overall support across the entire set (see Figure 3C). For diseases that displayed a sublinear increase in risk (consistent with possible miscoding errors), the additive model was supported over the combinatorial by a wide margin (see diabetes mellitus type II in Figure S5). Overall, this result provides additional and orthogonal support for the hypothesis that Mendelian-Mendelian comorbidities were driven by genetic interactions. It also suggests that certain complex diseases (such as Addisons disease, acute glomerulonephritis, and malignant brain neoplasms, but not the two forms of diabetes or bipolar disorder) have a nonadditive (epistatic) genetic architecture with respect to Mendelian disease variants.

DISCUSSION

Highly penetrant mutations have not been found for most common, complex diseases, despite intensive search. Although rare single-nucleotide and copy-number variants have been implicated in some complex disorders, including intellectual disability (Vissers et al., 2010), schizophrenia (Bassett et al., 2008) and autism (Iossifov et al., 2012), these results appear to be the exception rather than the norm. The fact that we observed widespread comorbidity among Mendelian and complex diseases suggests that rare, highly penetrant variants do in fact play a significant role in complex disease risk, but their deleterious effects do not result in single, isolated diseases. Instead, highly deleterious genetic variants likely induce a variety of pathological consequences, consistent with the Mendelian code displayed in Figures 2 and 3A. Such analysis resonates with the results of recent genetic dissections of oligogenic traits, such as Bardet-Biedl syndrome, which appears to harbor a diverse genetic architecture that produces a variety of clinical phenotypes (Katsanis et al., 2001; Zaghloul et al., 2010).

In addition to these direct associations, we also observed that common risk variants associated with complex diseases were specifically enriched in comorbid Mendelian loci. The most obvious explanation for this is that some of the patients included in GWA studies carried genetic variation that predisposed them to both the Mendelian and complex diseases. However, there are several reasons to be skeptical of this hypothesis. First, subjects with Mendelian disorders are typically, by design, excluded from GWAS (Zhao et al., 2010). Second, Mendelian diseases are rare and have overt clinical presentations, so the unintentional inclusion of such carriers in the studies is highly improbable. Finally, even if the rate of accidental sampling of Mendelian phenotypes were aberrantly high, we do not believe that ''synthetic'' genome-wide associations, in which the detected common variants are in linkage disequilibrium with Mendelian disease alleles, drive our results (Dickson et al., 2010). As discussed at length by others (Visscher et al., 2012), numerous empirical and theoretical analyses are simply not consistent with this interpretation.

As an alternative explanation, we and others (Lupski et al., 2011) propose that Mendelian genes carry both rare and

common deleterious variants, such that alleles from both ends of the frequency spectrum contribute to disease risk. Rare, highly penetrant variants cause Mendelian disorders, whereas common variants with milder effects contribute to the complex phenotypes. By design, GWAS detect only the latter end of the frequency spectrum, and the former is typically uncovered through linkage analysis and sequencing. When the Mendelian and complex phenotypes are similar, we can think of the two disorders as different ends of the same genetic and phenotypic spectrum, known as the allelic series hypothesis. In fact, there are several well-documented examples of this phenomenon, such as the familial and common forms of Parkinsonism and blood lipid disorders (Manolio et al., 2009).

However, aside from a few special cases, this straightforward definition of allelic series is not very helpful when explaining Mendelian and complex phenotypes that are comorbid and share genetic loci but are biologically dissimilar. For example, asthma and systemic primary carnitine deficiency share clinical risk and are both associated with variants in the *SLC22A5* locus, but there is no obvious relationship between the biology underlying these two diseases. Instead, we suggest a modification to the allelic series hypothesis that considers the multifactorial nature of gene function. On one end of the spectrum, we hypothesize that very rare, Mendelian disease variants completely or nearly completely abolish all of a gene's physiological functions. Therefore, their effects are highly penetrant and pleiotropic, resulting in overt pathologies (like Mendelian disease), while increasing a carrier's risk for a variety of other disorders. On the other end, less deleterious mutations may perturb the same genes, but their effects are more limited, perhaps modifying only a subset of a gene's functions. In such instances, the resulting deleterious effects may be quite subtle, allowing the variants to reach relatively high population frequencies. Moreover, their ultimate pathological manifestations may be very different than those that are observed in patients harboring Mendelian variants, reflecting the different subsets of physiological functions perturbed by each mutation type.

With this in mind, we hypothesize that the loci underlying comorbid Mendelian disorders represent strong candidates for harboring complex disease-predisposing genetic variants with moderate and weak effects, as the Mendelian associations have already suggested that the underlying gene is involved in the pathophysiology of the complex disorder. This theory is supported by our GWAS enrichment results, but we believe that it extends to rare variants with larger effects as well. Because they have already been shown to contain a variety of complex disease predisposing variants, we propose that the best candidates for testing this hypothesis are perhaps those loci that were found to contain both common risk and Mendelian disease-causing variants (see Table S1). Consistent with this hypothesis, we note that 4 out of the 7 neoplasms for which GWAS results were available were found to associate with both common and rare Mendelian genetic variants in the *TERT* locus, which encodes the human telomerase reverse transcriptase. Mendelian variants within this locus completely abolish reverse-transcriptase enzymatic activity, resulting in several overt, pathological symptoms (combined into a syndrome called dyskeratosis congenita) (Kirwan and Dokal, 2009). Recently, a rare germline variant in the promoter region of *TERT* was linked to a familial form of melanoma, although carriers of the allele may have increased risk for other neoplasms as well (Horn et al., 2013). In support, somatic variants within the promoter region of *TERT* were also found in a variety of human cancer cell lines (Huang et al., 2013) and solid tumors (Killela et al., 2013). Such results raise the intriguing possibility that a spectrum of *TERT*-associated variants, both rare and common, somatic and germline, increase one's risk for neoplastic disease.

Furthermore, our complex-Mendelian comorbidity analysis predicted that schizophrenia, bipolar disorder, autism, and depression are all associated with the following four Mendelian loci: *SYNE1*, *PRPF3*, *CACNA1C*, and *PPP2R2B*. Consistent with their hypothesized shared genetic architecture (Cross-Disorder Group of the Psychiatric Genomics Consortium et al., 2013), these four loci were also found to harbor common genetic variants that influence risk for this same set of diseases. Interestingly, exome sequencing in autism patients has uncovered both de novo and inherited potentially deleterious variants in *SYNE1* (O'Roak et al., 2011; Yu et al., 2013). We find this result particularly interesting, as it suggests that these four genes may also harbor rare variants that predispose carriers to multiple neuropsychiatric disorders. If this is correct, then pooling strategies that combine sequence data from patients with these different, but related, complex phenotypes could offer a simple approach for increasing the power to identify rare variants with modest effects.

In the second part of our study, we discovered approximately 450 comorbidity associations among pairs of Mendelian disorders, suggesting that genetic interactions among Mendelian variants are quite common. Consistent with this hypothesis, we used genetic modeling to demonstrate that epistatic effects could be detected in the complex disease risk profiles of patients diagnosed with multiple, comorbid Mendelian disorders. At the very least, our results suggest that strongly deleterious variants have a high propensity for modifying the effects of other deleterious alleles in functionally similar genes. However, the existence of nonadditive effects among rare genetic variants could have practical consequences as well. For example, undocumented epistasis among rare variants in distinct loci could negatively impact the power of targeted resequencing studies.

Although our inference of widespread, nonadditive genetic effects is novel, the fact that highly penetrant genetic variants are subject to modification by other alleles that exist *in trans* is well known. For example, at first glance, the Mendelian disorder retinitis pigmentosa appears to follow the "independent effects" assumption of genetic additivity quite well (Parmeggiani, 2011), as several, highly penetrant mutations in distinct genes have been associated with the phenotype. However, this disease was also one of the first Mendelian phenotypes with clearly demonstrated digenic inheritance (Kajiwara et al., 1994), and epistatic interactions among multiple loci have been reported for other Mendelian phenotypes as well, such as Bardet-Biedl syndrome (Badano et al., 2006). There are also known examples in which *trans* genetic variants modify the specific symptoms of Mendelian disorders. More specifically, several suspected genetic modifiers have been previously identified for cystic fibrosis (CF) (Cutting, 2010), a recessive disease caused by mutations in

the *CFTR* gene. CF patients display a variety of symptoms, including mucus congestion in the lungs, intestinal obstruction, diabetes, abnormal gut microflora, and liver disease, and nearly a dozen loci have been identified that appear to modulate the strength of these clinical symptoms (Cutting, 2010). For example, variation in *EDNRA* appears to affect the pulmonary function of CF patients, whereas *MSRA* alleles modulate intestinal obstruction.

In summary, we detected thousands of instances of comorbidity between complex-Mendelian and Mendelian-Mendelian disease pairs. The existence of such associations was not unexpected; however, their widespread nature was surprising. Furthermore, although there is a growing body of evidence that genetic interactions are common across both Mendelian and complex traits, such as Alzheimer's disease (Badano and Katsanis, 2002), facioscapulohumeral dystrophy type 2 (Lemmers et al., 2012), and Hirschsprungs disease (Wallace and Anderson, 2011), we believe that this is the first instance in which such relationships have been uncovered systematically across multiple complex diseases. Ultimately, we demonstrate that digital phenotypic data can be utilized to infer genetic and genomic architectures, potentially allowing for extensive, novel analyses in the field of human disease genetics. Moreover, this work highlights the importance of documenting a wider spectrum of Mendelian and other disease traits in a very large population of humans, perhaps the entire United States or even multiple countries, in order to uncover the pathophysiology associated with very rare genetic events.

EXPERIMENTAL PROCEDURES

Phenotype Curation and Billing Code Assignments

To identify the clinical phenotypes of interest, we used the disease codes provided by the International Disease Classification (ICD) system (WHO, 2010) (see Table 1). The mappings between billing codes (both ICD9 and ICD10) and diseases were obtained from Rzhetsky et al. (2007) and by manual curation, first by a PhD-level contractor trained in a biomedical field and second by two of the authors, iteratively. All billing code mappings for the complex and Mendelian diseases are provided in Tables S2 and S3, respectively. The billing codes enabled the identification of 65 specific complex disorders and 95 Mendelian disease groups (representing 213 disorders) (see Tables S2 and S3, respectively). Note, this reduction of 213-to-95 was not a choice of experimental design but was necessitated by the ICD9 code taxonomy. See Extended Experimental Procedures for additional details.

Clinical Record Analysis

Each clinical record database was first parsed (see Table 1), removing duplicate records and identifying patients that harbored the diseases of interest. In theory, a small fraction of these records could be shared between US and the other, smaller US data sets (CU, NYPH, SU, TX, UC) because some patients could have been documented in multiple databases. Because duplicate records would strongly bias the results for rare diseases, we decided against simply combining the information from different data sets into a single meta-analysis. Instead, we performed an independent statistical analysis for each data set and then combined the results according to a conservative procedure (see Extended Experimental Procedures for details). For the complex-Mendelian comorbidity analysis, any disease pair containing a complex or Mendelian disease that was specific to males or females (indicated by ♂ and ♀, respectively, in Figure 2) was analyzed after conditioning on the appropriate gender; gender-specific diseases were not included in the Mendelian-Mendelian analysis. The *MED* data set (Hidalgo et al., 2009; Lee et al., 2008) was excluded from the meta-analysis, as we were unable to consistently identify our phenotypes of interest. Specifically, the *MED* data set provides individual ICD9 code counts only, but many of the disorders used in our analysis map to multiple such codes. Additional details concerning our statistical procedures for the analysis of complex-Mendelian and Mendelian-Mendelian disease pairs are provided in the Extended Experimental Procedures.

Neighbor-Joining Tree Inference

The complex disease tree was constructed from the Mendelian comorbidity relationships using the neighbor-joining method (Saitou and Nei, 1987). See Extended Experimental Procedures for additional details.

GWAS Enrichment Analysis

To test for an enrichment of common, complex disease risk variants in Mendelian loci, we aligned legacy genome-wide association results (NIH, 2012) with the SNP-to-gene annotations provided by SCAN (Gamazon et al., 2010). Binomial tests that specifically controlled for gene length and SNP annotation biases were used to assess enrichment (see Extended Experimental Procedures for details).

The Additive and Nonadditive Genetic Models for Complex Disease Risk

In the main text, we briefly described two competing genetic models that specify distinct mechanisms for how multiple Mendelian disease variants combine to affect complex disease risk. Ultimately, the additive and combinatorial models make very different predictions with respect to the increase in complex disease risk as a function of the number of comorbid Mendelian phenotypes, allowing them to be differentiated within our massive clinical data sets. The mathematical details concerning this prediction are somewhat involved, and the interested reader should consult the Extended Experimental Procedures. In the following section, we simply introduce our competing genetic models using standard notation.

Consistent with common practice (Risch, 1990), each of our genetic models treats the genotype (g) and phenotype (ϕ) of an individual as random variables. Their joint probability is equivalent to the expected population frequency of individuals that possess both a particular genotype (G) and disease of interest (D). It is computed by taking the product of the genotype frequency and its corresponding penetrance:

$$P(\phi = D, g = G|\Theta) = P(g = G|\Theta)P(\phi = D|g = G, \Theta) = F(G) \times W_D(G),$$

where $F(G)$ is the probability of observing genotype G and $W_D(G)$ is the genetic penetrance of G with respect to phenotype D (i.e., the probability of D given G) (Risch, 1990). The overall expected prevalence of the disease within the population is computed by summing the previous probability over all possible genotypes:

$$P(\phi = D|\Theta) = \sum_G F(G) \times W_D(G).$$

Although not included for the sake of simplicity, environmental factors can be easily incorporated into this framework through the inclusion of additional random variables.

Our additive genetic model is specified within the previous framework by defining the following simple penetrance function (Risch, 1990):

$$W_D(G) = 1 - \prod_{i=1}^{n}[1 - W_D(G_i)],$$

where n is the number of independent loci affecting phenotype D, and $W_D(G_i)$ is the marginal penetrance function of the genotype at the i^{th} locus (Risch, 1990) that may take a variety of forms (dominant, recessive, additive, etc.). Technically, the model assumes that each locus contributes independently to complex disease risk, and this assumption generally underlies most "additive" models in human genetics. That said, it also approximates a stricter definition of "additivity," in which the probability of the complex disease is simply the linear combination of the penetrance probabilities of the individual loci (Risch, 1990).

Our nonadditive genetic model assumes that the deleterious genotypes belong to a different "communities" of loci that act coordinately, and at least one adverse genetic event must be present within multiple communities in order to generate significant complex disease risk. Because this model requires combinations of deleterious alleles, we call it the "combinatorial" model. To illustrate, imagine two disjoint groups of loci, or "communities," each harboring a set of genotypes that predispose an individual to the disease of interest. We denote the two communities using circle and square subscripts, such that $\{g_{\bigcirc,1}, g_{\bigcirc,2}, ..., g_{\bigcirc,n_\bigcirc}\}$ and $\{g_{\square,1}, g_{\square,2}, ..., g_{\square,n_\square}\}$ denote the genetic loci that belong to each community and $n_\bigcirc$ and $n_\square$ denote community sizes. To simplify notation, we will indicate either the square or the circle community, depending on context, using the $\mathcal{C}$ symbol $(\mathcal{C} = \{\bigcirc, \square\})$. Assuming an additive model within each community, the penetrance function for the two-community combinatorial model is

$$W_D(G) = \prod_{\mathcal{C} \in \{\bigcirc,\square\}} \left(1 - \prod_{i=1}^{n_\mathcal{C}} [1 - W_D(G_i)]\right).$$

Note that more general formulations of the model could allow for more than two communities and a variety of different community- and loci-specific penetrance functions.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.08.030.

ACKNOWLEDGMENTS

We are grateful to Steven Bagley, Richard R. Hudson, Ivan Iossifov, Ravinesh Kumar, Simon Lovestone, Fabiola Rivas, Gregory Gibson, Jason Pitt, Rita Rzhetsky, Michael Wigler, and anonymous reviewers for helpful comments on earlier versions of the manuscript. GeneXplain, GmbH, provided help with annotation of Mendelian disorders. This work was supported by grants (1P50MH094267, NHLBI MAPGen U01HL108634-01, P50GM081892-01A1, and 2T32GM007281-39) from the National Institutes of Health and by a Lever Award from the Chicago Biomedical Consortium.

Received: December 17, 2012
Revised: March 30, 2013
Accepted: August 16, 2013
Published: September 26, 2013

REFERENCES

Ashwood, P., Nguyen, D.V., Hessl, D., Hagerman, R.J., and Tassone, F. (2010). Plasma cytokine profiles in Fragile X subjects: is there a role for cytokines in the pathogenesis? Brain Behav. Immun. *24*, 898–902.

Badano, J.L., and Katsanis, N. (2002). Beyond Mendel: an evolving view of human genetic disease transmission. Nat. Rev. Genet. *3*, 779–789.

Badano, J.L., Leitch, C.C., Ansley, S.J., May-Simera, H., Lawson, S., Lewis, R.A., Beales, P.L., Dietz, H.C., Fisher, S., and Katsanis, N. (2006). Dissection of epistasis in oligogenic Bardet-Biedl syndrome. Nature *439*, 326–330.

Bassett, A.S., Marshall, C.R., Lionel, A.C., Chow, E.W., and Scherer, S.W. (2008). Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. Hum. Mol. Genet. *17*, 4045–4053.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *10*, 10008–10020.

Calderhead, B., and Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. Comput. Stat. Data Anal. *53*, 4028–4045.

Cross-Disorder Group of the Psychiatric Genomics Consortium, Smoller, J.W., Craddock, N., Kendler, K., Lee, P.H., Neale, B.M., Nurnberger, J.I., Ripke, S., Santangelo, S., and Sullivan, P.F. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet *381*, 1371–1379.

Cutting, G.R. (2010). Modifier genes in Mendelian disorders: the example of cystic fibrosis. Ann. N Y Acad. Sci. *1214*, 57–69.

De Boulle, K., Verkerk, A.J.M.H., Reyniers, E., Vits, L., Hendrickx, J., Van Roy, B., Van den Bos, F., de Graaff, E., Oostra, B.A., and Willems, P.J. (1993). A point mutation in the FMR-1 gene associated with fragile X mental retardation. Nat. Genet. *3*, 31–35.

De Hert, M., Steemans, D., Theys, P., Fryns, J.P., and Peuskens, J. (1996). Lujan-Fryns syndrome in the differential diagnosis of schizophrenia. Am. J. Med. Genet. *67*, 212–214.

De Sanctis, V., Zurlo, M.G., Senesi, E., Boffa, C., Cavallo, L., and Di Gregorio, F. (1988). Insulin dependent diabetes in thalassaemia. Arch. Dis. Child. *63*, 58–62.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. PLoS Biol. *8*, e1000294.

Gamazon, E.R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E.O., Nicolae, D.L., Dolan, M.E., and Cox, N.J. (2010). SCAN: SNP and copy number annotation. Bioinformatics *26*, 259–262.

Hidalgo, C.A., Blumm, N., Barabási, A.L., and Christakis, N.A. (2009). A dynamic network approach for the study of human phenotypes. PLoS Comput. Biol. *5*, e1000353.

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. Science *339*, 959–961.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science *339*, 957–959.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285–299.

Kajiwara, K., Berson, E.L., and Dryja, T.P. (1994). Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. Science *264*, 1604–1608.

Katsanis, N., Ansley, S.J., Badano, J.L., Eichers, E.R., Lewis, R.A., Hoskins, B.E., Scambler, P.J., Davidson, W.S., Beales, P.L., and Lupski, J.R. (2001). Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. Science *293*, 2256–2259.

Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L.A., Jr., Friedman, A.H., Friedman, H., Gallia, G.L., Giovanella, B.C., et al. (2013). TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. Proc. Natl. Acad. Sci. USA *110*, 6021–6026.

Kirwan, M., and Dokal, I. (2009). Dyskeratosis congenita, stem cells and telomeres. Biochim. Biophys. Acta *1792*, 371–379.

Kumar, R.A., KaraMohamed, S., Sudi, J., Conrad, D.F., Brune, C., Badner, J.A., Gilliam, T.C., Nowak, N.J., Cook, E.H., Jr., Dobyns, W.B., and Christian, S.L. (2008). Recurrent 16p11.2 microdeletions in autism. Hum. Mol. Genet. *17*, 628–638.

Lee, D.S., Park, J., Kay, K.A., Christakis, N.A., Oltvai, Z.N., and Barabási, A.L. (2008). The implications of human metabolic network topology for disease comorbidity. Proc. Natl. Acad. Sci. USA *105*, 9880–9885.

Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. *21*, 1109–1121.

Lemmers, R.J., Tawil, R., Petek, L.M., Balog, J., Block, G.J., Santen, G.W., Amell, A.M., van der Vliet, P.J., Almomani, R., Straasheijm, K.R., et al. (2012). Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nat. Genet. *44*, 1370–1374.

Lupski, J.R., Belmont, J.W., Boerwinkle, E., and Gibbs, R.A. (2011). Clan genomics and the complex architecture of human disease. Cell *147*, 32–43.

Lyles, R.H., and Allen, A.S. (2002). Estimating crude or common odds ratios in case-control studies with informatively missing exposure data. Am. J. Epidemiol. *155*, 274–281.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Min, W.W., Yuskaitis, C.J., Yan, Q., Sikorski, C., Chen, S., Jope, R.S., and Bauchwitz, R.P. (2009). Elevated glycogen synthase kinase-3 activity in Fragile X mice: key metabolic regulator with evidence for treatment potential. Neuropharmacology *56*, 463–472.

NIH. (2012). http://www.genome.gov/admin/gwascatalog.txt.

O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet. *43*, 585–589.

Parmeggiani, F. (2011). Clinics, epidemiology and genetics of retinitis pigmentosa. Curr. Genomics *12*, 236–237.

Podolsky, S., Leopold, N.A., and Sax, D.S. (1972). Increased frequency of diabetes mellitus in patients with Huntington's chorea. Lancet *1*, 1356–1358.

Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. Am. J. Hum. Genet. *46*, 222–228.

Ristow, M. (2004). Neurodegenerative disorders associated with diabetes mellitus. J. Mol. Med. *82*, 510–529.

Rzhetsky, A., Wajngurt, D., Park, N., and Zheng, T. (2007). Probing genetic overlap among complex human phenotypes. Proc. Natl. Acad. Sci. USA *104*, 11694–11699.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.

Sellers, T.A. (1997). Genetic factors in the pathogenesis of breast cancer: their role and relative importance. J. Nutr. *127*(5, Suppl), 929S–932S.

Shinawi, M., Liu, P., Kang, S.H., Shen, J., Belmont, J.W., Scott, D.A., Probst, F.J., Craigen, W.J., Graham, B.H., Pursley, A., et al. (2010). Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. J. Med. Genet. *47*, 332–341.

Sinibaldi, L., De Luca, A., Bellacchio, E., Conti, E., Pasini, A., Paloscia, C., Spalletta, G., Caltagirone, C., Pizzuti, A., and Dallapiccola, B. (2004). Mutations of the Nogo-66 receptor (RTN4R) gene in schizophrenia. Hum. Mutat. *24*, 534–535.

Strong, A., and Rader, D.J. (2012). Sortilin as a regulator of lipoprotein metabolism. Curr. Atheroscler. Rep. *14*, 211–218.

Tabet, A.C., Pilorge, M., Delorme, R., Amsellem, F., Pinard, J.M., Leboyer, M., Verloes, A., Benzacken, B., and Betancur, C. (2012). Autism multiplex family with 16p11.2p12.2 microduplication syndrome in monozygotic twins and distal 16p11.2 deletion in their brother. Eur. J. Hum. Genet. *20*, 540–546.

van Walraven, C., and Austin, P. (2012). Administrative database research has unique characteristics that can risk biased results. J. Clin. Epidemiol. *65*, 126–131.

Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.

Vissers, L.E., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M., et al. (2010). A de novo paradigm for mental retardation. Nat. Genet. *42*, 1109–1112.

Wallace, A.S., and Anderson, R.B. (2011). Genetic interactions and modifier genes in Hirschsprung's disease. World J. Gastroenterol. *17*, 4937–4944.

WHO. (2010). http://www.who.int/classifications/icd/en/.

Yu, T.W., Chahrour, M.H., Coulter, M.E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., Schmitz-Abe, K., Harmin, D.A., Adli, M., Malik, A.N., et al. (2013). Using whole-exome sequencing to identify inherited causes of autism. Neuron *77*, 259–273.

Zaghloul, N.A., Liu, Y., Gerdes, J.M., Gascue, C., Oh, E.C., Leitch, C.C., Bromberg, Y., Binkley, J., Leibel, R.L., Sidow, A., et al. (2010). Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. Proc. Natl. Acad. Sci. USA *107*, 10602–10607.

Zang, J.B., Nosyreva, E.D., Spencer, C.M., Volk, L.J., Musunuru, K., Zhong, R., Stone, E.F., Yuva-Paylor, L.A., Huber, K.M., Paylor, R., et al. (2009). A mouse model of the human Fragile X syndrome I304N mutation. PLoS Genet. *5*, e1000758.

Zhao, J., Bradfield, J.P., Zhang, H., Annaiah, K., Wang, K., Kim, C.E., Glessner, J.T., Frackelton, E.C., Otieno, F.G., Doran, J., et al. (2010). Examination of all type 2 diabetes GWAS loci reveals HHEX-IDE as a locus influencing pediatric BMI. Diabetes *59*, 751–755.

Chapter 7:
Concluding remarks

7.1 A systems approach to epidemiology

During the data-driven work with disease correlations, I found that lists of correlations that are candidates for further studies are of very little interest themselves. There must be some plan for the further studies of the candidate lists. They could be used to identify hypotheses for further examinations in a traditional epidemiological study. However, it is a time consuming task to identify interesting candidates among the thousands of significant correlations. Ultimately, it is questionable if this process is better or faster than the processes used to identify hypotheses in hypothesis driven research. The temporal analysis presented in manuscript I is an example of how a list of comorbidities can be converted into trajectories and used for discovering patterns across not only pairs of diseases, but groups of diseases. This is a better strategy to utilize the results of data-driven analyses. The clustering method presented in the work in progress with non-temporal correlations is also an approach to investigate the full system of correlations rather than just pairs of diseases.

Traditional epidemiology seeks to investigate the relation between one disease and a risk factor and eliminate the signal from all other diseases and risk factors. Regardless of the corrections made, this is a very difficult task. Diseases have a complex nature, and often there is no simple causal explanation to their development. The data-driven studies presented in this thesis are a part of a move towards doing epidemiological studies of the full system of correlations rather than attempting to reduce the outcome to a single risk factor.

7.1.1 Improvements of the correlation measures

Although the data-driven studies were epidemiological, they lack the stringency that is expected from a traditional epidemiological study. Future work on disease correlation needs to incorporate basic techniques from epidemiology [15].

In the work on both non-temporal correlations and the temporal correlations, the majority of correlations were positive. As discussed in chapter 3, this can be because of positive feedback: Sick and weak people get sicker. In manuscript I and the work in progress I, the frequency expected by random given no correlation between a pair of diseases was used to calculate RR. Thus, the null hypothesis is that diseases co-occur randomly when the age and gender signal is controlled for. This is a weak null-hypothesis, when diseases correlate positively even when there is no direct causality. This is especially a problem, if trying to make statistical inference on disease trajectories. A patient who has a history of four different diseases is likely to be weaker than a patient who is picked randomly out of the full population. To improve the null-model, patients need to be matched with control patients who have similar medical history. The control for inpatient, outpatient and ER encounters indirectly does this to some degree. Severe diseases are mainly treated in inpatient care and injuries. Other diseases of less severity

are treated in ERs, and in case of severe incidents, an ER or outpatient would be admitted to the hospital. The matching was done more stringent in the colectomy study, where colectomy patients were matched with patients having undergone other kind of surgical procedures.

Another issue is in the ICD-10 classification. A diagnosis such as I10 Essential hypertension is similar to I11 Hypertensive heart disease and I13 Hypertensive heart and renal disease in that all the diagnoses refer to the same condition: hypertension. It would be beneficial for both the statistics and the medical interpretation to combine these. While it is feasible to combine the diagnoses in a study of few diseases, it is a large and tedious task in a study of all diseases. The main problem is that there is no objective true set of basic diseases and conditions. The differences between classifications demonstrate this very well. Any attempt to reduce the full set of ICD-10 diagnoses to a smaller set of basic diseases will depend on personal judgment.

7.2 Opportunities and limitations

Several other interesting data sources can be used for analyzing correlations and disease progression. Patient health records are a source of medical knowledge that can be used in a similar fashion to the registries. With the transition from paper journals to electronic health records (EHR), performing large-scale studies on health records is becoming more widespread [77]. In addition to the information that is reported to disease registries, the EHRs provide access to structured data such as drug prescriptions and clinical text. This allows for much more detailed diagnostic description of patients. It is likely that data on more confounding factors can be found in this data.

Using the CPR data it is possible to link family members together, which would give the opportunity to investigate heritable diseases in a large scale. This could be used to assess which patients have a genetic variant of a disease. It is also important to have the link between mother and child when studying diagnoses related to pregnancy. Some codes that are relevant to the child are registered in the mother's records. Due to this, we did not analyze diagnoses regarding births in the data-driven correlation analyses.

It is well known that income and other socioeconomic factors are confounding factors for diseases (e.g. diabetes and obesity [80], and cancer [81]). Data on socioeconomic factors is available from Statistics Denmark and can contribute in further studies. The main challenge here is to do the large-scale analysis within the systems of Statistic Denmark.

Another interesting data source is drug prescriptions. Correlations can be a result of adverse drug effects if a drug is used to treat one disease, but cause another disease (or medical condition) as adverse effect. Removing the drug bias would be beneficial. Having these data for a large population would also give opportunity to do Comparative Effectiveness Research (CER). Different combinations of drugs and other medical treatment could be analyzed in order to assess the efficacy. Drugs have an obvious effect on disease progression, and prescription data could be a key variable in modeling disease trajectories. Other use of this data includes drug surveillance [82,83] and identification of adverse effects from drugs [84].

Even though NPR offers the full hospitalization history of patients, we lack data from the private practitioner. This is an important part of the medical history, as many contacts will never lead to hospitalization. Unfortunately, the Danish public health insurance registry does not contain diagnoses. It is used solely for reimbursement of the private practitioners, who are generally paid per contact regardless of the diagnosis. Therefore only very few codes in the registry point to a specific disease.

References

1 Morel NM, Holland JM, Van der Greef J, *et al.* Primer on Medical Genomics Part XIV: Introduction to Systems Biology—A New Approach to Understanding Disease and Treatment. *Mayo Clinic Proceedings* 2004;**79**:651–8.http://www.sciencedirect.com/science/article/pii/S0025619611622878 (accessed 16 Oct2013).

2 Weston AD, Hood L. Systems Biology , Proteomics , and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine Introduction: Paradigm Changes in Health Care. *Journal of Proteome Research* 2004;**3**:179–96.

3 Snoep J, Westerhoff H. From isolation to integration, a systems biology approach for building the Silicon Cell. In: *Topics in Current Genetics Volume 13*. Berlin: : Springer-Verlag 2005. 13–30. doi:10.1007/b106456

4 Hidalgo CA, Blumm N, Barabási A-L, *et al.* A dynamic network approach for the study of human phenotypes. *PLoS computational biology* 2009;**5**:e1000353. doi:10.1371/journal.pcbi.1000353

5 Roque FS, Jensen PB, Schmock H, *et al.* Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS Computational Biology* 2011;**7**:e1002141. doi:10.1371/journal.pcbi.1002141

6 Lage K, Karlberg EO, Størling ZM, *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology* 2007;**25**:309–16. doi:10.1038/nbt1295

7 Rzhetsky A, Wajngurt D, Park N, *et al.* Probing genetic overlap among complex human phenotypes. *Proceedings of the National Academy of Sciences of the United States of America* 2007;**104**:11694–9. doi:10.1073/pnas.0704820104

8 O'Brien SJ, Nelson GW. Human genes that limit AIDS. *Nature genetics* 2004;**36**:565–74. doi:10.1038/ng1369

9 Jain KK. *Textbook of Personalized Medicine*. 1st ed. New York, NY: : Springer New York 2009. doi:10.1007/978-1-4419-0769-1

10 Thaker PH, Han LY, Kamat A a, *et al.* Chronic stress promotes tumor growth and angiogenesis in a mouse model of ovarian carcinoma. *Nature medicine* 2006;**12**:939–44. doi:10.1038/nm1447

11 Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New biotechnology* 2012;**29**:613–24. doi:10.1016/j.nbt.2012.03.004

12 Hamburg MA, Collins FS. The Path to Personalized Medicine. *New England Journal of Medicine* 2010;**363**:301–4. doi:DOI: 10.1056/NEJMp1006304

13 Lage K, Karlberg EO, Størling ZM, *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology* 2007;**25**:309–16. doi:10.1038/nbt1295

14 Agusti A, Sobradillo P, Celli B. Addressing the complexity of chronic obstructive pulmonary disease: from phenotypes and biomarkers to scale-free networks,

systems biology, and P4 medicine. *American journal of respiratory and critical care medicine* 2011;**183**:1129–37. doi:10.1164/rccm.201009-1414PP

15 Khoury MJ, Gwinn ML, Glasgow RE, *et al.* A population approach to precision medicine. *American journal of preventive medicine* 2012;**42**:639–45. doi:10.1016/j.amepre.2012.02.012

16 Chen LL, Blumm N, Christakis NA, *et al.* Cancer metastasis networks and the prediction of progression patterns. *British Journal of Cancer* 2009;**101**:749–58. doi:10.1038/sj.bjc.6605214

17 Frank L. When an Entire Country Is a Cohort. *Science* 2000;**287**:2398–9.

18 Melbye M, Wohlfahrt J, Olsen JH, *et al.* Induced abortion and the risk of breast cancer. *New England journal of medicine* 1997;**336**:81–5.

19 Edgren G, Hjalgrim H, Reilly M, *et al.* Risk of cancer after blood transfusion from donors with subclinical cancer: a retrospective cohort study. *Lancet* 2007;**369**:1724–30. doi:10.1016/S0140-6736(07)60779-X

20 Russo J, Russo IH. Susceptibility of the mammary gland to carcinogenesis. II. Pregnancy interruption as a risk factor in tumor incidence. *American Journal of Pathology* 1980;**100**:497–512.

21 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association : JAMIA* 2013;**20**:117–21. doi:10.1136/amiajnl-2012-001145

22 Lynge E, Sandegaard JL, Rebolj M. The Danish National Patient Register. *Scandinavian journal of public health* 2011;**39**:30–3. doi:10.1177/1403494811401482

23 Malig C. The civil registration system in Denmark. 1996.http://cpr.dk/cpr_artikler/Files/Fil1/4404.pdf (accessed 15 Sep2013).

24 Arkiver S. Kirkebøger. 2013.http://www.sa.dk/content/dk/ao-forside/find_kirkeboger (accessed 15 Sep2013).

25 Thygesen SK, Christiansen CF, Christensen S, *et al.* The predictive value of ICD-10 diagnostic coding used to assess Charlson comorbidity index conditions in the population-based Danish National Registry of Patients. *BMC medical research methodology* 2011;**11**:83. doi:10.1186/1471-2288-11-83

26 Thomsen RW, Lange P, Hellquist B, *et al.* Validity and underrecording of diagnosis of COPD in the Danish National Patient Registry. *Respiratory medicine* 2011;**105**:1063–8. doi:10.1016/j.rmed.2011.01.012

27 Severinsen MT, Kristensen SR, Overvad K, *et al.* Venous thromboembolism discharge diagnoses in the Danish National Patient Registry should be used with caution. *Journal of clinical epidemiology* 2010;**63**:223–8. doi:10.1016/j.jclinepi.2009.03.018

28 Joensen AM, Jensen MK, Overvad K, *et al.* Predictive values of acute coronary syndrome discharge diagnoses differed in the Danish National Patient Registry. *Journal of clinical epidemiology* 2009;**62**:188–94. doi:10.1016/j.jclinepi.2008.03.005

29 Pedersen M, Klarlund M, Jacobsen S, *et al.* Validity of rheumatoid arthritis diagnoses in the Danish National Patient Registry. *European journal of epidemiology* 2004;**19**:1097–103.http://www.ncbi.nlm.nih.gov/pubmed/15678789

30 Mosbech J, Jørgensen J, Madsen M, *et al.* The national patient registry. Evaluation of data quality. *Ugeskrift for læger* 1995;**157**:3741–5.http://europepmc.org/abstract/MED/7631448 (accessed 12 Aug2013).

31 Ankjær-Jensen A, Rosling P, Bilde L. Variable prospective financing in the Danish hospital sector and the development of a Danish case-mix system. *Health Care Management Science* 2006;**9**:259–68. doi:10.1007/s10729-006-9093-1

32 Shaw JE, Sicree R a, Zimmet PZ. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes research and clinical practice* 2010;**87**:4–14. doi:10.1016/j.diabres.2009.10.007

33 Registerdeklaration for Det Nationale Diabetesregister. Sundhedsstyrelsen. http://www.ssi.dk/Sundhedsdataogit/Registre/~/media/Indhold/DK - dansk/Sundhedsdata og it/NSF/Registre/Diabetesregisteret/registerdeklaration det nationale diabetesregister1.ashx

34 Frederiksen K. Diabetes i Danmark – hvad siger Sundhedsstyrelsens registre? Nye tal fra Sundhedsstyrelsen. 1998.http://www.sst.dk/publ/tidsskrifter/nyetal/pdf/98_03_2.pdf

35 Morrey CP, Geller J, Halper M, *et al.* The Neighborhood Auditing Tool: A Hybrid Interface for Auditing the UMLS. *Journal of Biomedical Informatics* 2010;**42**:468–89.

36 Valderas JM, Starfield B, Sibbald B, *et al.* Defining Comorbidity: Implications for Understanding Health and Health Services. *Annals Of Family Medicine* 2009;**7**:357–63. doi:10.1370/afm.983.

37 Fortin M, Bravo G, Hudon C, *et al.* Relationship between multimorbidity and health-related quality of life of patients in primary care. *Quality of life research* 2006;**15**:83–91. doi:10.1007/s11136-005-8661-z

38 Fortin M, Hudon C, Bayliss EA, *et al.* Multimorbidity's many challenges. *BMJ (Clinical research ed)* 2007;**334**:1016–7. doi:10.1136/bmj.39212.467037.BE

39 Finkelstein J, Cha E, Scharf SM. Chronic obstructive pulmonary disease as an independent risk factor for cardiovascular morbidity. *International journal of chronic obstructive pulmonary disease* 2009;**4**:337–49.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2754086&tool=pmcentrez&rendertype=abstract (accessed 18 Jan2013).

40 Curkendall SM, DeLuise C, Jones JK, *et al.* Cardiovascular disease in patients with chronic obstructive pulmonary disease, Saskatchewan Canada cardiovascular disease in COPD patients. *Annals of epidemiology* 2006;**16**:63–70. doi:10.1016/j.annepidem.2005.04.008

41 Salisbury AC, Reid KJ, Spertus JA. Impact of chronic obstructive pulmonary disease on post-myocardial infarction outcomes. *The American journal of cardiology* 2007;**99**:636–41. doi:10.1016/j.amjcard.2006.09.112

42 Charlson ME, Pompei P, Ales KL, *et al.* A New Method of Classifying Prognostic in Longitudinal Studies : Development. *Journal of Chronical Diseases* 1987;**40**:373–83.

43 Nathan DM. Long-Term Complications of Diabetes Mellitus. *New England journal of medicine* 1993;**328**:1676–85. doi:10.1056/NEJM199306103282306

44 Audisio R a. Risk factors for morbidity and mortality after colectomy for colon cancer. *Techniques in coloproctology* 2001;**5**:177–9.http://www.ncbi.nlm.nih.gov/pubmed/11892032

45 Dongen S van. *Graph Clustering by Flow Simulation*. 2000.http://www.library.uu.nl/digiarchief/dip/diss/1895620/full.pdf

46 Bove R, Musallam A, Healy BC, *et al*. No sex-specific difference in disease trajectory in multiple sclerosis patients before and after age 50. *BMC neurology* 2013;**13**:73. doi:10.1186/1471-2377-13-73

47 Lin S, Chen Y, Yang L, *et al*. Pain, fatigue, disturbed sleep and distress comprised a symptom cluster that related to quality of life and functional status of lung cancer surgery patients. *Journal of clinical nursing* 2013;**22**:1281–90. doi:10.1111/jocn.12228

48 Murray SA, Boyd K, Kendall M, *et al*. Dying of lung cancer or cardiac failure: prospective qualitative interview study of patients and their carers in the community. *BMJ* 2002;**325**.

49 Murray SA, Kendall M, Boyd K, *et al*. Clinical review Illness trajectories and palliative care. *BMJ* 2005;**330**:1007–11.

50 Churchill GA. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* 1989;**51**:79–94.

51 White J V., Stultz CM, Smith TF. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Mathematical Biosciences* 1994;**119**:35–75.

52 Ohlsson M, Peterson C, Dictor M. Using Hidden Markov Models to Characterize Disease Trajectories. *4th International Conference on "Neural Networks and Expert Syststems in Medicine and Healthcare"* 2001;**11**:6–8.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.186&rep=rep1&type=ps (accessed 14 Sep2010).

53 Savage DC. Microbial Ecology of the Gastrointestinal Tract. *Annual Review of Microbiology* 1977;**31**:107–33. doi:10.1146/annurev.mi.31.100177.000543

54 Bäckhed F, Ding H, Wang T, *et al*. The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences of the United States of America* 2004;**101**:15718–23. doi:10.1073/pnas.0407076101

55 Everard A, Cani PD. Diabetes, obesity and gut microbiota. *Best practice & research Clinical gastroenterology* 2013;**27**:73–83. doi:10.1016/j.bpg.2013.03.007

56 Mendelsohn AR, Larrick JW. Dietary modification of the microbiome affects risk for cardiovascular disease. *Rejuvenation Research* 2013;**16**:241–4. doi:10.1089/rej.2013.1447

57 Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 1958;**53**:457–81.

58 Collett D. *Modelling Survival Data in Medical Research*. 2nd ed. Chapman & Hall 2003.

59 Cox DR, Society S, Methodological SB. Regression Models and Life-Tables. *Journal of the Royal Statistical Society* 1972;**34**:187–220.

60 Tang WH, Wang Z, Levison BS, *et al*. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *New England journal of medicine* 2013;**368**:1575–84.

61 Cani PD. Gut microbiota and obesity: lessons from the microbiome. *Briefings in Functional Genomics* 2013;**12**:381–7.

62 Cox LM, Blaser MJ. Pathways in microbe-induced obesity. *Cell Metabolism* 2013;**17**:883–94.

63 Nicholson JK, Holmes E, Kinross J, *et al.* Host-gut microbiota metabolic interactions. *Science* 2012;**336**:1262–7.

64 Tremaroli V, Backhed F. Functional interactions between the gut microbiota and host metabolism. *Nature* 2012;**489**:242–9.

65 Loscalzo J. Gut microbiota, the genome, and diet in atherogenesis. *New England journal of medicine* 2013;**368**:1647–9.

66 Cani PD, Osto M, Geurts L, *et al.* Involvement of gut microbiota in the development of low-grade inflammation and type 2 diabetes associated with obesity. *Gut Microbes* 2012;**3**.

67 Le CE, Nielsen T, Qin J, *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* 2013;**500**:541–6.

68 Burcelin R, Serino M, Chabo C, *et al.* Gut microbiota and diabetes: from pathogenesis to therapeutic perspective. *Acta Diabetologica* 2011;**48**:257–73.

69 Carvalho BM, Saad MJ. Influence of gut microbiota on subclinical inflammation and insulin resistance. *Mediators of Inflammation* 2013;**2013**:986734.

70 Mendelsohn AR, Larrick JW. Dietary modification of the microbiome affects risk for cardiovascular disease. *Rejuvenation Research* 2013;**16**:241–4.

71 Cani PD, Amar J, Iglesias MA, *et al.* Metabolic endotoxemia initiates obesity and insulin resistance. *Diabetes* 2007;**56**:1761–72.

72 Ahmadi-Abhari S, Luben RN, Wareham NJ, *et al.* Seventeen year risk of all-cause and cause-specific mortality associated with C-reactive protein, fibrinogen and leukocyte count in men and women: the EPIC-Norfolk study. *European Journal of Epidemiology* 2013.

73 Goh K, Cusick M, Valle D, *et al.* The human disease network. *PNAS* 2007;**104**:8685–90.http://www.pnas.org/content/104/21/8685.full (accessed 27 Aug2010).

74 Szklarczyk D, Franceschini A, Kuhn M, *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* 2011;**39**:D561–8. doi:10.1093/nar/gkq973

75 Lage K, Hansen NT, Karlberg EO, *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences of the United States of America* 2008;**105**:20870–5. doi:10.1073/pnas.0810772105

76 Chatr-Aryamontri A, Breitkreutz B-J, Heinicke S, *et al.* The BioGRID interaction database: 2013 update. *Nucleic acids research* 2013;**41**:D816–23. doi:10.1093/nar/gks1158

77 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature reviews Genetics* 2012;**13**:395–405. doi:10.1038/nrg3208

78 Hamosh A, Scott AF, Amberger JS, *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* 2005;**33**:D514–7. doi:10.1093/nar/gki033

79 Consortium TU. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research* 2013;**41**:D43–7. doi:10.1093/nar/gks1068

80 Hazuda HP, Haffner SM, Stern MP, *et al.* Effects of Acculturation and Socioeconomic Status on Obesity and Diabetes in Mexican Americans - The San Antonio Heart Study. *American journal of epidemiology* 1988;**128**:1289–301.

81 Bradley CJ, Given CW, Roberts C. Race, socioeconomic status, and breast cancer treatment and survival. *Journal of the National Cancer Institute* 2002;**94**:490–6.http://www.ncbi.nlm.nih.gov/pubmed/11929949

82 Behrman RE, Benner JS, Brown JS, *et al.* Developing the Sentinel System - a national resource for evidence development. *The New England journal of medicine* 2011;**364**:498–9. doi:10.1056/NEJMp1014427

83 Preciosa M. Coloma, Schuemie MJ, Trifirò G, *et al.* Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiology and Drug Safety* 2011;**20**:1–11. doi:10.1002/pds

84 Eriksson R, Jensen PB, Frankild S, *et al.* Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. *Journal of the American Medical Informatics Association : JAMIA* 2013;**20**:947–53. doi:10.1136/amiajnl-2013-001708

Supplementary materials

Supplementary table 1: List of ICD-10 diagnoses and clusters (work in progress 1). The listing shows the result of the clustering and meta-clustering of ICD-10 diagnoses. Clusters are listed under the meta-clusters they are part of. The meta-cluster numbers refers to the numbers shown in figure 3.3. The clusters are numbered from the largest cluster to the smallest.

Meta Cluster 1

Cluster 1
A00 Cholera
A08 Viral and other specified intestinal infections
A37 Whooping cough
B01 Varicella [chickenpox]
B05 Measles
B34 Viral infection of unspecified site
B43 Chromomycosis and phaeomycotic abscess
E40 Kwashiorkor
E71 Disorders of branched-chain amino-acid metabolism and fatty-acid metabolism
E74 Other disorders of carbohydrate metabolism
E75 Disorders of sphingolipid metabolism and other lipid storage disorders
E77 Disorders of glycoprotein metabolism
E86 Volume depletion
F44 Dissociative [conversion] disorders
F51 Nonorganic sleep disorders
F72 Severe mental retardation
F73 Profound mental retardation
G08 Intracranial and intraspinal phlebitis and thrombophlebitis
G11 Hereditary ataxia
G13 Systemic atrophies primarily affecting central nervous system in diseases classified elsewhere
G24 Dystonia
G31 Other degenerative diseases of nervous system, not elsewhere classified
G40 Epilepsy
G41 Status epilepticus
G47 Sleep disorders
G54 Nerve root and plexus disorders
G55 Nerve root and plexus compressions in diseases classified elsewhere
G80 Cerebral palsy
G81 Hemiplegia
G82 Paraplegia and tetraplegia
G83 Other paralytic syndromes
G90 Disorders of autonomic nervous system
G92 Toxic encephalopathy
G95 Other diseases of spinal cord
G96 Other disorders of central nervous system
G98 Other disorders of nervous system, not elsewhere classified
G99 Other disorders of nervous system in diseases classified elsewhere
H47 Other disorders of optic [2nd] nerve and visual pathways
H48 Disorders of optic [2nd] nerve and visual pathways in diseases classified elsewhere
I63 Cerebral infarction
J00 Acute nasopharyngitis [common cold]
J02 Acute pharyngitis
J03 Acute tonsillitis
J06 Acute upper respiratory infections of multiple and unspecified sites
J10 Influenza due to other identified influenza virus
J12 Viral pneumonia, not elsewhere classified
J15 Bacterial pneumonia, not elsewhere classified
J17 Pneumonia in diseases classified elsewhere
J18 Pneumonia, organism unspecified
J22 Unspecified acute lower respiratory infection
J36 Peritonsillar abscess
J39 Other diseases of upper respiratory tract
J40 Bronchitis, not specified as acute or chronic
J69 Pneumonitis due to solids and liquids
J96 Respiratory failure, not elsewhere classified
J99 Respiratory disorders in diseases classified elsewhere
K59 Other functional intestinal disorders
L60 Nail disorders
L62 Nail disorders in diseases classified elsewhere
L84 Corns and callosities
L89 Decubitus ulcer
M21 Other acquired deformities of limbs
M40 Kyphosis and lordosis
M41 Scoliosis
M42 Spinal osteochondrosis
M47 Spondylosis
M48 Other spondylopathies
M50 Cervical disc disorders
M51 Other intervertebral disc disorders
M53 Other dorsopathies, not elsewhere classified
M54 Dorsalgia
M62 Other disorders of muscle
M84 Disorders of continuity of bone
M89 Other disorders of bone
M96 Postprocedural musculoskeletal disorders, not elsewhere classified
M99 Biomechanical lesions, not elsewhere classified
Q02 Microcephaly
Q04 Other congenital malformations of brain
Q07 Other congenital malformations of nervous system
Q40 Other congenital malformations of upper alimentary tract
Q65 Congenital deformities of hip
Q66 Congenital deformities of feet
Q80 Congenital ichthyosis
Q81 Epidermolysis bullosa

Cluster 4
A07 Other protozoal intestinal diseases
A09 Diarrhoea and gastroenteritis of presumed infectious origin
A51 Early syphilis
A52 Late syphilis
A53 Other and unspecified syphilis
A55 Chlamydial lymphogranuloma (venereum)
A60 Anogenital herpesviral [herpes simplex] infection
A63 Other predominantly sexually transmitted diseases, not elsewhere classified

A64 Unspecified sexually transmitted disease
A81 Atypical virus infections of central nervous system
B00 Herpesviral [herpes simplex] infections
B02 Zoster [herpes zoster]
B07 Viral warts
B08 Other viral infections characterized by skin and mucous membrane lesions, not elsewhere classified
B09 Unspecified viral infection characterized by skin and mucous membrane lesions
B20 Human immunodeficiency virus [HIV] disease resulting in infectious and parasitic diseases
B21 Human immunodeficiency virus [HIV] disease resulting in malignant neoplasms
B22 Human immunodeficiency virus [HIV] disease resulting in other specified diseases
B23 Human immunodeficiency virus [HIV] disease resulting in other conditions
B24 Unspecified human immunodeficiency virus [HIV] disease
B33 Other viral diseases, not elsewhere classified
B35 Dermatophytosis
B36 Other superficial mycoses
B37 Candidiasis
B39 Histoplasmosis
B45 Cryptococcosis
B55 Leishmaniasis
B64 Unspecified protozoal disease
B68 Taeniasis
B71 Other cestode infections
B77 Ascariasis
B81 Other intestinal helminthiases, not elsewhere classified
B86 Scabies
C46 Kaposi's sarcoma
D64 Other anaemias
J13 Pneumonia due to Streptococcus pneumoniae
J14 Pneumonia due to Haemophilus influenzae
J16 Pneumonia due to other infectious organisms, not elsewhere classified
J20 Acute bronchitis
L21 Seborrhoeic dermatitis
L72 Follicular cysts of skin and subcutaneous tissue
L73 Other follicular disorders
L75 Apocrine sweat disorders

Cluster 5
A44 Bartonellosis
A46 Erysipelas
A49 Bacterial infection of unspecified site
A82 Rabies
B99 Other and unspecified infectious diseases
E54 Ascorbic acid deficiency
E87 Other disorders of fluid, electrolyte and acid-base balance
F05 Delirium, not induced by alcohol and other psychoactive substances
F10 Mental and behavioural disorders due to use of alcohol
F11 Mental and behavioural disorders due to use of opioids
F12 Mental and behavioural disorders due to use of cannabinoids
F13 Mental and behavioural disorders due to use of sedatives or hypnotics
F14 Mental and behavioural disorders due to use of cocaine
F15 Mental and behavioural disorders due to use of other stimulants, including caffeine
F16 Mental and behavioural disorders due to use of hallucinogens
F17 Mental and behavioural disorders due to use of tobacco
F18 Mental and behavioural disorders due to use of volatile solvents
F19 Mental and behavioural disorders due to multiple drug use and use of other psychoactive substances
F40 Phobic anxiety disorders
F55 Abuse of non-dependence-producing substances
F62 Enduring personality changes, not attributable to brain damage and disease
F63 Habit and impulse disorders
F65 Disorders of sexual preference
I72 Other aneurysm
I80 Phlebitis and thrombophlebitis
I82 Other venous embolism and thrombosis
I87 Other disorders of veins
J86 Pyothorax
K29 Gastritis and duodenitis
K30 Dyspepsia
L02 Cutaneous abscess, furuncle and carbuncle
L03 Cellulitis
L08 Other local infections of skin and subcutaneous tissue
M80 Osteoporosis with pathological fracture
M86 Osteomyelitis

Cluster 8
A23 Brucellosis
B67 Echinococcosis
D50 Iron deficiency anaemia
D53 Other nutritional anaemias
D62 Acute posthaemorrhagic anaemia
E42 Marasmic kwashiorkor
E43 Unspecified severe protein-energy malnutrition
E46 Unspecified protein-energy malnutrition
E51 Thiamine deficiency
E60 Dietary zinc deficiency
E85 Amyloidosis
I43 Cardiomyopathy in diseases classified elsewhere
I81 Portal vein thrombosis
I85 Oesophageal varices
I86 Varicose veins of other sites
I98 Other disorders of circulatory system in diseases classified elsewhere
K20 Oesophagitis
K25 Gastric ulcer
K26 Duodenal ulcer
K27 Peptic ulcer, site unspecified
K28 Gastrojejunal ulcer
K70 Alcoholic liver disease
K71 Toxic liver disease
K72 Hepatic failure, not elsewhere classified

K74 Fibrosis and cirrhosis of liver
K75 Other inflammatory liver diseases
K76 Other diseases of liver
K77 Liver disorders in diseases classified elsewhere
K92 Other diseases of digestive system
Q44 Congenital malformations of gallbladder, bile ducts and liver

Cluster 18
B66 Other fluke infections
E10 Insulin-dependent diabetes mellitus
E11 Non-insulin-dependent diabetes mellitus
E12 Malnutrition-related diabetes mellitus
E13 Other specified diabetes mellitus
E14 Unspecified diabetes mellitus
E15 Nondiabetic hypoglycaemic coma
E16 Other disorders of pancreatic internal secretion
E66 Obesity
E90 Nutritional and metabolic disorders in diseases classified elsewhere
G59 Mononeuropathy in diseases classified elsewhere
G63 Polyneuropathy in diseases classified elsewhere
H28 Cataract and other disorders of lens in diseases classified elsewhere
H36 Retinal disorders in diseases classified elsewhere
L14 Bullous disorders in diseases classified elsewhere
L97 Ulcer of lower limb, not elsewhere classified
M14 Arthropathies in other diseases classified elsewhere

Cluster 29
B80 Enterobiasis
B82 Unspecified intestinal parasitism
K50 Crohn's disease [regional enteritis]
K51 Ulcerative colitis
K56 Paralytic ileus and intestinal obstruction without hernia
K57 Diverticular disease of intestine
K62 Other diseases of anus and rectum
K63 Other diseases of intestine
K66 Other disorders of peritoneum
K91 Postprocedural disorders of digestive system, not elsewhere classified

Cluster 32
G12 Spinal muscular atrophy and related syndromes
G70 Myasthenia gravis and other myoneural disorders
G71 Primary disorders of muscles
G72 Other myopathies
G73 Disorders of myoneural junction and muscle in diseases classified elsewhere
M33 Dermatopolymyositis
M36 Systemic disorders of connective tissue in diseases classified elsewhere
M60 Myositis
M63 Disorders of muscle in diseases classified elsewhere

Cluster 35
B15 Acute hepatitis A
B16 Acute hepatitis B
B17 Other acute viral hepatitis
B18 Chronic viral hepatitis
B19 Unspecified viral hepatitis
B27 Infectious mononucleosis
E80 Disorders of porphyrin and bilirubin metabolism
K73 Chronic hepatitis, not elsewhere classified

Cluster 43
J61 Pneumoconiosis due to asbestos and other mineral fibres
J62 Pneumoconiosis due to dust containing silica
J64 Unspecified pneumoconiosis
J68 Respiratory conditions due to inhalation of chemicals, gases, fumes and vapours
J70 Respiratory conditions due to other external agents
J92 Pleural plaque

Cluster 49
D74 Methaemoglobinaemia
K42 Umbilical hernia
K43 Ventral hernia
K45 Other abdominal hernia
K46 Unspecified abdominal hernia
K67 Disorders of peritoneum in infectious diseases classified elsewhere

Cluster 60
I30 Acute pericarditis
I31 Other diseases of pericardium
I32 Pericarditis in diseases classified elsewhere
I40 Acute myocarditis

Cluster 64
D80 Immunodeficiency with predominantly antibody defects
D81 Combined immunodeficiencies
D83 Common variable immunodeficiency
D84 Other immunodeficiencies

Cluster 68
D55 Anaemia due to enzyme disorders
D56 Thalassaemia
D57 Sickle-cell disorders
D58 Other hereditary haemolytic anaemias

Cluster 70
I84 Haemorrhoids
K60 Fissure and fistula of anal and rectal regions
K61 Abscess of anal and rectal regions
L05 Pilonidal cyst

Cluster 85
L10 Pemphigus
L12 Pemphigoid
L13 Other bullous disorders

Cluster 86
C17 Malignant neoplasm of small intestine
E34 Other endocrine disorders
Q97 Other sex chromosome abnormalities, female phenotype, not elsewhere classified

Cluster 109
D68 Other coagulation defects
E72 Other disorders of amino-acid metabolism

Cluster 111
D66 Hereditary factor VIII deficiency
D67 Hereditary factor IX deficiency

Cluster 116
A80 Acute poliomyelitis
B91 Sequelae of poliomyelitis

Meta Cluster 2

Cluster 22
C51 Malignant neoplasm of vulva
C52 Malignant neoplasm of vagina
C60 Malignant neoplasm of penis
D01 Carcinoma in situ of other and unspecified digestive organs
D07 Carcinoma in situ of other and unspecified genital organs
D25 Leiomyoma of uterus
D26 Other benign neoplasms of uterus
D28 Benign neoplasm of other and unspecified female genital organs
M88 Paget's disease of bone [osteitis deformans]
N75 Diseases of Bartholin's gland
N77 Vulvovaginal ulceration and inflammation in diseases classified elsewhere
N84 Polyp of female genital tract
N85 Other noninflammatory disorders of uterus, except cervix
N90 Other noninflammatory disorders of vulva and perineum
N92 Excessive, frequent and irregular menstruation

Cluster 37
B88 Other infestations
D24 Benign neoplasm of breast
N60 Benign mammary dysplasia
N61 Inflammatory disorders of breast
N62 Hypertrophy of breast
N63 Unspecified lump in breast
N64 Other disorders of breast
Q83 Congenital malformations of breast

Cluster 39
N81 Female genital prolapse
N89 Other noninflammatory disorders of vagina
N94 Pain and other conditions associated with female genital organs and menstrual cycle
N96 Habitual aborter
N97 Female infertility
Q51 Congenital malformations of uterus and cervix
Q52 Other congenital malformations of female genitalia

Cluster 41
B87 Myiasis
C53 Malignant neoplasm of cervix uteri
C57 Malignant neoplasm of other and unspecified female genital organs
D06 Carcinoma in situ of cervix uteri
N82 Fistulae involving female genital tract
N87 Dysplasia of cervix uteri
N88 Other noninflammatory disorders of cervix uteri

Cluster 42
M22 Disorders of patella
M23 Internal derangement of knee
M76 Enthesopathies of lower limb, excluding foot
M77 Other enthesopathies
M92 Other juvenile osteochondrosis
M93 Other osteochondropathies
M94 Other disorders of cartilage

Cluster 69
F64 Gender identity disorders
L67 Hair colour and hair shaft abnormalities
L68 Hypertrichosis
Q84 Other congenital malformations of integument

Cluster 72
N70 Salpingitis and oophoritis
N71 Inflammatory disease of uterus, except cervix
N73 Other female pelvic inflammatory diseases
N74 Female pelvic inflammatory disorders in diseases classified elsewhere

Cluster 73
K35 Acute appendicitis
K36 Other appendicitis
K37 Unspecified appendicitis
K38 Other diseases of appendix

Cluster 87
C54 Malignant neoplasm of corpus uteri
C55 Malignant neoplasm of uterus, part unspecified
N95 Menopausal and other perimenopausal disorders

Cluster 93
E28 Ovarian dysfunction
L83 Acanthosis nigricans
N98 Complications associated with artificial fertilization

Cluster 94
A54 Gonococcal infection
A56 Other sexually transmitted chlamydial diseases
A74 Other diseases caused by chlamydiae

Cluster 95
N72 Inflammatory disease of cervix uteri
N86 Erosion and ectropion of cervix uteri

N93 Other abnormal uterine and vaginal bleeding

Cluster 113
F52 Sexual dysfunction, not caused by organic disorder or disease
F66 Psychological and behavioural disorders associated with sexual development and orientation

Cluster 114
A59 Trichomoniasis
N76 Other inflammation of vagina and vulva

Meta Cluster 3

Cluster 6
A78 Q fever
B85 Pediculosis and phthiriasis
I83 Varicose veins of lower extremities
L00 Staphylococcal scalded skin syndrome
L01 Impetigo
L20 Atopic dermatitis
L22 Diaper [napkin] dermatitis
L23 Allergic contact dermatitis
L24 Irritant contact dermatitis
L25 Unspecified contact dermatitis
L26 Exfoliative dermatitis
L27 Dermatitis due to substances taken internally
L28 Lichen simplex chronicus and prurigo
L29 Pruritus
L30 Other dermatitis
L42 Pityriasis rosea
L44 Other papulosquamous disorders
L45 Papulosquamous disorders in diseases classified elsewhere
L50 Urticaria
L51 Erythema multiforme
L53 Other erythematous conditions
L54 Erythema in diseases classified elsewhere
L55 Sunburn
L56 Other acute skin changes due to ultraviolet radiation
L74 Eccrine sweat disorders
L80 Vitiligo
L85 Other epidermal thickening
L86 Keratoderma in diseases classified elsewhere
L88 Pyoderma gangrenosum
L92 Granulomatous disorders of skin and subcutaneous tissue
L98 Other disorders of skin and subcutaneous tissue, not elsewhere classified
L99 Other disorders of skin and subcutaneous tissue in diseases classified elsewhere

Cluster 10
A03 Shigellosis
C74 Malignant neoplasm of adrenal gland
C75 Malignant neoplasm of other endocrine glands and related structures
D35 Benign neoplasm of other and unspecified endocrine glands
D44 Neoplasm of uncertain or unknown behaviour of endocrine glands
E20 Hypoparathyroidism
E21 Hyperparathyroidism and other disorders of parathyroid gland
E22 Hyperfunction of pituitary gland
E23 Hypofunction and other disorders of pituitary gland
E24 Cushing's syndrome
E25 Adrenogenital disorders
E26 Hyperaldosteronism
E27 Other disorders of adrenal gland
E29 Testicular dysfunction
E30 Disorders of puberty, not elsewhere classified
E31 Polyglandular dysfunction
E35 Disorders of endocrine glands in diseases classified elsewhere
E44 Protein-energy malnutrition of moderate and mild degree
E83 Disorders of mineral metabolism
E88 Other metabolic disorders
E89 Postprocedural endocrine and metabolic disorders, not elsewhere classified
M81 Osteoporosis without pathological fracture
M82 Osteoporosis in diseases classified elsewhere
M85 Other disorders of bone density and structure
N46 Male infertility
N91 Absent, scanty and rare menstruation
Q98 Other sex chromosome abnormalities, male phenotype, not elsewhere classified

Cluster 11
I00 Rheumatic fever without mention of heart involvement
M00 Pyogenic arthritis
M01 Direct infections of joint in infectious and parasitic diseases classified elsewhere
M02 Reactive arthropathies
M03 Postinfective and reactive arthropathies in diseases classified elsewhere
M05 Seropositive rheumatoid arthritis
M06 Other rheumatoid arthritis
M08 Juvenile arthritis
M09 Juvenile arthritis in diseases classified elsewhere
M11 Other crystal arthropathies
M12 Other specific arthropathies
M13 Other arthritis
M15 Polyarthrosis
M17 Gonarthrosis [arthrosis of knee]
M18 Arthrosis of first carpometacarpal joint
M19 Other arthrosis
M24 Other specific joint derangements
M25 Other joint disorders, not elsewhere classified
M61 Calcification and ossification of muscle
M65 Synovitis and tenosynovitis
M66 Spontaneous rupture of synovium and tendon
M67 Other disorders of synovium and tendon
M68 Disorders of synovium and tendon in diseases classified elsewhere

M70 Soft tissue disorders related to use, overuse and pressure
M71 Other bursopathies
M75 Shoulder lesions
M79 Other soft tissue disorders, not elsewhere classified

Cluster 40
G56 Mononeuropathies of upper limb
G57 Mononeuropathies of lower limb
G58 Other mononeuropathies
G60 Hereditary and idiopathic neuropathy
G61 Inflammatory polyneuropathy
G62 Other polyneuropathies
G64 Other disorders of peripheral nervous system

Cluster 47
A50 Congenital syphilis
K21 Gastro-oesophageal reflux disease
K22 Other diseases of oesophagus
K23 Disorders of oesophagus in diseases classified elsewhere
K44 Diaphragmatic hernia
Q39 Congenital malformations of oesophagus

Cluster 48
B30 Viral conjunctivitis
H13 Disorders of conjunctiva in diseases classified elsewhere
H16 Keratitis
H17 Corneal scars and opacities
H18 Other disorders of cornea
H19 Disorders of sclera and cornea in diseases classified elsewhere

Cluster 53
B26 Mumps
C07 Malignant neoplasm of parotid gland
D11 Benign neoplasm of major salivary glands
K11 Diseases of salivary glands
L11 Other acantholytic disorders

Cluster 58
L93 Lupus erythematosus
L95 Vasculitis limited to skin, not elsewhere classified
M32 Systemic lupus erythematosus
M35 Other systemic involvement of connective tissue

Cluster 77
L40 Psoriasis
M07 Psoriatic and enteropathic arthropathies
M73 Soft tissue disorders in diseases classified elsewhere

Cluster 88
M20 Acquired deformities of fingers and toes
Q69 Polydactyly
Q70 Syndactyly

Cluster 90
E73 Lactose intolerance
K58 Irritable bowel syndrome
K90 Intestinal malabsorption

Cluster 97
L94 Other localized connective tissue disorders
M34 Systemic sclerosis

Meta Cluster 4

Cluster 13
B06 Rubella [German measles]
B40 Blastomycosis
N10 Acute tubulo-interstitial nephritis
N11 Chronic tubulo-interstitial nephritis
N12 Tubulo-interstitial nephritis, not specified as acute or chronic
N13 Obstructive and reflux uropathy
N20 Calculus of kidney and ureter
N21 Calculus of lower urinary tract
N22 Calculus of urinary tract in diseases classified elsewhere
N23 Unspecified renal colic
N28 Other disorders of kidney and ureter, not elsewhere classified
N29 Other disorders of kidney and ureter in diseases classified elsewhere
N30 Cystitis
N32 Other disorders of bladder
N33 Bladder disorders in diseases classified elsewhere
N35 Urethral stricture
N39 Other disorders of urinary system
N40 Hyperplasia of prostate
N42 Other disorders of prostate
N99 Postprocedural disorders of genitourinary system, not elsewhere classified
Q42 Congenital absence, atresia and stenosis of large intestine
Q60 Renal agenesis and other reduction defects of kidney
Q61 Cystic kidney disease
Q62 Congenital obstructive defects of renal pelvis and congenital malformations of ureter
Q63 Other congenital malformations of kidney
Q64 Other congenital malformations of urinary system

Cluster 14
C40 Malignant neoplasm of bone and articular cartilage of limbs
C41 Malignant neoplasm of bone and articular cartilage of other and unspecified sites
C47 Malignant neoplasm of peripheral nerves and autonomic nervous system
C48 Malignant neoplasm of retroperitoneum and peritoneum
C49 Malignant neoplasm of other connective and soft tissue
C56 Malignant neoplasm of ovary
C58 Malignant neoplasm of placenta
C76 Malignant neoplasm of other and ill-defined sites
C78 Secondary malignant neoplasm of respiratory and digestive organs
C79 Secondary malignant neoplasm of other sites

C80 Malignant neoplasm without specification of site
C97 Malignant neoplasms of independent (primary) multiple sites
D17 Benign lipomatous neoplasm
D20 Benign neoplasm of soft tissue of retroperitoneum and peritoneum
D21 Other benign neoplasms of connective and other soft tissue
D27 Benign neoplasm of ovary
D36 Benign neoplasm of other and unspecified sites
D39 Neoplasm of uncertain or unknown behaviour of female genital organs
D48 Neoplasm of uncertain or unknown behaviour of other and unspecified sites
D63 Anaemia in chronic diseases classified elsewhere
D71 Functional disorders of polymorphonuclear neutrophils
M90 Osteopathies in diseases classified elsewhere
N80 Endometriosis
N83 Noninflammatory disorders of ovary, fallopian tube and broad ligament
Q50 Congenital malformations of ovaries, fallopian tubes and broad ligaments

Cluster 25
C34 Malignant neoplasm of bronchus and lung
C37 Malignant neoplasm of thymus
C38 Malignant neoplasm of heart, mediastinum and pleura
C39 Malignant neoplasm of other and ill-defined sites in the respiratory system and intrathoracic organs
C45 Mesothelioma
D15 Benign neoplasm of other and unspecified intrathoracic organs
D19 Benign neoplasm of mesothelial tissue
D38 Neoplasm of uncertain or unknown behaviour of middle ear and respiratory and intrathoracic organs
E32 Diseases of thymus
J90 Pleural effusion, not elsewhere classified
J91 Pleural effusion in conditions classified elsewhere
J93 Pneumothorax
J94 Other pleural conditions

Cluster 44
C64 Malignant neoplasm of kidney, except renal pelvis
C65 Malignant neoplasm of renal pelvis
C66 Malignant neoplasm of ureter
C68 Malignant neoplasm of other and unspecified urinary organs
D30 Benign neoplasm of urinary organs
D41 Neoplasm of uncertain or unknown behaviour of urinary organs

Cluster 50
C61 Malignant neoplasm of prostate
C62 Malignant neoplasm of testis
C63 Malignant neoplasm of other and unspecified male genital organs
D29 Benign neoplasm of male genital organs
D40 Neoplasm of uncertain or unknown behaviour of male genital organs
N43 Hydrocele and spermatocele

Cluster 51
C18 Malignant neoplasm of colon
C19 Malignant neoplasm of rectosigmoid junction
C20 Malignant neoplasm of rectum
C21 Malignant neoplasm of anus and anal canal
D12 Benign neoplasm of colon, rectum, anus and anal canal
D37 Neoplasm of uncertain or unknown behaviour of oral cavity and digestive organs

Cluster 65
N34 Urethritis and urethral syndrome
N36 Other disorders of urethra
N37 Urethral disorders in diseases classified elsewhere
N41 Inflammatory diseases of prostate

Cluster 96
A58 Granuloma inguinale
K40 Inguinal hernia
K41 Femoral hernia

Cluster 101
C15 Malignant neoplasm of oesophagus
C16 Malignant neoplasm of stomach

Cluster 105
C67 Malignant neoplasm of bladder
D09 Carcinoma in situ of other and unspecified sites

Meta Cluster 5

Cluster 9
F06 Other mental disorders due to brain damage and dysfunction and to physical disease
F09 Unspecified organic or symptomatic mental disorder
F20 Schizophrenia
F21 Schizotypal disorder
F22 Persistent delusional disorders
F23 Acute and transient psychotic disorders
F24 Induced delusional disorder
F25 Schizoaffective disorders
F28 Other nonorganic psychotic disorders
F29 Unspecified nonorganic psychosis
F30 Manic episode
F31 Bipolar affective disorder
F32 Depressive episode
F33 Recurrent depressive disorder
F34 Persistent mood [affective] disorders

F38 Other mood [affective] disorders
F39 Unspecified mood [affective] disorder
F41 Other anxiety disorders
F43 Reaction to severe stress, and adjustment disorders
F45 Somatoform disorders
F48 Other neurotic disorders
F50 Eating disorders
F53 Mental and behavioural disorders associated with the puerperium, not elsewhere classified
F59 Unspecified behavioural syndromes associated with physiological disturbances and physical factors
F60 Specific personality disorders
F61 Mixed and other personality disorders
F68 Other disorders of adult personality and behaviour
F69 Unspecified disorder of adult personality and behaviour
F92 Mixed disorders of conduct and emotions
F99 Mental disorder, not otherwise specified

Cluster 12
E67 Other hyperalimentation
F01 Vascular dementia
F07 Personality and behavioural disorders due to brain disease, damage and dysfunction
F54 Psychological and behavioural factors associated with disorders or diseases classified elsewhere
G43 Migraine
G44 Other headache syndromes
G45 Transient cerebral ischaemic attacks and related syndromes
G46 Vascular syndromes of brain in cerebrovascular diseases
G91 Hydrocephalus
G93 Other disorders of brain
G94 Other disorders of brain in diseases classified elsewhere
G97 Postprocedural disorders of nervous system, not elsewhere classified
H82 Vertiginous syndromes in diseases classified elsewhere
I60 Subarachnoid haemorrhage
I61 Intracerebral haemorrhage
I62 Other nontraumatic intracranial haemorrhage
I64 Stroke, not specified as haemorrhage or infarction
I65 Occlusion and stenosis of precerebral arteries, not resulting in cerebral infarction
I66 Occlusion and stenosis of cerebral arteries, not resulting in cerebral infarction
I67 Other cerebrovascular diseases
I68 Cerebrovascular disorders in diseases classified elsewhere
I69 Sequelae of cerebrovascular disease
Q00 Anencephaly and similar malformations
Q01 Encephalocele
Q03 Congenital hydrocephalus
Q75 Other congenital malformations of skull and face bones

Cluster 27
A15 Respiratory tuberculosis, bacteriologically and histologically confirmed
A16 Respiratory tuberculosis, not confirmed bacteriologically or histologically
A17 Tuberculosis of nervous system
A18 Tuberculosis of other organs
A19 Miliary tuberculosis
A31 Infection due to other mycobacteria
B69 Cysticercosis
B79 Trichuriasis
B90 Sequelae of tuberculosis
J85 Abscess of lung and mediastinum
K93 Disorders of other digestive organs in diseases classified elsewhere
M83 Adult osteomalacia

Cluster 34
F02 Dementia in other diseases classified elsewhere
G10 Huntington's disease
G20 Parkinson's disease
G21 Secondary parkinsonism
G22 Parkinsonism in diseases classified elsewhere
G23 Other degenerative diseases of basal ganglia
G25 Other extrapyramidal and movement disorders
G26 Extrapyramidal and movement disorders in diseases classified elsewhere
I02 Rheumatic chorea

Cluster 56
E53 Deficiency of other B group vitamins
E55 Vitamin D deficiency
E56 Other vitamin deficiencies
E58 Dietary calcium deficiency
E61 Deficiency of other nutrient elements

Cluster 71
F00 Dementia in Alzheimer's disease
F03 Unspecified dementia
F04 Organic amnesic syndrome, not induced by alcohol and other psychoactive substances
G30 Alzheimer's disease

Cluster 108
D51 Vitamin B12 deficiency anaemia
D52 Folate deficiency anaemia

Cluster 112
M16 Coxarthrosis [arthrosis of hip]
M91 Juvenile osteochondrosis of hip and pelvis

Cluster 115
A88 Other viral infections of central nervous system, not elsewhere classified
H81 Disorders of vestibular function

Meta Cluster 6

Cluster 2
A42 Actinomycosis
C00 Malignant neoplasm of lip
C01 Malignant neoplasm of base of tongue
C02 Malignant neoplasm of other and unspecified parts of tongue
C03 Malignant neoplasm of gum
C04 Malignant neoplasm of floor of mouth
C05 Malignant neoplasm of palate
C06 Malignant neoplasm of other and unspecified parts of mouth
C08 Malignant neoplasm of other and unspecified major salivary glands
C09 Malignant neoplasm of tonsil
C10 Malignant neoplasm of oropharynx
C11 Malignant neoplasm of nasopharynx
C12 Malignant neoplasm of piriform sinus
C13 Malignant neoplasm of hypopharynx
C14 Malignant neoplasm of other and ill-defined sites in the lip, oral cavity and pharynx
C32 Malignant neoplasm of larynx
C33 Malignant neoplasm of trachea
C50 Malignant neoplasm of breast
C77 Secondary and unspecified malignant neoplasm of lymph nodes
D00 Carcinoma in situ of oral cavity, oesophagus and stomach
D02 Carcinoma in situ of middle ear and respiratory system
D05 Carcinoma in situ of breast
D10 Benign neoplasm of mouth and pharynx
D14 Benign neoplasm of middle ear and respiratory system
E41 Nutritional marasmus
E47
E63 Other nutritional deficiencies
E64 Sequelae of malnutrition and other nutritional deficiencies
G50 Disorders of trigeminal nerve
I89 Other noninfective disorders of lymphatic vessels and lymph nodes
J04 Acute laryngitis and tracheitis
J05 Acute obstructive laryngitis [croup] and epiglottitis
J35 Chronic diseases of tonsils and adenoids
J37 Chronic laryngitis and laryngotracheitis
J38 Diseases of vocal cords and larynx, not elsewhere classified
J95 Postprocedural respiratory disorders, not elsewhere classified
K00 Disorders of tooth development and eruption
K01 Embedded and impacted teeth
K02 Dental caries
K03 Other diseases of hard tissues of teeth
K04 Diseases of pulp and periapical tissues
K05 Gingivitis and periodontal diseases
K06 Other disorders of gingiva and edentulous alveolar ridge
K07 Dentofacial anomalies [including malocclusion]
K08 Other disorders of teeth and supporting structures
K09 Cysts of oral region, not elsewhere classified
K10 Other diseases of jaws
K12 Stomatitis and related lesions
K13 Other diseases of lip and oral mucosa
K14 Diseases of tongue
L43 Lichen planus
L58 Radiodermatitis
L59 Other disorders of skin and subcutaneous tissue related to radiation
L66 Cicatricial alopecia [scarring hair loss]
M87 Osteonecrosis
Q18 Other congenital malformations of face and neck
Q31 Congenital malformations of larynx
Q32 Congenital malformations of trachea and bronchus

Cluster 16
H60 Otitis externa
H62 Disorders of external ear in diseases classified elsewhere
H65 Nonsuppurative otitis media
H66 Suppurative and unspecified otitis media
H67 Otitis media in diseases classified elsewhere
H68 Eustachian salpingitis and obstruction
H69 Other disorders of Eustachian tube
H70 Mastoiditis and related conditions
H71 Cholesteatoma of middle ear
H72 Perforation of tympanic membrane
H73 Other disorders of tympanic membrane
H74 Other disorders of middle ear and mastoid
H80 Otosclerosis
H83 Other diseases of inner ear
H90 Conductive and sensorineural hearing loss
H91 Other hearing loss
H92 Otalgia and effusion of ear
H93 Other disorders of ear, not elsewhere classified
H94 Other disorders of ear in diseases classified elsewhere
H95 Postprocedural disorders of ear and mastoid process, not elsewhere classified
J65 Pneumoconiosis associated with tuberculosis

Cluster 45
E84 Cystic fibrosis
J01 Acute sinusitis
J30 Vasomotor and allergic rhinitis
J31 Chronic rhinitis, nasopharyngitis and pharyngitis
J32 Chronic sinusitis
J33 Nasal polyp

Cluster 61

C69 Malignant neoplasm of eye and adnexa
D31 Benign neoplasm of eye and adnexa
H11 Other disorders of conjunctiva
H31 Other disorders of choroid

Cluster 75
H02 Other disorders of eyelid
H04 Disorders of lacrimal system
Q10 Congenital malformations of eyelid, lacrimal apparatus and orbit

Cluster 89
C44 Other malignant neoplasms of skin
D04 Carcinoma in situ of skin
L57 Skin changes due to chronic exposure to nonionizing radiation

Cluster 98
C30 Malignant neoplasm of nasal cavity and middle ear
C31 Malignant neoplasm of accessory sinuses

Cluster 104
C43 Malignant melanoma of skin
D03 Melanoma in situ

Meta Cluster 7

Cluster 3
A04 Other bacterial intestinal infections
A05 Other bacterial foodborne intoxications, not elsewhere classified
A27 Leptospirosis
A41 Other septicaemia
A48 Other bacterial diseases, not elsewhere classified
A98 Other viral haemorrhagic fevers, not elsewhere classified
B48 Other mycoses, not elsewhere classified
D65 Disseminated intravascular coagulation [defibrination syndrome]
E70 Disorders of aromatic amino-acid metabolism
E78 Disorders of lipoprotein metabolism and other lipidaemias
E79 Disorders of purine and pyrimidine metabolism
I10 Essential (primary) hypertension
I11 Hypertensive heart disease
I12 Hypertensive renal disease
I13 Hypertensive heart and renal disease
I15 Secondary hypertension
I22 Subsequent myocardial infarction
I70 Atherosclerosis
I73 Other peripheral vascular diseases
I74 Arterial embolism and thrombosis
I77 Other disorders of arteries and arterioles
I95 Hypotension
J09 Influenza due to identified avian influenza virus
J11 Influenza, virus not identified
J80 Adult respiratory distress syndrome
J81 Pulmonary oedema
K52 Other noninfective gastroenteritis and colitis
K55 Vascular disorders of intestine
K65 Peritonitis
L87 Transepidermal elimination disorders
M10 Gout
M30 Polyarteritis nodosa and related conditions
M31 Other necrotizing vasculopathies
N00 Acute nephritic syndrome
N01 Rapidly progressive nephritic syndrome
N02 Recurrent and persistent haematuria
N03 Chronic nephritic syndrome
N04 Nephrotic syndrome
N05 Unspecified nephritic syndrome
N06 Isolated proteinuria with specified morphological lesion
N07 Hereditary nephropathy, not elsewhere classified
N08 Glomerular disorders in diseases classified elsewhere
N14 Drug- and heavy-metal-induced tubulo-interstitial and tubular conditions
N15 Other renal tubulo-interstitial diseases
N16 Renal tubulo-interstitial disorders in diseases classified elsewhere
N17 Acute renal failure
N18 Chronic renal failure
N19 Unspecified renal failure
N25 Disorders resulting from impaired renal tubular function
N26 Unspecified contracted kidney
N27 Small kidney of unknown cause

Cluster 7
D82 Immunodeficiency associated with other major defects
I20 Angina pectoris
I21 Acute myocardial infarction
I23 Certain current complications following acute myocardial infarction
I25 Chronic ischaemic heart disease
I26 Pulmonary embolism
I27 Other pulmonary heart diseases
I28 Other diseases of pulmonary vessels
I36 Nonrheumatic tricuspid valve disorders
I37 Pulmonary valve disorders
I42 Cardiomyopathy
I44 Atrioventricular and left bundle-branch block
I45 Other conduction disorders
I46 Cardiac arrest
I47 Paroxysmal tachycardia
I48 Atrial fibrillation and flutter
I49 Other cardiac arrhythmias
I50 Heart failure
I51 Complications and ill-defined descriptions of heart disease
I52 Other heart disorders in diseases classified elsewhere
I97 Postprocedural disorders of circulatory system, not elsewhere classified

I99 Other and unspecified disorders of circulatory system
Q20 Congenital malformations of cardiac chambers and connections
Q21 Congenital malformations of cardiac septa
Q22 Congenital malformations of pulmonary and tricuspid valves
Q23 Congenital malformations of aortic and mitral valves
Q24 Other congenital malformations of heart
Q25 Congenital malformations of great arteries
Q26 Congenital malformations of great veins
Q33 Congenital malformations of lung
Q34 Other congenital malformations of respiratory system
Q96 Turner's syndrome

Cluster 17
A99 Unspecified viral haemorrhagic fever
B25 Cytomegaloviral disease
B44 Aspergillosis
B49 Unspecified mycosis
B96 Other bacterial agents as the cause of diseases classified to other chapters
C91 Lymphoid leukaemia
C92 Myeloid leukaemia
C93 Monocytic leukaemia
C94 Other leukaemias of specified cell type
C95 Leukaemia of unspecified cell type
D46 Myelodysplastic syndromes
D59 Acquired haemolytic anaemia
D60 Acquired pure red cell aplasia [erythroblastopenia]
D61 Other aplastic anaemias
D69 Purpura and other haemorrhagic conditions
D70 Agranulocytosis
D72 Other disorders of white blood cells
D73 Diseases of spleen
D76 Certain diseases involving lymphoreticular tissue and reticulohistiocytic system
D77 Other disorders of blood and blood-forming organs in diseases classified elsewhere

Cluster 20
A40 Streptococcal septicaemia
B95 Streptococcus and staphylococcus as the cause of diseases classified to other chapters
I01 Rheumatic fever with heart involvement
I05 Rheumatic mitral valve diseases
I06 Rheumatic aortic valve diseases
I07 Rheumatic tricuspid valve diseases
I08 Multiple valve diseases
I09 Other rheumatic heart diseases
I24 Other acute ischaemic heart diseases
I33 Acute and subacute endocarditis
I34 Nonrheumatic mitral valve disorders
I35 Nonrheumatic aortic valve disorders
I38 Endocarditis, valve unspecified
I39 Endocarditis and heart valve disorders in diseases classified elsewhere
I41 Myocarditis in diseases classified elsewhere

Cluster 21
A28 Other zoonotic bacterial diseases, not elsewhere classified
A32 Listeriosis
B59 Pneumocystosis
B97 Viral agents as the cause of diseases classified to other chapters
C81 Hodgkin's disease
C82 Follicular [nodular] non-Hodgkin's lymphoma
C83 Diffuse non-Hodgkin's lymphoma
C84 Peripheral and cutaneous T-cell lymphomas
C85 Other and unspecified types of non-Hodgkin's lymphoma
C88 Malignant immunoproliferative diseases
C96 Other and unspecified malignant neoplasms of lymphoid, haematopoietic and related tissue
D89 Other disorders involving the immune mechanism, not elsewhere classified
I88 Nonspecific lymphadenitis
L04 Acute lymphadenitis
L41 Parapsoriasis

Cluster 23
A70 Chlamydia psittaci infection
J21 Acute bronchiolitis
J41 Simple and mucopurulent chronic bronchitis
J42 Unspecified chronic bronchitis
J43 Emphysema
J44 Other chronic obstructive pulmonary disease
J45 Asthma
J46 Status asthmaticus
J47 Bronchiectasis
J63 Pneumoconiosis due to other inorganic dusts
J66 Airway disease due to specific organic dust
J67 Hypersensitivity pneumonitis due to organic dust
J82 Pulmonary eosinophilia, not elsewhere classified
J84 Other interstitial pulmonary diseases
J98 Other respiratory disorders

Cluster 74
D45 Polycythaemia vera
D75 Other diseases of blood and blood-forming organs
E45 Retarded development following protein-energy malnutrition

Cluster 100
C90 Multiple myeloma and malignant plasma cell neoplasms
D47 Other neoplasms of uncertain or unknown behaviour of lymphoid, haematopoietic and related tissue

Meta Cluster 8

Cluster 19
A01 Typhoid and paratyphoid fevers
A02 Other salmonella infections
A06 Amoebiasis
A68 Relapsing fevers
A79 Other rickettsioses
A90 Dengue fever [classical dengue]
B50 Plasmodium falciparum malaria
B51 Plasmodium vivax malaria
B52 Plasmodium malariae malaria
B53 Other parasitologically confirmed malaria
B54 Unspecified malaria
B65 Schistosomiasis [bilharziasis]
B74 Filariasis
B76 Hookworm diseases
B78 Strongyloidiasis
B83 Other helminthiases
B89 Unspecified parasitic disease

Cluster 63
H10 Conjunctivitis
H15 Disorders of sclera
H20 Iridocyclitis
H22 Disorders of iris and ciliary body in diseases classified elsewhere

Cluster 78
I78 Diseases of capillaries
L70 Acne
L71 Rosacea

Cluster 82
L63 Alopecia areata
L64 Androgenic alopecia
L65 Other nonscarring hair loss

Cluster 102
D86 Sarcoidosis
L52 Erythema nodosum

Cluster 110
H00 Hordeolum and chalazion
H01 Other inflammation of eyelid

Cluster 117
A75 Typhus fever
A77 Spotted fever [tick-borne rickettsioses]

Meta Cluster 9

Cluster 24
I71 Aortic aneurysm and dissection
I79 Disorders of arteries, arterioles and capillaries in diseases classified elsewhere
J34 Other disorders of nose and nasal sinuses
M95 Other acquired deformities of musculoskeletal system and connective tissue
Q30 Congenital malformations of nose
Q35 Cleft palate
Q36 Cleft lip
Q37 Cleft palate with cleft lip
Q38 Other congenital malformations of tongue, mouth and pharynx
Q79 Congenital malformations of the musculoskeletal system, not elsewhere classified
Q87 Other specified congenital malformation syndromes affecting multiple systems
Q89 Other congenital malformations, not elsewhere classified
Q91 Edwards' syndrome and Patau's syndrome

Cluster 54
M72 Fibroblastic disorders
Q71 Reduction defects of upper limb
Q72 Reduction defects of lower limb
Q73 Reduction defects of unspecified limb
Q74 Other congenital malformations of limb(s)

Cluster 84
D16 Benign neoplasm of bone and articular cartilage
Q77 Osteochondrodysplasia with defects of growth of tubular bones and spine
Q78 Other osteochondrodysplasias

Cluster 103
Q16 Congenital malformations of ear causing impairment of hearing
Q17 Other congenital malformations of ear

Cluster 119
Q67 Congenital musculoskeletal deformities of head, face, spine and chest

Cluster 120
Q68 Other congenital musculoskeletal deformities

Cluster 121
M43 Other deforming dorsopathies

Meta Cluster 10

Cluster 33
A39 Meningococcal infection
A87 Viral meningitis
G00 Bacterial meningitis, not elsewhere classified
G02 Meningitis in other infectious and parasitic diseases classified elsewhere
G03 Meningitis due to other and unspecified causes
G06 Intracranial and intraspinal abscess and granuloma
G07 Intracranial and intraspinal abscess and granuloma in diseases classified elsewhere
G09 Sequelae of inflammatory diseases of central nervous system
H75 Other disorders of middle ear and mastoid in diseases classified elsewhere

Cluster 46
A85 Other viral encephalitis, not elsewhere classified

A86 Unspecified viral encephalitis
A89 Unspecified viral infection of central nervous system
B94 Sequelae of other and unspecified infectious and parasitic diseases
G04 Encephalitis, myelitis and encephalomyelitis
G05 Encephalitis, myelitis and encephalomyelitis in diseases classified elsewhere

Cluster 52
A35 Other tetanus
G35 Multiple sclerosis
G36 Other acute disseminated demyelination
G37 Other demyelinating diseases of central nervous system
H46 Optic neuritis

Cluster 55
A26 Erysipeloid
A69 Other spirochaetal infections
A84 Tick-borne viral encephalitis
G01 Meningitis in bacterial diseases classified elsewhere
G51 Facial nerve disorders

Cluster 81
M45 Ankylosing spondylitis
M46 Other inflammatory spondylopathies
M49 Spondylopathies in diseases classified elsewhere

Meta Cluster 11

Cluster 31
C70 Malignant neoplasm of meninges
C71 Malignant neoplasm of brain
C72 Malignant neoplasm of spinal cord, cranial nerves and other parts of central nervous system
D32 Benign neoplasm of meninges
D33 Benign neoplasm of brain and other parts of central nervous system
D42 Neoplasm of uncertain or unknown behaviour of meninges
D43 Neoplasm of uncertain or unknown behaviour of brain and central nervous system
G32 Other degenerative disorders of nervous system in diseases classified elsewhere
Q85 Phakomatoses, not elsewhere classified

Cluster 66
D22 Melanocytic naevi
D23 Other benign neoplasms of skin
L81 Other disorders of pigmentation
L82 Seborrhoeic keratosis

Cluster 76
D18 Haemangioma and lymphangioma, any site
Q27 Other congenital malformations of peripheral vascular system
Q82 Other congenital malformations of skin

Cluster 118
Q28 Other congenital malformations of circulatory system

Meta Cluster 12

Cluster 36
F81 Specific developmental disorders of scholastic skills
F82 Specific developmental disorder of motor function
F83 Mixed specific developmental disorders
F90 Hyperkinetic disorders
F91 Conduct disorders
F93 Emotional disorders with onset specific to childhood
F94 Disorders of social functioning with onset specific to childhood and adolescence
F98 Other behavioural and emotional disorders with onset usually occurring in childhood and adolescence

Cluster 38
B47 Mycetoma
N47 Redundant prepuce, phimosis and paraphimosis
N48 Other disorders of penis
Q53 Undescended testicle
Q54 Hypospadias
Q55 Other congenital malformations of male genital organs
Q56 Indeterminate sex and pseudohermaphroditism

Cluster 91
Q41 Congenital absence, atresia and stenosis of small intestine
Q43 Other congenital malformations of intestine
Q45 Other congenital malformations of digestive system

Cluster 107
F42 Obsessive-compulsive disorder
F95 Tic disorders

Meta Cluster 13

Cluster 26
H21 Other disorders of iris and ciliary body
H25 Senile cataract
H26 Other cataract
H27 Other disorders of lens
H33 Retinal detachments and breaks
H40 Glaucoma
H42 Glaucoma in diseases classified elsewhere
H44 Disorders of globe
H59 Postprocedural disorders of eye and adnexa, not elsewhere classified
Q11 Anophthalmos, microphthalmos and macrophthalmos
Q12 Congenital lens malformations
Q13 Congenital malformations of anterior segment of eye
Q15 Other congenital malformations of eye

Cluster 62
B58 Toxoplasmosis
H30 Chorioretinal inflammation
H32 Chorioretinal disorders in diseases classified elsewhere
H35 Other retinal disorders

Cluster 80
H34 Retinal vascular occlusions
H43 Disorders of vitreous body
H45 Disorders of vitreous body and globe in diseases classified elsewhere

Meta Cluster 14

Cluster 30
E05 Thyrotoxicosis [hyperthyroidism]
E06 Thyroiditis
E07 Other disorders of thyroid
E50 Vitamin A deficiency
E52 Niacin deficiency [pellagra]
H03 Disorders of eyelid in diseases classified elsewhere
H05 Disorders of orbit
H06 Disorders of lacrimal system and orbit in diseases classified elsewhere
H58 Other disorders of eye and adnexa in diseases classified elsewhere

Cluster 67
C73 Malignant neoplasm of thyroid gland
D34 Benign neoplasm of thyroid gland
E01 Iodine-deficiency-related thyroid disorders and allied conditions
E04 Other nontoxic goitre

Cluster 92
E00 Congenital iodine-deficiency syndrome
E02 Subclinical iodine-deficiency hypothyroidism
E03 Other hypothyroidism

Meta Cluster 15

Cluster 15
A38 Scarlet fever
E76 Disorders of glycosaminoglycan metabolism
F70 Mild mental retardation
F71 Moderate mental retardation
F78 Other mental retardation
F79 Unspecified mental retardation
F80 Specific developmental disorders of speech and language
F84 Pervasive developmental disorders
F88 Other disorders of psychological development
F89 Unspecified disorder of psychological development
H50 Other strabismus
H51 Other disorders of binocular movement
H52 Disorders of refraction and accommodation
H53 Visual disturbances
H54 Blindness and low vision
H55 Nystagmus and other irregular eye movements
H57 Other disorders of eye and adnexa
H61 Other disorders of external ear
Q14 Congenital malformations of posterior segment of eye
Q86 Congenital malformation syndromes due to known exogenous causes, not elsewhere classified
Q90 Down's syndrome
Q92 Other trisomies and partial trisomies of the autosomes, not elsewhere classified
Q93 Monosomies and deletions from the autosomes, not elsewhere classified
Q95 Balanced rearrangements and structural markers, not elsewhere classified
Q99 Other chromosome abnormalities, not elsewhere classified

Cluster 79
G52 Disorders of other cranial nerves
G53 Cranial nerve disorders in diseases classified elsewhere
H49 Paralytic strabismus

Meta Cluster 16

Cluster 28
C22 Malignant neoplasm of liver and intrahepatic bile ducts
C23 Malignant neoplasm of gallbladder
C24 Malignant neoplasm of other and unspecified parts of biliary tract
C25 Malignant neoplasm of pancreas
C26 Malignant neoplasm of other and ill-defined digestive organs
D13 Benign neoplasm of other and ill-defined parts of digestive system
K31 Other diseases of stomach and duodenum
K80 Cholelithiasis
K81 Cholecystitis
K82 Other diseases of gallbladder
K83 Other diseases of biliary tract

Cluster 83
K85 Acute pancreatitis
K86 Other diseases of pancreas
K87 Disorders of gallbladder, biliary tract and pancreas in diseases classified elsewhere

Meta Cluster 17

Cluster 99
E65 Localized adiposity
E68 Sequelae of hyperalimentation

Cluster 106
L90 Atrophic disorders of skin
L91 Hypertrophic disorders of skin

Meta Cluster 18

Cluster 57
N44 Torsion of testis
N45 Orchitis and epididymitis

N49 Inflammatory disorders of male genital organs, not elsewhere classified
N50 Other disorders of male genital organs
N51 Disorders of male genital organs in diseases classified elsewhere

Meta Cluster 19

Cluster 59
N31 Neuromuscular dysfunction of bladder, not elsewhere classified
Q05 Spina bifida
Q06 Other congenital malformations of spinal cord
Q76 Congenital malformations of spine and bony thorax