DOI: 10.1111/cgf.13891

COMPUTER GRAPHICS *forum*
Volume 39 (2020), number 1 pp. 543–580

A Survey of Visual Analytics for Public Health

Bernhard Preim[1] and Kai Lawonn[2]

[1]Department of Simulation and Graphics, University of Magdeburg, Germany
bernhard.preim@ovgu.de
[2]Faculty of Mathematics and Computer Science, University of Jena, Germany
kai.lawonn@uni-jena.de

Abstract

We describe visual analytics solutions aiming to support public health professionals, and thus, preventive measures. Prevention aims at advocating behaviour and policy changes likely to improve human health. Public health strives to limit the outbreak of acute diseases as well as the reduction of chronic diseases and injuries. For this purpose, data are collected to identify trends in human health, to derive hypotheses, e.g. related to risk factors, and to get insights in the data and the underlying phenomena. Most public health data have a temporal character. Moreover, the spatial character, e.g. spatial clustering of diseases, needs to be considered for decision-making. Visual analytics techniques involve (subspace) clustering, interaction techniques to identify relevant subpopulations, e.g. being particularly vulnerable to diseases, imputation of missing values, visual queries as well as visualization and interaction techniques for spatio-temporal data. We describe requirements, tasks and visual analytics techniques that are widely used in public health before going into detail with respect to applications. These include outbreak surveillance and epidemiology research, e.g. cancer epidemiology. We classify the solutions based on the visual analytics techniques employed. We also discuss gaps in the current state of the art and resulting research opportunities in a research agenda to advance visual analytics support in public health.

Keywords: medical imaging, visualization, visual analytics, visualization

ACM CCS: • Computer Applications → Life and Medical Sciences

1. Introduction

Visual analytics—the science of analytical reasoning facilitated by interactive visual interfaces [TC05]—has a great potential to support the whole health care system, including decision support for clinical medicine and rehabilitation as well as the public health (PH) care sector which is the focus of this survey. Interactive visual interfaces enable filtering, i.e. to restrict the amount of information to be displayed, flexible combinations of different aspects or layers of information and an adaptation of the visual representation, e.g. to switch between various levels of aggregation.

According to the centre of disease control, 'Public health is the science of protecting and improving the health of people and their communities ... by promoting healthy lifestyles, researching disease and injury prevention, and detecting, preventing and responding to infectious diseases'. https://www.cdcfoundation.org/what-public-health. This definition is in line with the classic definition from Amory [Amo20]. 'Public health is the science and art of preventing disease, prolonging life and promoting human health through organized efforts and informed choices of society, organizations, public and private, communities and individuals'. Despite changes in the history of PH [Ros15], including a stronger focus on environmental health, we consider this definition as still valid.

PH activities aim at concrete measures to maintain or improve health, e.g. with vaccination campaigns, screening programmes to detect severe diseases early or measures to improve the safety in traffic or at work. PH academics acknowledge the 'immense capacity' of visualization tools 'to examine various dimensions of PH data including spatial, temporal and other attributes ... beyond the capacity of statistical analysis' [JAK*17]. PH experts consider visual analytics also as a means to improve the 'ability to communicate findings and key messages' [MOSB16]. Despite this potential, adoption of visual analytics in PH is slow [OS14]. PH academics also raise concerns, e.g. that 'visualization is misleading users due to misinterpretation or cognitive load' [CAD*14].

As a basis for disease understanding, epidemiological research employs clinical data and population-based studies where a representative set of participants in a region is involved. The identification of risk factors, the analysis of the relative risk of single factors as well as their combined influence, the so-called *interaction*, are primary research goals in preventive health care. We also consider urgent problems, related to the increased frequency of a health problem, e.g. in case of a food-borne or infectious disease that spread stronger than expected based on seasonal patterns. In these cases, urgent decision support based on current data, e.g. over-the-counter sales of drugs, is needed.

So far, there is no survey article on visual analytics in PH. A survey article from Rind *et al.* [RWA*13] discussed the use of electronic health records with information visualization and visual analytics. The significant difference to this paper is their focus on clinical decision-making instead of PH activities. Shneiderman *et al.* [SPH13] provide a discussion of trends in health care and the role of visualization. They describe PH as one component of 'Health 2.0' (personal health and clinical decision-making being the two other components) with a focus on user needs and tasks. Carroll *et al.* [CAD*14] focused on software for infectious disease epidemiology. Visual analytics is considered along with other issues, e.g. databases, security, user needs and usability. In contrast, this paper analyses software for PH through the lens of visualization and analytics. In the chapter 'Visual Analytics of Image-Centric Cohort Studies in Epidemiology' [PKH*16], a specific problem, namely the analysis of population-based cohort study data, is discussed. The this paper is broader in scope and considers more recent work.

Scope of the survey. We focus on visual analytics solutions where the user is in the loop and their reasoning process is supported. Thus, we exclude pure machine learning solutions. We restrict to *structured data*, e.g. measured data or categorical data. We do not consider unstructured text and related text mining methods that are rarely used for PH-related research.

With our focus on *public health*, we exclude publications that deal with decision support for the treatment of *one* patient. Since visual analytics methods for epidemiology are partially based on such techniques, we mention as selective examples the CareCruiser [GAK*11], VisuExplore [RAM*11] and the pioneering work on LifeLines [PMR*96]. The major contribution of these papers is the interactive visualization of time-dependent data characterizing the person's health at various dimensions, e.g. concerning laboratory values, medication, hospital stays and interventions.

We also exclude approaches that enable patients to analyse their health data. We consider epidemiologists and environmental health specialists as primary target users. Due to our focus on PH activities, we also exclude papers focused on drug development (see [SBCvdS04] for an example). Visual analytics in basic neuroscience research is not considered since it is not explicitly linked to preventive medicine. We do not consider animal health. However, we include visual analytics related to *zoonotic* diseases that may be transmitted to humans.

Selection strategy. A comprehensive search in digital libraries from computer science and medicine was performed. We searched in the following digital libraries (last update, July 4, 2019): Eurographics DL, IEEE DL using the keywords 'epidemiology', 'pandemic', 'PH' and 'prevention'. We found 21 papers in the IEEE DL and Eurographics DL and refined our selection according to the following criteria: we excluded poster presentations and most short papers unless they are particularly relevant for this survey. This resulted in nine publications. The search in the ACM Digital library with the same keywords lead to 17 further publications that seemed relevant based on their title. A considerable portion only discussed the simulation models for predicting the course of outbreaks and static visualizations were used to convey the simulation model. After removing these papers, five further papers, from conferences and workshops such as ACM SIGCHI and ACM Advanced Visual Interfaces, were added. To identify medical journal papers, where visual analytics is applied to PH problems, we searched in PubMed using the keywords 'Visual Analytics'. There we found 494 papers, mostly in medical journals, where the term 'Visual analytics' is used surprisingly often. Many of these papers relate to clinical decision support, or medical research based on genomic data and are thus not considered. Others comprise PH but are only loosely connected to visual analytics. Since all IEEE TVCG papers are indexed by PubMed, many 'Visual analytics' papers without any relation to medicine are listed there. The PubMed search led to additional 24 papers. Furthermore, we searched the keywords 'Visual analytics' combined with 'PH', 'Epidemiology' or 'Pandemics' on Google Scholar, which lead to 11 further publications.

With our initial search, we found several papers that discussed the influence of air quality, which we had not anticipated. To make sure that we do not miss anything important related to the influence of air quality on human health, we explicitly searched for further papers using the keywords 'Visual Analytics Air Quality' since this topic gained a lot of attention and clearly fits in the PH scope. This search resulted in 11 additional papers (in total 14 air quality-related publications). We also searched for 'Visual analytics water quality' but did not find any additional paper.

Besides, we manually searched in all papers presented at the IEEE Visual Analytics in Healthcare Workshop (2010-2018). However, only two papers of that workshop with a focus on PH were found, since clinical medicine is the focus of this workshop. Further papers were identified by analysing the references and citations of the selected papers. Most papers are from *IEEE TVCG* (20), *IEEE VAST* (6), *IEEE CGA* (4), *Information Visualization* (3) as well as the *Journal of the American Medical Informatics Association* (5), *Online Journal of Public Health Informatics* (3), the *International Journal of Environmental Research and Public Health* (3) and *Biomedical Informatics* (3). The supplementary material lists 15 other venues where two papers were considered.

Organization. In Section 2, we describe the scope of PH activities, the essential stakeholders, high-level tasks and requirements for visual analytics support. In Section 3, we describe data that are frequently used in PH with a focus on research in epidemiology. In Section 4, we discuss visualization and interaction techniques that are widely used in PH. These commonly used techniques include geospatial views and time-oriented visualizations.

We describe specific applications in different branches of PH in the following sections (5–7). In these applications, we largely see the techniques introduced in Section 4 but often combined in a specific

manner based on characteristics of the data and the requirements specific for this application. The detection of disease outbreaks and response management is discussed in Section 5. In Section 6, we analyse a wide range of epidemiology research questions, e.g. related to cancer epidemiology, air quality and injuries. In Section 7, we discuss the analysis of population-based cohort study data aiming at the assessment of risk factors for frequent health disorders. In these studies, healthy participants are assessed and followed over time to characterize health risks and their influence on the initiation and development of diseases.

An essential aspect of visual analytics in PH is the evaluation, ideally based on using the systems by real users doing actual tasks. In Section 8, we discuss evaluation strategies and selected results. The analysis of the current state of the art also reveals a number of gaps and research opportunities. We discuss a corresponding research agenda in Section 9.

2. Public Health

PH activities typically start with gathering information about a potential health problem, e.g. after an alert, including the exploration of the available data, e.g. recent cases of a reportable disease, and go on with statistical analysis and presentation of results, e.g. a set of diagrams. **Incidence and prevalence.** We focus on health problems in a narrower sense and PH activities dealing with long-term developments of *incidence* and *prevalence*. The incidence is the proportion of a population that newly acquires a disease in a certain period, typically in 1 year. The prevalence is the proportion of a population affected by a disease, i.e. for chronic diseases, such as diabetes, the incidence is rather low but the prevalence is high.

A focus of preventive health care is a better understanding of *avoidable* lifestyle-related risk factors, e.g. obesity, low level of physical exercises or poor nutrition and environmental factors, such as air quality. A related aspect is the analysis of disease networks, i.e. if a certain disease frequently co-occurs with another one, or whether the outbreak of a disease involves a higher risk for the outbreak of another, often more severe disease. Health care data often exhibit quality problems, such as noisy, unreliable or missing data. Thus, visual analytics solutions should consider potential quality problems and provide remedies.

PH institutions exist at various levels: from community authorities to the World Health Organization (WHO). These institutions are engaged in the comprehensive surveillance of major issues related to the health of populations. In the United States, the centres for disease control and prevention (https://www.cdc.gov/) also provide up-to-date information for the general public and PH experts. The National Health Service plays a similar role in the United Kingdom [TRL*17]. As Zakkar and Sedig point out, the available data from these sources are enormous. However, no sophisticated tools are provided to filter, sort or associate data [ZS17].

Interventions to improve health need to be justified by an in-depth analysis of data, which are to a large extent collected for the purpose of informing health policy. Thus, there is a growing demand for *evidence-based health measures* [BGL99, OS14] which includes the need for re-evaluation whether certain measures are as effective as supposed when they were established. While the effect on human health is the dominant criterion, other issues, in particular cost-effectiveness, are also considered and re-evaluated. Despite the trend towards evidence-based PH, policy development is also based on media attention for a health problem [ZS17].

2.1. Epidemiology

Epidemiology, a term that is related to *epidemic*, originally dealt with the outbreak of infectious diseases [Win20]. John Snow's detection of a water pump as a source of a Cholera outbreak in London in the 1850s was a landmark event. Snow depicted the home of the patients on a map and thus became aware of a spatial cluster centred around a pump that was found to be contaminated. Today, epidemiology, as an essential part of PH, aims at evidence-based knowledge related to the distribution of diseases. The *demographics*, e.g. the characterization of the patients in terms of age, gender, race, income levels and family status, the spatial distribution of patients and temporal developments are core aspects of the distribution of diseases. Epidemiology also investigates *exposures*, i.e. factors that may influence the health status. Environmental conditions, poisoned air or water, or a genetic variant are factors to which a part of the population is exposed. Epidemiology research aims at identifying relations between exposures and diseases or injuries in a *defined population*, i.e. the population of a particular region eventually further restricted to an age group. A representative sample of this population is defined in a randomized manner and invited to participate in a study.

If a disease is correlated with an exposure, research follows to assess whether correlations imply a causal effect. Often, this is not the case, e.g. because a confounding variable is responsible for the observed effect. A famous example is a strong association of shoe size with life expectancy: people with larger shoe size die earlier. The confounding variable here is the gender: people with large shoe size are typically men and men die earlier than women [FF11].

Epidemiology research is often triggered by observations from clinical medicine that lead to the generation of hypotheses. Hypothesis-based testing, the use of confidence intervals, the statistical significance of correlations and the computation of effect sizes are specific examples for this statistical basis. As a further ingredient, epidemiologists employ *biological and medical knowledge* to derive hypotheses and to assess the plausibility of findings.

Subfields. Epidemiology is similarly specialized according to organ systems and related health indicators and diseases, i.e. an epidemiologist is often an expert for a subdiscipline such as:

- *Neuroepidemiology*, the field that aims, for example, at the prevention of neurodegenerative diseases, such as Morbus Alzheimer and Morbus Parkinson, analysing the influence of nutrition, physical exercises, social relations, cardiovascular risk factors and genetic variants on disease outbreak.
- *Pharmacoepidemiology*, the field that analyses the use of drugs and their effects on human health, including adverse effects.
- *Cancer epidemiology*, the field that deals with tumour diseases and the influence of lifestyle-related variables and genetic variants. An important observation is that some risk factors affect a variety of tumour diseases, whereas others, e.g. some virus types, are associated with the specific risk for one type of cancer.

This list only describes selected examples. The specialization of epidemiologists is necessary since the interpretation of any statistical finding requires background knowledge related to the biological and physiological processes that may explain such phenomena. At the same time, strong specialization may restrict the results of epidemiology research. Modern large-scale epidemiology is interdisciplinary, involving experts from different disciplines, to identify and further study more complex relations, e.g. between mental illness and nutrition [FST*18].

In Section 6, we present visual analytics solutions for the three subfields mentioned above.

2.2. Epidemiological instruments

Despite the specifics of the subfields, the instruments that epidemiologists employ for answering scientific questions are broadly applied. The selection of an instrument or *study type* is based on the research questions that should be answered as well as on available resources. According to [Pea12] and [HM*87], study types include:

- *case series*,
- *case-control studies*,
- *cross-sectional studies* and
- *cohort studies*.

A *case series* comprises patients who suffer from a disease or persons exposed to a risk. These persons are monitored over time to study patterns of the development of their health status.

Within a *case-control study*, a second group, namely the control group, is added. The control group is as similar as possible to the case group in terms of major demographic factors, such as age, gender and health status. However, the members of the control group are not exposed to the risk. The British doctor's study is a famous example. Already in 1956, this study clearly indicated that tobacco smoking (case group) significantly increases the risk of getting lung cancer [DH56]. The major result of case-control studies is the *Odds ratio* that characterizes how the chance of getting a disease is affected by an exposure. An odds ratio significantly above 1 is a risk factor. A special type of a case-control study is the *interventional study*, where one group is treated with an intervention and the control group that is not treated.

A *cross-sectional study*, also called health survey, has only one point in time where data related to the health status of a group of patients or participants are gathered. Cross-sectional studies often serve to analyse the prevalence of diseases and are therefore often called *prevalence studies*.

Cohort studies serve to understand how the prevalence and incidence of diseases develop in a group of participants. Since the participants are followed for a period of time, Pearce [Pea12] uses the term *longitudinal study* as a synonym. A cohort study can be seen as a series of linked cross-sectional studies. Cohort studies are appropriate for a wider range of research questions. The analysis of the temporal development of the health status of participants may give hints to causal relations. If persons were exposed to a risk before getting a disease, the likelihood for a causal relation is larger compared to the pure coincidence.

Cross-sectional and cohort studies may be based on data of one or more hospitals or it may involve data from participants who are representative for a *defined population*. The latter are referred to as *population-based* studies. Population-based studies involve primarily healthy participants, i.e. the prevalence of a disease is low.

Visual analytics solutions are particularly useful for population-based cross-sectional and cohort studies since they typically involve many variables, and thus, may reveal surprising associations. These study types enable a broad analysis of risk factors, whereas case studies, case-control and interventional studies are restricted to a specific disease (and a specific treatment). These disease-specific studies are assessed with statistical methods.

2.3. Task analysis and requirements

In the following, we discuss the target user groups, tasks and requirements relevant to them. Revere *et al.* [RTM*07] distinguish between *PH experts* and *PH academics*. PH experts have to solve routine tasks, sometimes also urgent tasks related to an unusual situation. PH academics, on the other hand, are free from the time constraints of PH experts and are able to focus on exploratory investigations and more complex data analysis.

The major results of their activities are scientific publications related to new insights, e.g. about risk factors. A more fine-grained analysis of stakeholders reveals among others [OS14, RTM*07, MRH*08]:

- *Epidemiologists* analyse data with an in-depth understanding of statistics often to identify and assess potential risk factors.
- *Communicable disease specialists* are involved if an infectious disease spreads and contributes with their experience related to possible interventions to limit the effects of an outbreak.
- *Specialists for nutrition* provide their knowledge of food and diseases potentially related to food.
- *Environmental health scientists* analyse the air or collect samples from the ground to investigate contaminations.
- *Health policy makers* aim at changing health policy, e.g. with respect to new screening procedures.

Masoodian *et al.* [MLK16] also mention *veterinarians* for supporting an understanding of insect vectors that are particularly relevant for tropical diseases. The knowledge and analytical skills of these experts need to be efficiently integrated.

Task analysis. Thew *et al.* [TSP*09] report on a cognitive task analysis, including interviews and observations from meetings where epidemiologists discussed their approach to analyse data. They started with 22 unstructured (exploratory) interviews. An analyst then took part in seven meetings where upcoming work was discussed (1–2 h) and four longer meetings (up to 4 h) where decisions were taken.

A lot of these discussions are related to the validity and credibility of data. Thew *et al.* observed many discussions related to potential confounding variables. As an example, the potential influence of obesity on asthma is discussed. Age as well as gender are examined as confounding variables. Low credibility and information overload are considered major pain points by PH experts according to the

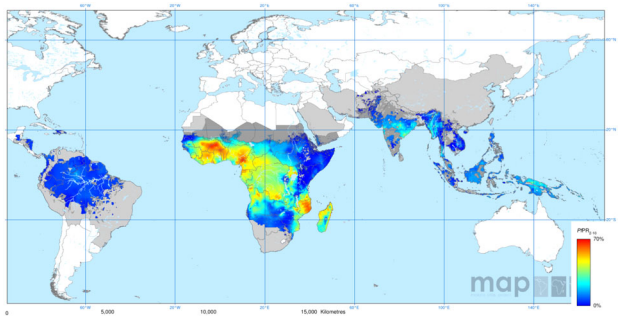

Figure 1: *The global distribution of malaria in 2010 is displayed with a heatmap (from: Wikipedia).*

systematic review by Revere *et al.* [RTM*07]. This paper is based on a literature search and summarizes findings from 31 publications. It provides a discussion of the information needs, sources of information and quality criteria to include them for decision-making. However, it is slightly biased towards PH in the United States, discussing the role of specific institutions, such as the National Institute of Health.

Carroll *et al.* [CAD*14] cite nine papers where the need for 'interactive graphics . . . to dynamic review their data at various levels' is emphasized. In the same article, ten papers are mentioned where 'users demonstrated high interest in tools with multiple panels, enabling them to review their data from multiple perspectives'. Carroll *et al.* [CAD*14] observed PH experts doing actual tasks and asked them to explain their decisions. They also analysed differences in education and experience, and investigated diseases among the PH workforce to provide the right level of support.

When reporting results, epidemiologists prefer numbers over images. Thus, epidemiologists gain substantial understanding from statistical information, whereas images, e.g. diagrams, are considered ambiguous or at least less accurate. In the design of visual analytics systems, statistics should be integrated. Only recently, data mining and visual analytics methods are used in addition to come up with new findings, i.e. these methods serve to *generate new hypotheses* [TSP*09]. Especially in the early stages of the analysis, visualizations may help to understand the data, e.g. with respect to distributions and potential abnormalities, and to select interesting subpopulations with visual query mechanisms [TSP*09].

Ola *et al.* [OS14] also discuss computer support available for PH experts and highlight STATA (https://www.stata.com/, TABLEAU (https://www.tableau.com/), and SPOTFIRE (https://spotfire.tibco.com/) as general statistics and data visualization tools, but also interactive visualizations that map diseases spatially, e.g. maps showing the malaria distribution (see Figure 1).[1] Further information on task analysis for supporting visual PH solutions can be found in Gesteland *et al.* [GLG*12].

[1] https://en.wikipedia.org/wiki/Malaria_Atlas_Project Wikipedia, Creative Commons

2.3.1. *Tasks*

On an abstract level, PH experts face typical analytical tasks, such as identifying relations, testing assumptions, generating hypotheses and supporting conclusions with sufficient evidence. On a more concrete level, the tasks of PH experts involved include [OS14, TSP*09, GLG*12, MLK16]:

- T1 *Exploration.* Gathering information about a health problem, exploring available data, e.g. recent cases of reportable disease, aggregating and displaying these data.
- T2 *Assessment and pattern identification.* Analysing a health problem, e.g. drill down to vulnerable subpopulations, such as children. This includes an analysis of the distribution of a disease in a population with respect to gender and comorbidities.
- T3 *Associations.* Finding and analysing associations between lifestyle-related factors, environmental factors, health risks and diseases. Considering different strengths of associations and emphasizing stronger associations.
- T4 *Verification.* The *verification* of an assessment, pattern or association relates to the quality of the data and the significance of results. A strong type of verification is the transfer of derived knowledge to another cohort to assess whether it can be 1 replicated there.
- T5 *Comparisons.* Support comparisons between populations, e.g. case and control group. Comparisons may be part of the verification process, e.g. when current data are compared with historic data to assess plausibility.
- T6 *Policy development.* Design of interventions to prevent health problems or limit their effect, including priority setting and assessment of involved costs.
- T7 *Dissemination.* Informing and educating the public about health problems and strategies to avoid them; ensure awareness for potentially severe health problems.

These tasks may involve *cooperative* situations where multidisciplinary teams jointly analyse the data. The cooperative analysis may be supported by shared large displays as well as coupled and decoupled modes of interaction [MLK16].

Syndromic surveillance comprises the collection and analysis of health data for outbreak detection and response management [AAA*16]. For syndromic surveillance, e.g. related to infectious diseases, the following tasks need to be supported [MLR*11]:

- T8 *Preparedness*. Regional authorities should be trained how to respond to a major outbreak and the health care system should provide sufficient resources, e.g. hospital beds.
- T9 *Outbreak detection.* Based on the monitoring of available data and assumptions, related to seasonal patterns and disease types, an outbreak should be detected as early as possible and assessed regarding its severity.
- T10 *Spatio-temporal assessment.* An outbreak needs to be monitored in space and time to understand paths of disease spreading. An analysis in different spatial scales and for selected temporal intervals is essential.
- T11 *Prediction.* Based on the available data, the progress is simulated under different assumptions, e.g. with or without certain

interventions established. Predictions involve uncertainty that needs to be conveyed.

Ola *et al.* [OS14] discuss PH problems based on a food poisoning scenario that occurs in an unseasonable period of the year: PH experts analysed confirmed cases, and displayed these data on a map along with sources of water that may serve mosquitos to breed. Thus, map-based visualizations are essential for a wide range of PH tasks (T10).

Livnat *et al.* [LRS12] make general statements about visual analytics support for syndromic surveillance. It should support *convergent* and *divergent* thinking, i.e. on the one hand, some tasks may be supported with guidance, but, on the other hand, a visual analytics system should encourage users to consider alternative decisions and assess potential consequences. Thus, a system could counteract cognitive biases that lead to a narrow decision space.

The tools currently available to PH experts are not sufficient to effectively support these tasks. We found research prototypes addressing tasks T1–T3 and T5 but no visual analytics system that explicitly supports T6 and T7. Only one system supports T4 [AHN*17a]. T8 to T11 are explicitly addressed by visual analytics (VA) systems for syndromic surveillance (Section 5). While some of these systems focus on *Preparedness* (T8), others provide support for detection (T9) and prediction (T11). Basically, all of them support a spatio-temporal assessment (T10). Among the means to support epidemiologists, Thew *et al.* [TSP*09] discuss query mechanisms and a statistics wizard that *guide* users to statistical tests suggests appropriate tests for the specific data and issues warnings when problems occur that may prevent a reliable result. Furthermore, Thew *et al.* revealed the interest in complex analysis questions where the combined influence of variables is analysed.

2.3.2. *Requirements*

We briefly discuss requirements based on Thew *et al.* [TSP*09], Sedig *et al.* [SPDO12] and our own experience [KOL*14, KLG*16]. PH experts benefit from direct support for the tasks discussed in Section 2.3.1 and the involved data management problems, e.g. the integration of data from various sources. Since PH data are often noisy or exhibit other quality problems, careful processing is needed. Interactive analysis and visualization techniques need to be adapted to PH scenarios.

As general requirement, PH systems should be designed in a user-centric way to enable users to transform the heterogeneous health data to *actionable information* [AHFSP17, GLG*12]. We extract the following specific requirements:

- R1 *Provide an overview of the data.*
- R2 *Enable analysts to integrate expert knowledge.*
- R3 *Provide familiar visualizations.*
- R4 *Provide integrated information.*
- R5 *Provide visual support for association analysis.*
- R6 *Provide visual support for comparisons.*

This set of requirements is only a starting point and needs to be extended for any specific application. Additional requirements may arise, for example, with respect to analytical components, spatial and temporal visualizations. The techniques discussed to fulfil R1 to R4 are primarily related to task T1 (exploration). R5 is related to T3 (finding associations). R6 is related to T5. In the following, we discuss the specific meaning of these general requirements for visual analytics in PH.

Provide an overview of the data. Users benefit from an overview of the available dimensions and the distribution of values for *exploratory analysis*. For a small number of dimensions, histograms may serve as an overview. For spatial data, frequency should be presented along with a map. Time-line-based visualizations, e.g. box plots over time, display a temporal development. For large numbers of dimensions or huge number of patients, more abstract visualizations are necessary, e.g. a diagram that shows for each dimension the skewness of the distribution and the interquartile range (see [TLLH13] for such an overview of dimensions). An overview that emphasizes pairs of variables with strong correlations is also desirable [KLG*16].

Enable analysts to integrate expert knowledge. In particular for hypothesis-based analysis, it is essential that analysts can specify which datasets and dimensions they are interested in. This may include the specification of (socio-)demographic features and the selection of participants based on lifestyle or variables characterizing their medical history. Moreover, analysts should be enabled to exclude participants they consider as outliers. This requirement is strongly related to task 2.

Provide familiar visualizations. A number of visualizations, e.g. certain map-based visualizations, time-line-based visualizations, Kaplan–Meier curves and Mosaic plots are widely used in PH research. Kaplan–Meier curves display the portion of patients who survive after diagnosis for a certain amount of time. They may also be used to compare the survival between subpopulations, e.g. with different health status or to compare the survival for different kinds of treatment, e.g. to report on a case-control study. Familiar visualization techniques should be preferred over techniques primarily known to visual analytics experts. Visualization techniques should also be given names that are familiar in PH, e.g. a time-line-based visualization that indicates how the frequency of a disease changes over time is called an Epi(demic) curve (see Section 4.5).

Visualization techniques based on age pyramids are widely used to show which age groups are affected and whether there is a gender effect. An age pyramid consists of two vertical histograms that indicate the number of women and men in this age group in a certain region. The number of people is shown for age groups of 5 years or for age groups of 1 year. They can also be used to display the incidence or prevalence of a disease in this age group (see Figure 2), which is referred to as *outcome pyramid* [CWCN11].

Provide integrated information. The complexity of PH tasks and the underlying data requires the design of a set of coordinated views. A careful selection of individual views, an appropriate spatial layout that supports the user in assessing relations and synchronization techniques, e.g. synchronized emphasis of information in different views, may support complex surveillance tasks. Overview types of visualization, in-depth display of selected datasets or dimensions, displays of relations between dimensions and map-based visualizations are typical components of visual analytics systems (not only) for PH.

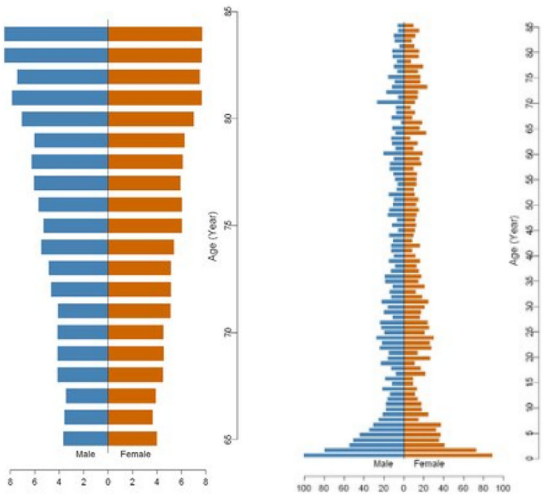

Figure 2: *Age pyramid-inspired visualizations of disease frequency. Left: Salmonellosis in the whole United States elderly population (aged 65–85). Right: Salmonellosis cases in Massachusetts with strongly increased frequency for small children and a smaller second peak among young adults (from: [CWCN11]).*

Provide visual support for association analysis. Association analysis (T3) requires dedicated support. A visual analytics system may compute correlations for all pairwise attributes or a user-specified subset. Matrix views may colour-code correlation coefficients to direct the user to strong (positive or negative) correlations. Different correlation measures may be incorporated. The rank-by-feature framework [SS05] may serve as orientation and Klemm *et al.* [KLG*16] as a solution developed for epidemiology research.

Provide visual support for comparisons. Since comparisons are essential in PH activities (T5), specific support needs to be provided. We already discussed age pyramids as a tool for gender and age comparison. Comparison support is often also needed for temporal developments, e.g. data from a current outbreak should be compared to historic data. Temporal alignment and synchronization are essential issues. Many techniques are available to support comparisons in side-by-side overviews or in an integrated manner, where different datasets are overlaid [PP95].

The requirements and tasks discussed in this section are quite general for PH. Thus, a researcher aiming at supporting PH may use them as a starting point. For a specific PH problem, e.g. injury or cancer prevention, further more specific requirements arise. These more specific requirements are discussed directly with the specific applications (Sections 5–7).

3. Data for Public Health

We now describe the data that are acquired or already available for PH activities. The large amount of data, related to patients' symptoms, diagnoses and treatment in hospitals and insurance companies can be used for research, in particular for disease understanding and for preventing diseases or further complications of an already diagnosed disease. PH-related data have unique properties that make the analysis difficult.

- The *high number of dimensions*, e.g. in population-based cohort study data often several thousand dimensions, hamper a comprehensive analysis.
- Data are *heterogeneous*, including scalar, ordinal, categorical and binary variables. Diagnosis, previous treatments and prescribed drugs are examples for categorical data. There are even hierarchies of categories, e.g. respiratory diseases and cardiovascular diseases as high-level categories. Measures derived from blood or urine samples are scalar data. The stage or severity of disease is an example for ordinal data, e.g. non-hypertensive, mild, moderate and severe hypertension. Epidemiological studies often consider whether a patient is exposed to a risk or not–yielding binary data, e.g. being a smoker.
- Data often have a *temporal* and a *spatial dimension*, e.g. data related to the outbreak of infectious or food-borne diseases.
- Despite measures to ensure data quality, it is often far from perfect. The amount of *missing data*, e.g. patients dropping out of follow-ups in cohort studies, typically is too high to restrict the analysis to complete cases.
- Data are not perfectly reliable, in particular self-reported data, e.g. on nutrition behaviour, alcohol and tobacco consumption is often biased towards social expectations, i.e. people pretend to follow a healthier lifestyle than they actually do.

In summary, PH-related data are often of *complex* nature and the analysis may benefit from flexible and tailored VA solutions (see also [AHFSP17]). A wide variety of data are systematically collected to study *public health indicators* [ZS17], including:

- *public health status*, e.g. prevalence of diseases, such as diabetes,
- *health risks*, e.g. high blood pressure or obesity,
- *outcomes of health care programmes*, such as cancer screening,
- *health equity indicators* that indicate how similar the health status is for population groups defined by social, economic or regional factors. Health equity is an essential goal of PH activities.
- *Performance indicators*, e.g. waiting times for certain diagnostic procedures or surgical intervention.

Severe diseases, such as all types of cancer, are reportable and all known cases are collected in specialized registers. Cancer registers comprise, e.g. age, gender, location of the tumour, stage [BZKF13] and in case of clinical registers also diagnostic and therapeutic procedures as well as their outcome. The same applies to severe infectious diseases, heart diseases and neurodegenerative diseases. Also, severe injuries are recorded in *trauma registries*.

For the analysis of temporal changes, it is essential to relate the data to regulations and legislative issues, e.g. the introduction of a screening programme or the prohibition of smoking in restaurants [RTM*07]. The reported cases may strongly differ from the actual cases. Thus, when the (reported) numbers change over time, analysts need to understand whether the likelihood has changed that a disease is actually reported, e.g. based on media attention, to avoid wrong conclusions. Prescription data and over-the-counter sales are timely data and thus useful to analyse short- and long-term trends. Research studies also create data that *may be* relevant for PH activities. However, credibility and validity need to be ensured in particular for industry-sponsored research [ZS17].

3.1. Population-based cohort study data

Population-based cohort studies are the ideal research instrument to answer research questions related to the 'combined effects of lifestyle, occupation, and environment, social and psychological factors and genetic predisposition on disease development' [Con14]. General goals of population-based studies are:

- estimations of prevalence and incidence of diseases,
- better understanding of the differences between healthy ageing and beginning pathologies,
- identification of risk factors for severe diseases,
- identification of pathways from risk factors to chronic diseases,
- evaluation of markers for diseases in a preclinical stage to foster specific prevention measures and
- assessment of geographic and socio-economic differences in health status in different regions.

Population-based studies involve a representative sample of the population in the target region (recall Section 2.1). Typically, a random sampling of registry data is performed to determine the sample of the persons to be invited, i.e. the epidemiological researchers have to cooperate with the local administration to get access to the relevant data, including addresses. Various measures are combined to achieve a high participation rate. Strict quality insurance policies apply to all types of data acquisition. As an example, all enrolled physicians and study nurses are instructed *how* to perform a measurement, e.g. of blood pressure, to achieve the highest possible degree of reproducibility. For the same reason, the hardware and software of MR scanners employed in a cohort study needs to be kept constant in the whole 4-year period of a cycle. Publications on cohort studies in epidemiology journals therefore dedicate a large portion to 'Quality control' [JHLea01]. The strong emphasis on data quality and the unbiased selection of participants are major differences to the retrospective analysis of data acquired in the clinical routine. Not all examinations that are desirable from a research point of view are actually integrated in cohort studies. Ethics, data protection and cost-effectiveness lead to constraints for the design of cohort studies.

While most of the goals can be achieved with regional and mid-sized studies (several thousand participants), in particular the last goal requires large, nation-wide or international studies. Cohort studies are carefully planned, including an in-depth discussion of the instruments to be used, i.e. the specific choice of examinations, questions to be answered by the participants, laboratory tests as well as imaging. The recruiting of participants is centred around the following questions [TSP*09]:

- How many participants are needed to reliably identify risk factors for selected diseases?
- What are the specific criteria to select participants, e.g. with respect to age, gender and geographic area?
- How to invite participants and render the participation attractive?

Related to the privacy of data, data need to be consequently anonymized, which means that personally identifiable information, such as the names of patients or participants are encrypted or deleted. Also, the exact birth date is removed. Only the year of birth is typically registered, which is sufficient to categorize participants with respect to age groups. Moreover, it should be avoided that a de-anonymization is possible. As an example, a computed tomography (CT) head scan with high resolution could enable to recognize the person.

Image data. In modern cohort study data, imaging is regularly used to detect early subclinical signs of diseases or precursors thereof. Hepatic steatosis (fatty liver) or left-ventricular function are examples for such subclinical signs [BKWea15]. In population-based studies, primarily magnetic resonance imaging (MRI) and ultrasound data are used. X-ray or CT involve ionizing radiation and are thus potentially harmful, representing an ethical problem in case of healthy volunteers. For the same reason, MRI is used without a contrast agent [BKWea15].

In the following, we describe selected cohort studies:

- Rotterdam study,
- Study of Health in Pomerania (SHiP) and
- UK Biobank.

While the first two studies are population-based, the UK Biobank is not. However, the invitation process is the only major difference. Thus, we mention this important study here as well.

Rotterdam study. This study was started in 1989 with a cohort of 7983 persons aged over 55 years [HvdKHea06, IBM*17]. This cohort was repeatedly examined to understand age-related effects of health with a focus on diseases with high prevalence in the elderly. This includes the coronary heart disease, neurodegenerative diseases and eye diseases, such as Glaucoma. To enable comparability, the set of investigations was rather stable over time. Only few new tests were included in later cycles, e.g. because of wider availability or increased interest in certain diseases. The initial cohort was examined in six cycles until 2015 when the cohort was reduced to 1153 persons. New cohorts with younger participants were established in 2001, 2006 and 2014 [IBM*17].

Key objectives of the study include the prevention of the first cardiovascular event, e.g. cardiac infarction, secondary prevention following a first event and the prevention of chronic diseases. Gender-specific differences were considered, e.g. the concept of a *healthy menopause* was developed for characterizing women's health.

A large variety of age- and gender-specific effects was reported based on the Rotterdam study [IBM*17]. To give a few examples, it was observed that men are at a higher risk to develop coronary heart disease as first cardiovascular event, whereas women are at a higher risk to develop cerebrovascular disease with the risk of getting an ischemic stroke. The study also enabled the analysis of the combined influence of endocrine, inflammatory and other factors on disease initiation. Thus, risk markers for disease monitoring and early detection could be identified.

SHiP. This study is carried out in the northeastern part of Germany. It was initiated in the 1990s when this region suffered from a high unemployment rate and below-average health and life expectancy after the German re-unification. It was known that differences between the East and West German population existed in terms of health indicators, such as allergies. The SHiP aimed

at extending this knowledge with comprehensive data acquisition. In contrast to the Rotterdam study, a large age range of adults (20–79 years) was considered [VASea11]. The SHiP aims at a broad range of diseases. As an example, complex dental and medical examinations were carried out and new hypotheses for relations between the dental status and a range of diseases were derived. In-depth interviews were carried out covering many aspects of the participants' social life, family history of health-related events and working conditions [JHLea01].

The SHiP encompasses two cohorts: SHiP (aka SHiP-core) started with 4308 participants in 1997 (SHiP-0) with follow-up investigations every 5 years and is currently at its fourth wave SHiP-3 (1700 participants, 2014–2016). The second cohort, SHiP-Trend, started with 4420 participants in parallel with SHiP-2.

UK Biobank. The UK Biobank involves 500 000 participants (aged 40–70 years) and a broad range of measures to understand the well-being and health status of these participants, including genetic samples, comprehensive self-reported health data and dietary intake data [GI15, SGAea15]. Three hundred twenty nine variables related to physical measures, such as blood pressure, and 471 variables, e.g. related to lifestyle and health history, were acquired during interviews. In contrast to the previously described studies, the cohort is *not* a representative sample of the population because it consists of *healthy volunteers*. Therefore, a selection bias occurs. The baseline phase of the UK Biobank was between 2006 and 2010 and follow-up investigations are ongoing. While a basic genetic phenotyping was performed for all participants, 100 000 UK Biobank participants have worn a 24-h activity monitor for a week. Online questionnaires were used to study the participants' cognitive function and work history. Imaging data for 100 000 participants are available covering the brain, the heart, the abdominal region and the skeletal anatomy [PMBea13]. The UK Biobank is linked to many other sources, e.g. electronic health records that characterize hospital information. The study data lead to the discovery of complex relationships. As an example, Firth *et al.* [FST*18] found that severe mental illnesses, e.g. schizophrenia, and bipolar disorder are related to a low-quality diet and poor nutritional status.

Joint analysis of cohort study data. Although the above-mentioned cohort studies are large and enable statistically significant results for frequent health disorders and moderate effect sizes, they are often not sufficiently large to study subtler effects or less frequent diseases. Moreover, it is scientifically more convincing if the effect that was observed in one study can be replicated in another one. This kind of confirmation could largely exclude specific local effects, e.g. due to nutrition patterns and socio-demographic specifics. The joint analysis of cohort study data is challenging, because the examinations are not completely standardized. Despite these difficulties, there are meta-studies. The most recent report on the Rotterdam study (recall [IBM*17]) contained references to such studies. To better facilitate such joint analysis, harmonization of data collection is aimed at [Con14].

Summary. Comprehensive cohort study data are acquired for medical research. Most data are available for all examination cycles; however, some were added or removed based on changing research priorities and availability of instruments. The data from the Rotterdam study and the UK Biobank were widely used for image analysis research. The first examination cycle of the German National Cohort is announced to be completed in summer 2019 and is not yet used for in-depth analysis. So far, only the SHiP data were used for visual analytics research [KOL*14, KLG*16] (see Section 7).

3.2. Clinical data

Clinical data are not the focus of most PH activities, but they add valuable information in case of urgent health problems. *Electronic Medical Records* contain all information related to the hospital stay of each patient, e.g. all diagnostic results and treatments. They primarily serve for billing purposes. Thus, information that would be interesting for research may not be available or is tedious to extract since no filtering of cohorts or statistical analysis of relations is supported [BSM*15]. Recently, tools were developed that support research based on medical data, e.g. the Observational Health Data Sciences and Informatics (OHDSI) system [HDS*15]. The OHDSI system was used by an international team with 11 partners summarizing data from 250 million patients mapped to a joint standard [HRD*16]. Such tools have a great potential, but the large majority of medical researchers are faced with hospital information systems optimized for billing purposes without support for research tasks. Hospital admission and emergency department (ED) data including basic demographics and major symptoms are further sources of information often used for the analysis of outbreaks [GLG*12, MTJ*07].

3.3. Other data for public health

A wide variety of sources are employed for solving PH tasks. These include national census data and health surveys. Moreover, the temporal development of web queries related to symptoms and diseases, prescription data and results of laboratory tests, including microbiological testing for respiratory and gastrointestinal pathogens, are useful for analysing an acute epidemic [GLG*12]. Ali *et al.* [AAA*16] mention emergency calls, school absentees and ambulatory data as further sources for infectious disease outbreak detection. Mortality data involving a precise classification of the cause of death are essential for monitoring long-term trends, related, e.g. to chronic diseases, cause of injuries or cancer epidemiology.

A *cancer registry* is an information system where comprehensive data on cancer patients are in a standardized manner to support statistical analysis and answer questions related to trends in particular types of cancer, e.g. changes in the incidence or survival rates. Cancer registry data are available in most countries and include gender- and age-specific incidence and mortality rates (the rate of people dying from a disease) that is spatially referenced [CRN*08]. These registries are population-based, i.e. *all* cancer cases in the respective region or country are considered avoiding a selection bias. While some cancer registries only represent diagnostic information (type of cancer, stage of the disease and tumour grading), others also represent the specific treatments and their timing supporting an analysis of the effectiveness of treatments. The spatial reference of patients is encoded with zip codes, electoral wards, cities and districts and thus enables an analysis at various spatial scales.

Data that are relevant for PH differ in timeliness and reliability. As an example, web queries are timely indications for an epidemic. The GoogleFlu project (https://www.google.org/flutrends/about/) aimed at predicting the development of a flu and dengue fever outbreak based on web queries faster than with conventional methods. It was not successful since the web queries are not reliable enough. Confirmed lab tests for a specific pathogen represent more reliable information.

3.4. Data preparation and data management

The data from different sources need to be loaded, validated, cleaned and integrated. This is often a time-consuming process that is not fully automated since the variety of formats is large [MOSB16]. A subsequent step is the storage of the data in an appropriate way that enables fast and convenient access. A classic relational database is not ideal for storing the complex and heterogeneous data of cohort studies, since it is not sufficiently flexible and leads to performance problems in case of queries that require table joins [AOH*14]. Data cubes, optimized for online analytical processing, enable faster access. Angelelli *et al.* [AOH*14] enhanced this concept that is used for the POLARIS system [STH02]. They implemented it with *n*-dimensional in-memory arrays.

A data dictionary is typically provided along with a cohort study. The SHiP, for example, contains a data dictionary with specific information about each variable, data type, admissible range and consistency rules that define admissible combinations between variables. We do not focus on data preparation and data management since the visual analytics publications rarely discuss these issues.

4. Commonly Used Visual Analytics Techniques

In this section, we describe rather general techniques that we found in a larger number of systems. In later sections (5–7), we describe specific problems and systems that use such general techniques but may add some special techniques or combine the general techniques in a special way. Thus, if a new PH problem is tackled, the general techniques described in this section should be considered first. Then, it may be useful to consider which of the problems described in Sections 5–7 is similar and to use the specific discussion of a similar problem as further inspiration.

4.1. Dashboards and multiple coordinated views

In most applications, several views are combined and coordinated to give an overview on heterogenous data. General strategies for multiple coordinated views (MCWs) [Rob07] also apply for PH care applications. Chui *et al.* [CWCN11] present a variant of MCWs that is applicable to a wide range of PH tasks. They combine three views to enable mental integration:

- an *age* or *outcome pyramid*, where age is depicted on the vertical axis,
- an *age-time image plot*, where age is depicted on the vertical axis and the horizontal axis represents time,
- a *timeline*, where the horizontal axis represents time and the vertical axis represents the incidence of a disease.

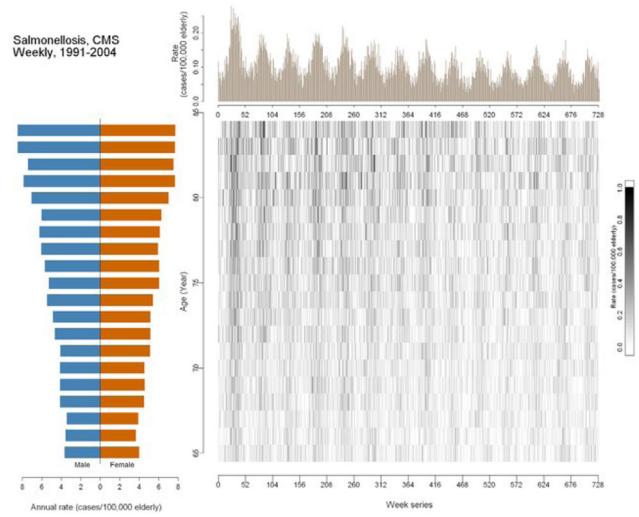

Figure 3: *An outcome pyramid (left), an image plot (lower right) and a time-line-based view (upper right) are combined. Note the common axis and the alignment of the views (from: [CWCN11]).*

The image plot is a two-dimensional (2D) histogram, where the joint frequency of a disease and an age group is counted and mapped to brightness, saturation or another one-dimensional (1D) colour scale. Such image plots may have characteristic patterns, e.g. oblique regions with increased values representing a subpopulation that suffered from a disease early and carries the disease along time when they age. The three plots are aligned such that the correlation between the outcome pyramid and the image plot in terms of the common axis 'age' and the correlation between the image plot and the timeline-based visualization along the common axis 'time' is easily perceived (see Figure 3).

While MCVs are the established term in the visual analytics field, PH experts frequently use the term *dashboard*. A dashboard presents all relevant information of a particular process or for a particular task in an integrated manner [Few06]. Many dashboards, like MCVs, are created with frameworks, such as Tableau. A difference relates to the coordination between the views: While such a coordination is mandatory for MCVs, the individual views in a dashboard are often not coordinated. Another slight difference relates to the complexity of the individual views and their combination. PH publications include dashboards that are designed for simplicity, i.e. rather simple charts or map-based visualizations are used [JAK*17]. MCVs introduced at Visual Analytics venues have a stronger tendency to include novel and quite complex designs. We use the term dashboard in particular if the authors used this term.

The individual views in dashboards and MCVs often comprise scatterplots, partially enhanced with regression lines [AOH*14, SMvB*10]. Scatterplots and scatterplot matrices often display two classes, e.g. participants with a health risk and those without (see Figure 4, left) or woman and man. The number of patients or participants is typically not huge. Therefore, overplotting is not considered. While scatterplots represent scalar values, mosaic plots are used for nominal or binary data (Figure 4, right). Graph and network visualizations are employed to study associations (task T3),

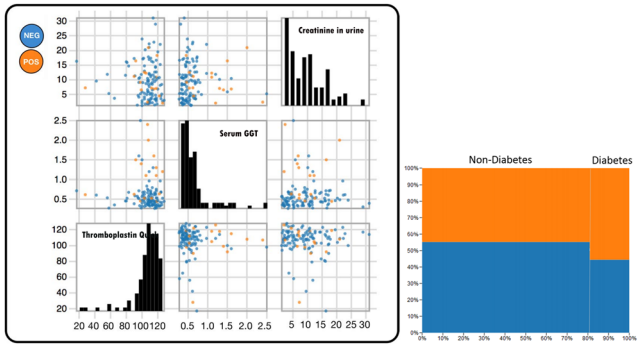

Figure 4: *A subpopulation of the SHiP data is analysed with respect to hepatic steatosis (fatty liver). The scatterplot matrix (left) displays laboratory values for participants with and without fatty liver. The mosaic plot (right) indicates that participants with a diabetes diagnosis are at higher risk for fatty liver. In both images, orange represents participants with fatty liver (Courtesy of Shiva Alemzadeh, University of Magdeburg).*

e.g. between different diseases or between diseases and exposures [BDD14]. Treemaps are occasionally used to indicate the relative frequency of events [MHD*14, TRL*17]. Histograms and other summary views may also be incorporated [LRS12].

In the EPINOME system, Livnat *et al.* [LRS12] also enable to add or remove views. They discuss, however, that the typical approach to update *all* views affected by a selection for example is often not desirable and advocate *loosely coordinated views*, where only summary views are adapted.

Web-based solutions are frequently used to support easy access of various stakeholders [ANI*17, JAK*17, LRS12, KLG*16]. Also, Carroll *et al.* [CAD*14] list numerous web-based systems for assessing infectious disease-related data.

4.2. Interactive subpopulation definition

Traditionally, health data are separately analysed for women and men and often also for different age groups to understand if certain populations are particularly affected by a disease. In an interactive system, this is supported by *demographic filtering*. Often, analysts are interested in *subpopulations* that share some risk factors or other health attributes and investigate the prevalence of diseases for them. Subpopulations may be defined in an interactive manner or by means of analytic techniques, such as clustering or decision trees. Basic interactive selection techniques include range sliders for numerical values, checkboxes or radio buttons for nominal data, such as gender. Since often several hundred variables are available, some user guidance is essential.

The challenge is to present the essential interactive facilities in an easily accessible manner. The analysis of subpopulations with statistical methods is only meaningful if the subpopulations' size is not too small. Therefore, this information should be easily recognizable whenever the filter changes. When subpopulations in a cohort study or case-control study are selected, this is referred to as *cohort construction* [KPS15].

More complex specifications may be useful, e.g. to select patients who experienced adverse drug effects (ADEs) within a certain interval after taking a drug. Thus, there are certain *events*, typically categorized, e.g. as begin/end of symptoms, begin/end of treatments, admission/readmission in the hospital, which may be used to filter the data.

Thus, epidemiologists may drill down further, e.g. with respect to co-occurring drugs or the indication or severity of medical problems. While the initial event-type analytics was focused on single patients, later smaller and even larger groups of patients could be analysed by a combination of query methods and event simplification strategies. Advanced visual query methods have been developed for such temporal event data based on temporal logic by Allen and Ferguson [AF94] who described the 13 unique relations between two intervals. Examples for such query methods include [GWP14, MLdO*13, WLB*17, ZGP15]. These methods address the specification of an initial query, the presentation of the results and the refinement of a query in a convenient manner (without the necessity to completely reformulate). Graphical methods that present the temporal relations in an intuitive manner are considered useful by epidemiologists [MLdO*13]. While temporal event specification is essential in clinical health care and some branches of PH, it is less important for cohort study data with a few points in time only.

4.3. Analytical methods for subpopulation definition

Vulnerable subpopulations with a strongly increased risk for a disease are interesting. Thus, when the risk of a subpopulation differs strongly from the global mean, epidemiologists want to understand the features that characterize such subpopulations. Analytical methods may reveal such subpopulations where purely interactive methods are not powerful enough. Hrovat *et al.* [HSKO14] and Niemann *et al.* [NSVK14a] employed *association rules*. Niemann *et al.* combined them with decision trees that were automatically computed yielding a hierarchy of split attributes. Alemzadeh *et al.* [AHN*17a] employed subspace clustering. All three methods are steered by some input parameters and yield quite complex results. Interactive visualization is required to support the adjustment of the input parameters and inspect the results.

Association rules may characterize subpopulations with certain characteristics, e.g. an increased risk, and were therefore used for PH data [HSKO14, NSVK14a]. Association rule mining algorithms, such as Apriori, effectively search for all association rules. To reduce the amount of rules, the search may be restricted, e.g. to rules with minimum support (ensuring that the number of affected persons is not too low) and a minimum confidence (ensuring that the rule leads to a correct conclusion in the majority of persons). Since even with such filters, often many rules result, they may be sorted according to further interestingness measures, such as the maximum length of the rules. Typically, PH academics are interested in rather short rules avoiding a complex description of a subpopulation. Association rules may be shown as glyphs in a scatterplot where the position in x and y directions, colour and size visually represent up to four interestingness measures. The user may select rules to inspect them in detail in a second view. This particular visualization is only a proof-of-concept and not widely used. We include it since association rules were used in data mining research related to epidemiology.

For time-dependent data, *temporal association rule mining* may represent how subpopulations change over time. Temporal association rule mining was employed by Hrovat *et al.* [HSKO14] to analyse about seven million hospital discharge data, including up to 15 diagnoses along with demographic data. Niemann *et al.* [NSVK14a] present the InteractiveMedicalMiner that combines decision tree classification with association rules and presents the results graphically to support the adjustment of parameters for decision tree and association rule computation. Their system was applied to a subset of the SHiP data (recall Section 3.1).

Clustering is frequently applied to determine groups of patients or participants who are similar to each other and dissimilar to others. Clustering is a useful technique if the data do not involve too many dimensions. Kwon *et al.* [KEV*18] described a clustering method supported by various visual aids to enable an understanding of alternative clusterings and adjust parameters in a goal-directed method. In a case study, they applied their method to clinical data of patients with heart failure. Clinical scientists used this method to extend their initial selection of three patient groups, e.g. one group with obese patients and diabetes and one group with elderly patients chronic kidney disease. If an epidemiologist manually breaks down a larger set of data to a rather low number of dimensions, this technique may also be useful in PH.

Subspace clustering is useful for analysing truly high-dimensional data (>20 dimensions) where global clustering is not promising due to the curse of dimensionality. For PH data, subspace clustering is beneficial, since persons are likely to be similar in some dimensions but not in all. Subspace clustering is typically a two-stage process: clusterable subspaces, where some regions have a very high density are identified in the first stage and a clustering method, such as DB-Scan [EKS*96] or Optics [ABKS99], is applied to these clusterable subspaces in the second stage. As an example, for the first stage, Kailing *et al.* [KKKW03] provide a ranking of subspaces and enable the analyst to choose, for example, the top N subspaces for clustering.

Subspace clustering involves several parameters that influence the search and ranking of subspaces as well as the clustering within the selected subspaces. Thus, Niemann *et al.* [NSVK14b] considered the application to cohort study data as not advisable since the results are too sensitive to a number of parameters. As a consequence, the same group developed a constrained-based technique, where the clustering is guided by a small set of constraints, given by an expert [HNP*18]. As an example, for a few pairs of participants diagnosed with fatty liver, the expert specifies that these participants must be in the same clusters, whereas some other pairs of participants are forced to be in different clusters since one in each pair is diagnosed with the disorder and the other is not. This semi-supervised subspace clustering turned out to yield relevant results for epidemiologists [HNP*18].

Alemzadeh *et al.* [AHN*17a] used this semi-supervised subspace clustering for the high-dimensional data of the SHiP. They described the visual exploration of subspace clustering results for the SHiP data (recall [JHLea01] and Section 3.1). As an overview, subspace clusters are displayed in a 2D view where multi-dimensional scaling was applied to map the similarity between the clusters to spatial proximity (Figure 5, left). Similarity for subspace clusters relates to the overlap between the dimensions and the instances of subspaces [AKMS07]. As an example, age is a dimension that contributes to different subspace clusters. Alemzadeh *et al.* [AHN*17b] discuss the use of this method in an extensive case study, involving the validation of the determined subpopulations in an independent cohort.

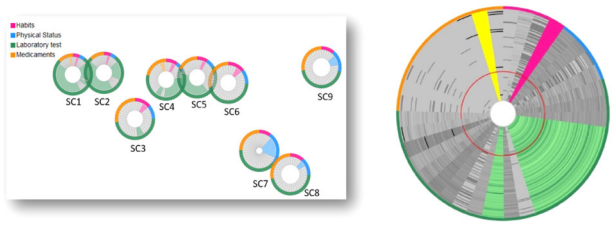

Figure 5: *The results of subspace clustering applied to cohort study data are shown in an overview (left). Each subspace cluster is shown as a donut where donuts with a larger inner circle (hole) represent clusters with few members only. Grey values represent dimensions that do not contribute to this subspace cluster. The four colours represent dimensions of different categories, e.g. medication and laboratory values. The detail view (right) reveals information on the participants. Darker colours represent greater values and black represents missing values (from: [AHN*17a]).*

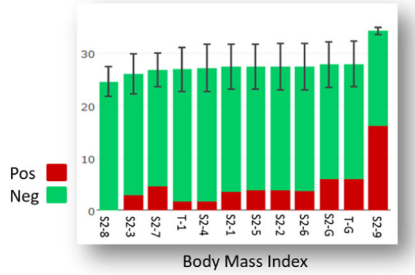

Figure 6: *For all subspace clusters, red indicates the portion of participants with a positive outcome, e.g. fatty liver, and green represents the healthy participants. The y-axis indicates the average (and variance) of the body mass index (BMI) in these subpopulations. Members of subspace cluster S2-9 with a BMI index of about 30 have an increased risk for fatty liver. Such visualizations are provided for all dimensions that contribute to several subspace clusters (from: [AHN*17a]).*

For selected subspace clusters, details are presented in additional views (Figure 5, right). The colours represent different categories of the data, e.g. laboratory values, medication, physical status and habits. To explore a selected subspace cluster in detail, scatterplot matrices are also available (recall Figure 4, left). For supporting an overview, scaled bar charts were also employed. They reveal for a health risk, such as high blood pressure, the portion in the different clusters. Thus, it becomes obvious when a cluster (representing a subpopulation) exhibits a risk that is strongly increased compared to the global mean (see Figure 6). Such design decisions were based on discussions with epidemiologists with respect to good overview visualizations.

Subspace clusters may have arbitrary shapes. Epidemiologists, however, prefer hyper-rectangular clusters, such that a subpopulation may be described by a set of intervals, related, e.g. to some laboratory values, the body mass index and alcohol consumption. A hyper-rectangle is the generalization of a rectangle to N dimensions. Alemzadeh *et al.* [AHN*17a] therefore support the transformation from arbitrary shapes to hyper-rectangular clusters. Since epidemiologists aim at a verification of findings from any data mining technique (task T4), the authors also supported the validation of the subpopulations in an independent cohort.

4.4. Spatial epidemiology

The *spatial context* of health data, e.g. regional differences in demographics, social status, health risks or prevalence of diseases, is essential to understand spatial correlations. Spatial epidemiology or health geography, as it is also called, aims at an understanding of the link between environmental factors and human health, pathways of spatial distribution and resulting health risks [BAHJ08]. Spatial epidemiology is an instance of geo-visual analytics and thus also has its roots in cartography. Atmospheric pollution and the consequences of climate change are some examples for topics addressed in spatial epidemiology [JGK10]. *Spatial modelling* comprises visualization, exploration and statistical analysis of geo-referenced health data and aims to assess whether certain deviations of local health risks are significant and require actions.

Map data are needed to relate information to geographic destinations accurately. As a consequence, geographic information systems (GIS) may be enhanced to provide support for PH tasks that involve spatial information. Jerret *et al.* [JGK10] give an overview involving aspects, such as meteorological dispersion of pollutants and behavioural changes that interact with each other and influence health risks. Again, finding a strong correlation of local regions with a health risk may not represent a causal relation. Various factors, such as pollution, infectious pathogens but also genetic susceptibility vary locally and also interact with each other.

Disease mapping. Map-based views are useful if there is a rather constant background risk for getting a disease and a peak frequency is likely attributed to a source of contamination. Maps are also used to detect *disease clusters*, i.e. frequent co-occurrences of diseases. Elliott and Wartenberg [EW04] mention as an example that the spatially increased occurrence of infectious diseases, such as Hepatitis B, and some types of cancer lead to the hypothesis that these infections increase the risk of cancer. An important area is *disease mapping* where the local frequency of diseases is displayed to identify regions with excessive disease load. Elliott and Wartenberg [EW04] mention examples, including atlas data related to cancer mortality or more general to causes of death.

Layer-based visualization. Map-based data are often represented in different *layers*, e.g. a background layer with major cities, rivers and administrative borders and various layers that may be combined with the background layer, e.g. locations of hospitals, diseased persons or water reservoirs as possible sources for infectious diseases. An overlay of different layers is typically superior to side-by-side displays of the individual map layers, as already discussed by Jaques Bertin [Ber66]. Luz and Masoodian [LM14] discuss the use of a semi-transparent foreground layer to improve the interpretability of the map-based data. They use examples from epidemiology to discuss the appropriateness of three transparency levels depending on the background complexity. Colours are widely used in all types of map-based visualization, including health geography. The ColorBrewer [HB03] provides guidance for a careful selection of colours for map displays. It is also available online.

Malaria Atlas project. As an example for a global activity, the Malaria Atlas project aims at integrating and communicating information on parasite rates, parasite types and epidemiological data, e.g. population at risk, malaria morbidity and mortality [GHL*07] (recall Figure 1). The density of information sources is quite diverse and a map-based visualization conveys this essential information. Furthermore, various charts are potentially useful, e.g. how many sources of parasite information are available per country or the land coverage (forest, rivers, cities, etc.). Spatial queries, e.g. the search for the k nearest neighbours, often are supported [AAA*16]. In conclusion, Jerret *et al.* [JGK10] stated that GIS and related systems are established in PH. Also, Beale *et al.* [BAHJ08] give an overview of tools that support mapping health data along with spatial statistics. In this subsection, we discuss general aspects of spatial epidemiology. In Sections 5 and 6, we shall discuss specific examples, e.g. related to infectious disease outbreak and air quality surveillance.

4.4.1. *Data*

Spatial epidemiology relies on detailed geo-referenced health data along with precise population data. Only if these two sources of information are accurate, summary measures per area, such as standardized mortality rates (SMRs), are meaningful. SMR, the number of cases in a study group divided by the number of cases from the general population, is one of the essential characteristics of the local dispersion of a disease. The underlying spatial data, e.g. various health indicators, environmental factors and population data, are rarely independent. Thus, spatial autocorrelation has to be taken into account [BAHJ08].

4.4.2. *Small area epidemiology*

Elliott and Wartenberg [EW04] discriminate *small-area statistics* and more global statistics. Small-area statistics, e.g. local indicators of spatial association, are employed to understand the effect of socio-economic differences or pollution on health. Small area statistics identify and summarize regions with an unusual incidence of diseases. Bayesian models and generalized mixed linear models are frequently applied to study spatial effects taking into account confounding variables, e.g. the effect of air pollution on respiratory diseases with smoking behaviour as confounder [JGK10]. Small-scale spatial epidemiology is sensitive to data protection issues. The publication of small-scale information related to pollution and disease frequency, for example, may influence property values and thus may be against the interest of the related population.

4.4.3. *Visualization techniques*

In this subsection, we introduce map types that are widely used in PH.

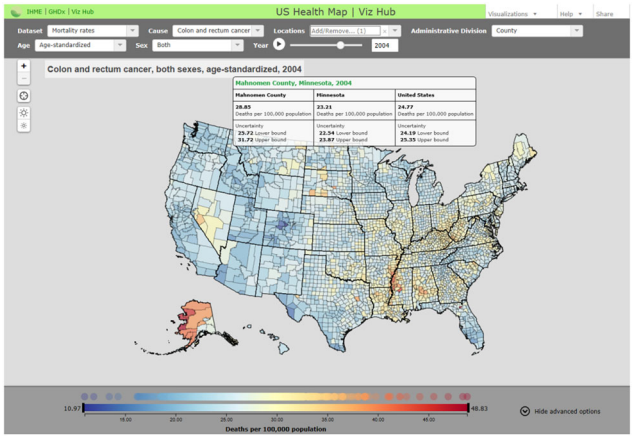

Figure 7: *A choropleth map indicates the mortality due to colon cancer on a county level. The maps can be generated for a certain year and for different age groups. Users may see details, including uncertainty for the selected county as tooltip (Screenshot generated from http://www.healthdata.org/data-visualization/us-health-map at 20/7/2019).*

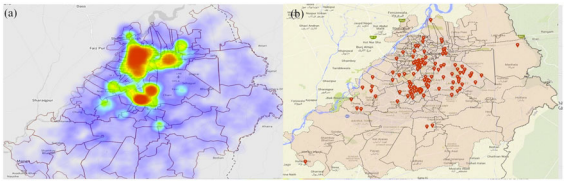

Figure 8: *A heatmap (left) shows the cases of haemorrhagic fever in Pakistan with a hot spot in the northeast. The right view contains markers for the most recent cases (from: [AAA*16]).*

Choropleth maps are widely used to visualize area-based geographic information along with associated data, such as influenza cases in a certain age group per region (see Figure 7). The data underlying a choropleth map are an aggregation within an administrative unit, such as a county. However, administrative units are often not ideal for such aggregation, since they may strongly differ in size and do not necessarily represent physical boundaries [BAHJ08]. Choropleth maps are sensitive to misinterpretation in case of incomplete or sparse data [CAD*14]. For example, they may be misleading in case of smaller regions with a low absolute number of diseases. In such cases, Castronova *et al.* [CCN09] suggest to re-aggregate the data, i.e. to merge adjacent districts until a significant number of cases is achieved. However, aggregation bears the risk to aggregate low- and high-risk regions and thus ignore this information by averaging. Elliott and Wartenberg [EW04] as well as Beale *et al.* [BAHJ08] discuss *smoothing* methods that create *interpretable risk surfaces* to avoid wrong conclusions based on small numbers.

Heatmaps are also used to display area-based geographic information. In contrast to choropleth maps, they assume continuous data over a surface and provide more precise information (see Figure 8). Discrete or continuous colour scales are used to map nominal or quantitative data. Since data are typically not available for every

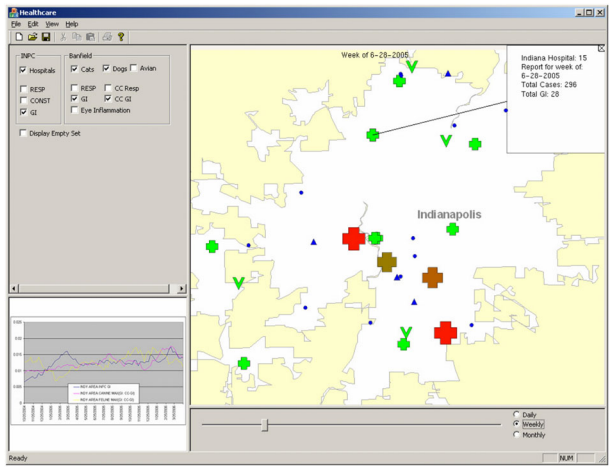

Figure 9: *A map view is part of the LAHVA system. The spatial distribution of emergency departments (red crosses) and veterinary hospitals (green crosses) is shown in a dotplot. The size of the glyphs represents the number of cases. In the lower left view, the temporal changes of cases is shown (from: [MTJ*07]).*

Table 1: *Mapping techniques and their properties.*

Technique	Sparse	Dense	Continuous	Discrete
Choropleth maps		x		x
Heat maps		x	x	
Isopleth maps	x		x	
Dotplots	x			x

unit of the map, e.g. in sparsely populated areas, missing values are often interpolated. Jerret *et al.* [JGK10] discuss various interpolation methods for spatial data and their properties. Maciejewski *et al.* [MRH*08] employ kernel density estimation, in particular a variable-sized kernel that adapts to the population density. This estimation improves heatmap generation and was used for the assessment of alerts from an outbreak detection algorithm.

Isopleth maps represent continuous 2D data using isolines. In spatial epidemiology, they are used frequently [JGK10].

Dotplots represent point-based health data, e.g. positions of health care institutions or cases of a severe disease [CAD*14] (see Figure 9). In contrast to choropleth and heatmaps, dotplots and isopleth maps are sparse representations that leave room to display the underlying geographic data. Like a scatterplot, a dotplot may suffer from overplotting. Additive opacity may reduce this effect [MRH*08]. The exact depiction of patient data in a dotplot at a map with street-level scale involves privacy issues. Therefore, aggregating, e.g. at zip-code level, is recommended [MRH*08]. Table 1 summarizes essential properties of these mapping techniques.

Multivariate maps. Spatial epidemiology comprises several aspects, such as susceptibility of persons, exposures to risk and actual diseases. Stacked multivariate maps, where each layer represents one of these aspects [LHM*07], integrate different types of information. To avoid visual clutter, layers of information can be displayed

or hidden. The combination of two dense mapping techniques is at a greater risk of visual clutter. The simplest multivariate maps represent for two variables the binary state whether or not a threshold is exceeded. Thus, an exposure and disease frequency may be mapped with appropriate colours. MacEachren *et al.* [MGP*04] use such visualizations, e.g. to compare regions with respect to the income level (low/high) and the frequency of diseases, such as acquired immunodeficiency syndrome (AIDS) (normal/elevated frequency). DiBiase *et al.* [DRK*94] give an overview of multivariate map displays.

Focus-and-context visualization. Chen *et al.* [CRN*08] present maps where only a circular region contains sharp high-contrast information, while the remaining map is shown blurred and with low contrast. Such a focus-and-context visualization may enable the focused analysis of regions of interest. The term *focusing* in the context of health maps typically means that only regions where a variable exceeds a threshold are mapped to colour, whereas other regions are shown in grey [MGP*04].

4.4.4. *Uncertainty quantification and visualization*

Uncertainty quantification and visualization is widely discussed in cartography and geo-visualization, see, e.g. the book by Zhang and Goodchild [ZG02]. Thus, also most papers on spatial epidemiology discuss the uncertainty due to sampling variability, biased information or low absolute numbers, e.g. in case of rare diseases. Map-based visualizations often include interpolation, simplification or binning of data—transformations that may reduce precision and affect interpretation. Appropriate uncertainty visualization may increase trust in the data and has an influence on decisions, e.g. in health policy. Uncertainty in map-based data occurs with respect to positions (*location uncertainty*), attributes (*value uncertainty*), completeness and time [MRH*05]. The combined influence of these uncertainty types may considerably affect trends detected in spatio-temporal analysis.

Uncertainty quantification. Most spatial analysis in PH is related to population data, e.g. the incidence and prevalence of diseases per 100 000 inhabitants (recall Section 2). This requires up-to-date and reliable population data—a requirement that is often only partially fulfilled in developing countries or in case of stronger recent migration. Spatial statistics provides a reasonable basis to quantify the uncertainty in map-based data [ZG02].

Confidence intervals should be computed, in particular for all area-level statistics [BAHJ08]. This information must be conveyed at least as temporarily available information, e.g. via tooltips. Confidence intervals assume approximate normality of the data. This assumption is typically not fulfilled for sparse data. In this case, a boxplot with interquartile ranges may better capture the uncertainty. Fuzzy set theory and probability theory are also used to model and quantify uncertainty in spatial data [MRH*05].

Uncertainty visualization. Various methods were introduced to display the uncertainty and thus to reveal local differences. Most of them depict the most likely interpretation of the data along with its uncertainty. As an alternative, different interpretations of the data may also be displayed in sequential frames or combined in an animation [Fis93]. User-adjustable uncertainty thresholds are also used to restrict the display of map-based data to regions where the certainty exceeds the threshold [HM96]. Uncertainty visualizations (in maps) may be *intrinsic* such that the presentation of the data is adapted or *extrinsic* where additional symbols (uncertainty glyphs) are added to the display [MRH*05].

Beale *et al.* recommend to analyse map displays with and without smoothing to understand the effects of statistical smoothing. Monmonier [Mon06] suggests to consider uncertainty as a second dimension and encodes it with bivariate choropleth maps (an intrinsic visualization). Uncertainty in maps is often mapped to the saturation of a colour, the transparency or blur [MRH*05].

The interpretation of map-based views depends on the chosen colour scale and on the spatial resolution, e.g. how spatial districts are summarized. By changing these parameters, different valid interpretations may arise [EW04]. Also, Chen *et al.* [CRN*08] argue for analysing and displaying health data at different scales to derive valid conclusions. They introduced *reliability maps* that indicate the certainty of the statistical measures. They employed spatial clustering with slightly different parameters to identify hot spots—regions with an elevated risk for some type of cancer. While some clusters are determined reliably with widely varying parameters, others are sensitive to small parameter changes.

Evaluation. The effect of uncertainty visualization on the map users' perception, interpretation and actual decision-making needs to be analysed in evaluations. Intuitiveness and trust in decisions are criteria in such evaluations, see MacEachren *et al.* [MRO*12] for an example of such an evaluation.

4.5. Temporal visualizations

The temporal aspect of health care data is often displayed with timelines. In contrast to patient-specific and discrete event-based clinical data, PH data are aggregated for populations and primarily continuous, e.g. the number of cases for reportable diseases is continuously monitored. Timelines are typical components of dashboards for PH experts. Several time-dependent data may be shown simultaneously to enable a comparison, e.g. the number of cases for different diseases, or the number of influenza cases along with the number of hospitalized influenza cases. Temporal visualizations may be guided by the restriction to certain regions in a spatial context.

4.5.1. *Epidemic curves*

Timeline-based visualizations of diseases-related data are often referred to as *epidemic curves* or short *EpiCurves* [LRS12]. They have a characteristic shape depending on the reason for an outbreak. One peak is typical for poisoned food that is consumed at one point in time, whereas for a continuous contamination, like a water pump, the number of new cases decreases slowly. EpiCurves, representing infectious diseases, typically show an increase for a longer period. The scaling of such curves is important. To display temporal intervals with low and very high values simultaneously, e.g. influenza cases, logarithmic scaling is recommended [CWCN11]. The temporal axis should be properly labelled. Since most surveillance data are collected on a weekly basis, labelling based on weeks is favourable [CWCN11]. For the exploration of longer time series, temporal zoom and stronger aggregated values are possible.

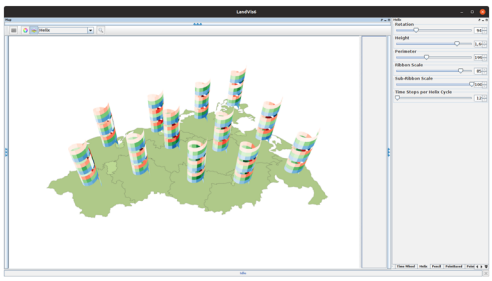

Figure 10: *Helical icons indicate the incidence of influenza in different months in different districts of a part of Germany (Courtesy of Christian Tominski, University of Rostock).*

4.5.2. *Static visualizations*

Calendar-based views may be employed to display the incidence of diseases on a daily basis. A clock metaphor may be used to display the disease frequency for the 12 months of a year in different regions of a map display. Tominski *et al.* [TSS05] extended this idea towards helical icons (see Figure 10) that integrate the information from several years and can thus convey the periodicity of the data. In static visualizations, scalar values are mapped to colour, typically with green denoting low frequency of a disease and yellow, orange and red denoting higher frequency. Typically, time is considered along a linear 1D scale. However, the incidence of diseases may follow seasonal patterns and thus visualizations that consider the *periodic character* of time are useful [AAA*16, MTJ*07].

4.5.3. *Animation*

The second major type of time-based visualizations is animation, where one or a few variables are displayed on a map and an animation indicates changes over time. Animations of health-related data gained popularity by Hans Roslings' GapMinder that integrates animated bubble charts showing, e.g. how life expectancy and child mortality developed in different countries. Inspired by the GapMinder, Robertson *et al.* [RFF*08] analysed the potential and limitations of animations for trend visualization. In PH, animations are rarely employed. The Motion charts, introduced by Al-Aziz *et al.* [AACD10], represent a notable exception.

The analysis of spatio-temporal patterns may benefit from *dynamic maps*, i.e. animations that enable an observation of changes of relevant variables in space and time. Castronovo *et al.* [CCN09] applied principles from cartography animation [Har03] in order to generate dynamic maps that do not overwhelm the analyst. Fabrikant and Goldsberry [FG05] emphasize the necessity to quantify the magnitude of change between different frames to adjust speed and duration of an animation. They suggest to carry out eye tracking studies to understand which changes are detected. They also suggest to employ a model of *perceptual salience* to predict how attention is focused on different parts of a map and eventually adapt the map display such that relevant changes are likely perceived.

Dynamic maps may be used to compare developments, e.g. the rate of Salmonella infections with the temperature development or the concurrent development of morbidity and mortality related to a disease. Dynamic maps may reveal relationships that are not recognizable in static visualizations. However, such animations need to be observed several times to enable an appropriate interpretation [CCN09]. Dynamic maps are appropriate for data with a high temporal frequency and require a careful selection of the temporal scale. Users may scroll in the temporal domain and thus select a single point in time. They may also advance time incrementally with a certain step size, e.g. a week, and restrict the dynamic map generation to a temporal interval [MRH*08]. Castronova *et al.* and Harrower [Har03] discuss principles for dynamic map generation as well as its application for environmental health data. This includes a discussion of the complexity, comprehensibility and confidence in the observations gained from viewing such animations.

5. Analysis and Control of Epidemics

This is the first from three sections to discuss a particular class of PH problems and solutions. The outbreak of infectious diseases is a severe problem leading to rigorous surveillance activities in PH institutions. Worldwide infectious diseases are the second most frequent cause of death (25%), only exceeded by cardiovascular diseases [Gon00]. An epidemic is defined as an excess of a severe illness far beyond normal expectancy [MV13]. Epidemics relate to communicable diseases. While air-borne infections dominate in the Western world, water-borne and vector-borne infections, such as Malaria, are most frequent in developing countries. Infectious diseases may have devastating effects, such as the influenza epidemic in 1918 that killed more people than World War 1. More recent examples include the human immunodeficiency viruses (HIV) endemics and SARS as a special type of flu. The 1918 epidemic serves PH authorities to prepare for worst-case scenarios [MLR*11].

Due to technical developments and the increased mobility of people as well as demographic changes (urbanization), the potential for a rapid spread of infectious diseases has increased [Gon00, GLG*12]. To assess an outbreak of a tropical disease, such as dengue fever, a multitude of information needs to be combined. Masoodian *et al.* [MLK16] mention the geographic distribution of human populations, patterns of land use, location of forests and water reservoirs and weather information along with disease case reports.

Often, streaming data from various sources are analysed in real time to identify an outbreak as early as possible [AAA*16]. Maciejewski *et al.* [MRH*08] mention symptoms reported by patients in EDs as essential to detect outbreaks before a large number of confirmed diagnoses are available. Diseases may not only spread locally, but connections between cities and the flow of people between them may be relevant. Therefore, any forecast needs to be compared regularly to currently available data. The likelihood for the outbreak of infectious, air-borne and water-borne diseases depends on seasonal patterns and actual weather conditions [AAA*16]. Thus, syndromic surveillance requires to monitor multiple streams of heterogeneous information. *Outbreak alert* algorithms are used, but produce many false-positive alarms, since at an early state, an outbreak is hard to discriminate from natural variations of disease frequency. Visual analytics solutions have the potential to effectively analyse such alerts, including hypotheses generation and testing [MRH*08].

The simulation of epidemics primarily serves the support of decisions about interventions. Interventions include

- *pharmaceutical measures*, such as vaccination, and
- *non-pharmaceutical measures*, such as information campaigns, school closures, travel restriction policies or cancellation of events that attract a lot of persons [Guo07].

These interventions are carried out to minimize the negative effects, e.g. travel between distant regions is restricted to avoid that the disease spreads in a so far unaffected region, whereas the more frequent traffic within a spatial cluster may remain unaffected [Guo07].

While the analysis of cohort study data leads to *descriptive models*, the control of an epidemic requires *generative simulation models* for predicting the further development [MV13]. Simulation results include the number of persons who are expected to get ill, to get hospitalized or to die. A test bed for experiments enables simulations under different scenarios to understand the effect of various interventions. The temporal aspect is crucial, i.e. the question of how much time is available for a certain measure to be effective. For example, even if vaccination is possible, the question is whether a sufficient amount of vaccine may be supplied fast enough.

5.1. Interactive visualization

The most straightforward support for the analysis of epidemics is an interactive visualization of all relevant information with support for overview and detail visualization along with filtering mechanisms according to Shneiderman's mantra [Shn03].

Cooperative visual analysis. Masoodian *et al.* [MLK16] described interactive visualizations with a focus on tropical diseases. Geo-referenced disease data are displayed on cartographic and satellite maps. The maps are annotated with comprehensive case reports, including information on patient demographics and housing conditions. The speciality of their nu-View system is the support for synchronized co-located collaborative analysis, i.e. several users analyse the data together at the same place. Based on a careful task analysis, they designed the integration of a shared display and private displays, where parts of the private display may be shared. While this system provides strong cooperation support, it does not include any simulation or prediction of the course of the outbreak.

Mapping travel route information. A system, introduced by Dunne *et al.* [DMPM15], provides visualization support focused on travel routes at different scales since these routes are potentially important to understand diffusion patterns. The authors combine different visualizations: global flight connections are displayed as arcs but also travel routes at lower scale, in particular daily commuting patterns, are displayed. Borders between communities that belong to frequent community pathways are emphasized and (automatic) labelling the resulting visualizations is also discussed. A special feature, motivated by the goal to better recognize detail, is the transformation of a map with community borders to a Voronoi tessellation that provides sufficient space to embed symbols for each community, e.g. to encode the population size.

5.2. Simulation of spreading

The simulation of outbreaks is a large research area on its own [MV13]. We only touch this area to understand the interface of such simulation engines, namely the input and output space. Simulations are based on assumptions and input parameters. As an example, communicable diseases are characterized by

- a *transmission rate*, i.e. the likelihood that the disease spreads to healthy persons,
- an *incubation time*, i.e. the time after infection until the disease leads to symptoms,
- an *infection time*, i.e. the time after an infection when the patient may spread the disease and
- the duration of the disease until the patient is cured or died.

The infection time often starts before the patient develops symptoms, e.g. within the incubation time. The incubation and infection times are modelled as a probability distribution. For mosquito-borne diseases, for example, the local differences in mosquito count are essential but can only be guessed based on land coverage data (parks have a higher count than office buildings) [BWMM15]. For infectious diseases, differences in the local population density affect the development. Such parameters are not precisely known and thus need to be estimated, e.g. based on earlier outbreaks.

Modelling mobility and transmission. The mobility of people and the contact between individuals in geographic space determine the spread of a communicable disease. As an example for a simulation model, Eubank *et al.* [Eub02] modelled the course of an epidemic by a *contact graph* that represents people (as nodes) and whether they had contact (as edges). This results in a graph representation of the social network of persons.

Activity graphs represent locations that are connected when people (frequently) move between them. This information is integrated in a *diffusion* model that represents how the disease spreads over time. Such a diffusion-based simulation can be applied to a wide range of diseases. Simulations aim at identifying specific locations, e.g. restaurants, or age groups that are part of a *critical transmission path*. While such a simulation allows for accurate modelling, it requires a lot of information that is typically not available. Still, Eubank *et al.* demonstrated its general feasibility, using diagrams to convey the results, e.g. the number of infected and contagious people over time, the age distribution and properties of the contact graph. Guo *et al.* [Guo07] employed the simulation model from Eubank *et al.* for a large-scale simulation involving 1.6 million people. Based on data from the Bureau of Transportation statistics, they achieve a reliable estimate of traffic, e.g. how many persons move between certain locations. Although the data are aggregated, it helps to achieve realistic simulations.

Visualization. The simulation of spreading with detailed transportation statistics and *contact graphs* yields huge data that need aggregation to be displayed. Flow maps are a useful technique for displaying the movements of people. For scalability, locations may be spatially clustered or only considered if the number of people moving between them exceeds a threshold [Guo07]. The display of an activity graph also requires strategies to prevent visual clutter. Guo *et al.* suggest to employ matrix-based visualizations as well

as node-link diagrams and partition them according to spatial proximity in roughly equally sized subgroups of manageable size. The visualization aims at identifying spatial disease clusters in order to consider interventions to limit this influence.

5.3. Predictive analytics for the simulation of outbreaks

Simulation results critically depend on the choice of input parameters, such as the transmission rate. Since these parameters are estimated, they carry some uncertainty. Instead of simulating the outbreak only once with precise input values, multiple simulations runs with slightly changed input parameters reveal the space of possible developments [BWMM15]. The resulting ensemble data reflect the temporal and spatial development. Bryan *et al.* [BWMM15] discuss how the large variety of results that arise from varying several input parameters can be explored for assessing the resulting predictions. Due to performance reasons, not every configuration of input parameters is actually used for a full simulation. Instead, interpolations are employed to *emulate* with predictive views that enable an exploration and comparison of different simulation runs. As a case study, they present the simulation of a mosquito-borne disease in the Washington DC area with about 500 000 persons involved. The simulation results to be explored are multi-variate and spatio-temporal.

The EpiCanvas. The EpiCanvas [GLG*12] provides comprehensive information related to the development of regional infectious diseases. The focus is on an integrated display of the various streams of information. Since their design is largely inspired by weather maps, they refer to their system as 'EpiCanvas infectious disease weather map'. The central idea is to design a tag cloud visualization that connects relevant concepts (see Figure 11). The tag cloud in the concept view can be steered by selecting tag groups and specific tags relevant for the current task. The concept view supports task T3, namely to show associations (recall Section 2.3). For the layout, various graph drawing algorithms are provided.

The size of a tag intuitively shows its frequency, e.g. of gastrointestinal pathogens in a certain temporal interval. The EpiCanvas also contains time-series graphs to display how diseases developed over time. The system is based on ED visit data from a children's hospital in Utah [GLG*12]. Data for a 1-year period was available, representing different outbreaks of respiratory and gastrointestinal infectious diseases. In total, 44 848 datasets were tagged with one to eight pathogens. Demographic data, including location (zip code, city) and age group, were available in addition to symptoms and microbiological test results.

The evaluation with 10 expert users involved the free exploration, e.g. with different temporal intervals. The following questionnaire was based on the *unified theory of acceptance* [VMDD03], i.e. standardized questions and scales were used to analyse whether the target users intend to use such a system.

Infectious disease viewer (idViewer). This system is used to process real-time streaming data from ED visits, emergency calls and drug sales to detect outbreaks of epidemics in Pakistan [AAA*16]. An essential component is the classification of symptoms and complaints with respect to the most likely disease. Since in Pakistan, the likelihood for diseases depends on seasonal patterns, the seasonal character (rainfall and temperature curves) is taken into account

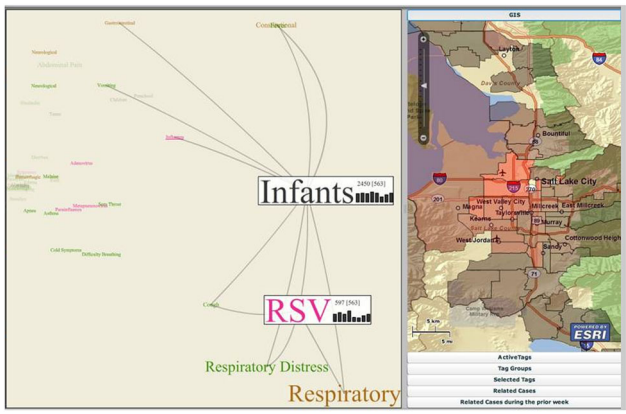

Figure 11: *The* EpiCanvas *can be configured with respect to a temporal interval and map display options. The central part contains a tag cloud with 'infant' and 'respiratory syncytial virus' selected. The choropleth map displays the local incidence of the disease (from: [GLG*12]).*

along with the symptoms for the classification of a disease. This input is processed by a neuronal net. Taking into account, the weather information increased the overall sensitivity for outbreak detection from 89% to 94% for nine major diseases. As an example, a set of symptoms that may relate to dengue fever in summer is much more likely to indicate a respiratory disease in winter. The frequency of diseases is colour-coded in heatmaps (recall Figure 8). A prediction component is incorporated to directly support response management. At the time of the publication, the authors reported on a 5-year period where the system was already in routine use.

5.4. Zoonotic diseases

Animal diseases may be important to human health if the companion pets of a human are affected, e.g. by influenza. Thus, data related to diseases of cats and dogs may be analysed to create warnings for a human outbreak. The linked animal–human health visual analytics (LAHVA) system was a pioneering work with respect to the joint analysis of human and animal health data [MHR*09, MTJ*07]. They employed ED visit data, data of pet owners and data from veterinary hospitals, e.g. related to respiratory diseases of dogs. The LAHVA system (recall Figure 9) is an example for the fruitful combination of a statistics component and a visualization component. The number of cases over time is presented as timeline, to which data transformation, i.e. logarithmic transform, was applied. Diseases are categorized as respiratory, gastrointestinal and eye inflammation. Seasonal trends were detected with seasonal trend decomposition. A spatial view shows the positions of EDs and veterinary hospitals with glyphs that are scaled according to the number of cases in the selected temporal interval. Respiratory symptoms occur in dogs approximately 10 days earlier than in accompanying humans.

5.5. Training of outbreak response

PH experts cannot rely solely on their personal experience with previous outbreaks. Severe outbreaks are rare and do not represent a learning opportunity. Similar to the education of pilots and surgeons, training environments are essential to learn how to respond in case

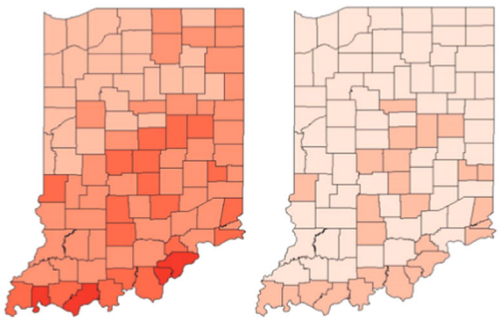

Figure 12: *The course of an influenza outbreak near Chicago is simulated on a national level. The number of diseased persons in various parts of Indiana is displayed as choropleth maps and compared for a situation without any intervention (left) and with media alerts (right) (from: [MLR*11]).*

of critical situations, complications or failures of devices. Robinson *et al.* [RMR11] described a user needs study among PH experts. The need for the training of specific workflows was emphasized. For outbreak control, comprehensive training tools are essential. Such tools should provide realistic data, e.g. based on historical outbreaks and allow to simulate the unfolding of an outbreak. They should also include various decision points, where PH experts can enable interventions, such as contact tracing and vaccination of close contacts in case of new infections. Storing these decisions and 'replaying' the simulation eventually with alternative decisions is a powerful means to train appropriate response. A visual analytics framework is required to steer the simulation and to observe and analyse the effects of interventions. In the following, we describe two training systems developed to improve *preparedness* of PH authorities (recall task T8, Section 2.3).

PanViz. The PanViz system [MLR*11] supports state and local communities to be prepared to a severe outbreak of influenza. This visual analytics toolkit analyses the effects of measures implemented during a simulated pandemic influenza scenario. Age distribution and population density are taken into account when simulating a pandemic starting at a certain origin. PH experts can explore the effects of the pandemic on the population. Major traffic routes are considered, e.g. the traffic between the 15 most important airports. Based on this information, the spread of an outbreak on a national level may be predicted. Once the disease reaches the nearest airport, it spreads on the following day along the airway connections. The core component is an interactive spatio-temporal view, where users can move through time and insert decision points. Thus, the impact of decisions can be displayed and compared to a situation without any intervention (see Figure 12). Major quantitative results of the simulation (number of sick, hospitalized and dead persons) are linked to the visualization allowing users to flexibly adjust parameters and enable intervention measures. The tool was deployed as an educational tool.

Epinome. Livnat *et al.* [LRS12] present the Epinome training system. They incorporated the simulation of a pertussis outbreak using detailed representations of social networks and contact patterns. The underlying simulation is stochastic, i.e. multiple runs of the simulation lead to different results, thus representing uncertainty in the prediction. The visual analytics system combines a rather large number of views along with considerable flexibility to adjust the layout, i.e. add or remove views depending on the suitability for the current task. In addition to map-based views, EpiCurves, and a list with the latest new cases are displayed for an in-depth analysis. The authors discussed the value and potential problems of coordination between views. An elaborate filtering mechanism was included to enable a focus on geographic regions or demographic groups. However, the views are *loosely* coordinated, i.e. only summary views are adapted, when the user brushes a region. Too many adjustments after filtering were found to be confusing.

Table 2 summarizes essential visual analytics systems for outbreak detection.

6. Visual Analytics for Epidemiological Research

In this section, we give an overview of visual analytics solutions for epidemiological research—a second category of important PH problems. The target user group for the solutions described in the following is *PH academics* doing exploratory analysis to generate new hypotheses in selected applications. In contrast, the solutions described in the past section address PH experts performing routine tasks or, in case of an outbreak, urgent problem solving. Since these situations are fundamentally different, we dedicate special sections to both.

The solutions cover a wide range from studying cognitive ageing, to the prevention of child injuries and the surveillance on air and soil quality. While some application areas are discussed more extensively than others, this does not imply that these areas are more important but that visual analytics research—according to our paper selection strategy—more often tackled this problem.

6.1. Study of cognitive ageing

The high prevalence of neurodegenerative diseases motivates ageing studies, i.e. an understanding of normal and pathologic changes in the brain in elderly persons. This understanding may help to identify persons with a high risk to develop neurodegenerative diseases earlier. This type of research belongs to *neuroepidemiology* (recall Section 2.1).

Besides basic demographic data, cognitive and psychologic tests as well as MRI data are typically used to study age-related effects. MRI data incorporate structural imaging to assess different compartments of the brain with respect to volume and shape descriptors. Diffusion tensor MRI may be used for analysing white matter tracts with respect to various measures, such as fractional anisotropy, which is a measure for white matter integrity [AOH*14]. The Norwegian cognitive ageing study is based on the previous knowledge of morphological and structural changes of the ageing brain, e.g. related to cortical thickness or hippocampal volume [WFR*05, YELL10]. Comprehensive data from 100 healthy individuals aged between 50 and 84 were acquired.

Using the Norwegian cognitive ageing study, Angellini *et al.* [AOH*14] describe a visual analytics system (see Figure 13) that enables PH academics to efficiently test hypotheses and

Table 2: *Major visual analytics systems for outbreak detection.*

System	Key features	Key publications
ID Viewer	Combined analysis of diseases and weather conditions	[AAA*16]
Epinome	Ensemble simulation, map-based views, loosely coordinated views and EpiCurves	[LRS12]
Outbreak training	Data of historic outbreaks used for decision support	[RMR11]
PanViz	Use of transportation statistics, predict consequences of interventions, choropleth maps	[MLR*11]
LAHVA	Combined analysis of human and animal health, dotplots	[GLG*12]
EpiCanvas	Tag cloud visualization of concepts to study associations, graph drawing	[MTJ*07, MHR*09]
nu-view	Analysis of tropical diseases, collaborative visualization	[MLK16]

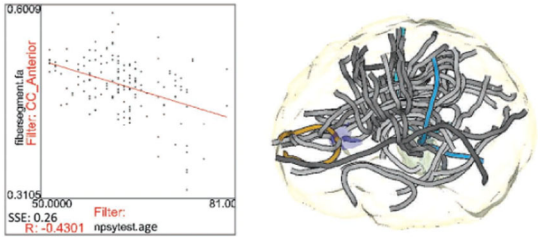

Figure 13: *A multiple coordinated view framework with a measure browser, histograms, scatterplots, parallel coordinates and a spatial view was provided to analyse age-related effects on the structure and morphology of the brain in the Norwegian cognitive ageing study. The scatterplot (left) shows the age-related decline of white matter integrity for the entire brain. The spatial view displays the brain and representatives for each white matter tract extracted from diffusion tensor imaging (from: [AOH*14]).*

eventually generate new hypotheses. The support for 'open-ended exploration' is considered the major requirement. The underlying data are stored in a hierarchical manner with the subject and demographic properties on top of the hierarchy. According to part of relations, the brain, different compartments in the brain, whiter matter tracts and their segments are further lower level components. The system provides a *measure browser* that provides a list of the available dimensions. Users can select measures and pairs of measures for which scatterplots are generated and correlation coefficients are computed. A certain analysis workflow may be stored and automatically applied to other parts of the hierarchical model. As an example, the loss of fractional anisotropy can be analysed for the entire brain but also for any compartment or white matter tract to better understand which regions are more stable over time. Based on such an analysis, PH academics can derive hypotheses about which brain functions, e.g. language-related functions, are more stable for elderly persons. An essential aspect of their system is that any selection is also visually represented, e.g. the analysed fibre tracts are also shown in a brain visualization. A brain atlas is used to assign brain compartments to the data of an individual patient.

The system was evaluated with a neuropsychologist and a neurologist. The experts gave feedback with respect to the usefulness of the system, the flexibility needed to fully answer their research questions as well as the efficiency. It turned out that the visual analytics solution indeed enables a much faster analysis of the data compared to the tools used so far. This efficiency is partially due to the specific database architecture that enabled a significant speed-up in access times compared to a standard relational database.

6.2. Food-borne diseases

Food-borne diseases typically represent an urgent PH problem [SIC*11]. However, we discuss long-term aspects, e.g. how agricultural methods and soil quality affect toxic concentrations of substances. As an example, we discuss efforts to reduce the risk of arsenic concentrations in food. This problem is particularly severe in some regions of India and Bangladesh where the arsenic concentration found in human samples was an order of magnitude higher than the highest values found in a control group in Australia. What Sims *et al.* [SIC*11] consider a visual analytics solution is actually a rather simple TABLEAU-based information visualization of arsenic food concentrations [JCIA10]. Arsenic concentration is colour-coded and the different types of food can be sorted and grouped flexibly. However, it is easy to envision a more powerful system that integrates analytic capabilities, such as clustering and time series analysis, to analyse how arsenic concentrations in food develop in different regions. Johnson *et al.* [JCIA10] introduced a TABLEAU-based solution that enables flexible filtering, e.g. related to different types of food. This paper also explains the health disorders that are associated with high arsenic concentrations in food.

6.3. Prevention of injuries

Injuries represent a major problem with striking local differences in the causes of injuries. An analysis of the injury situation, the identification of trends, gender and age differences is essential for improving injury prevention, e.g. related to playground design. Information about severe injuries and injury causes is recorded, e.g. in fire departments and crime reports. Based on such information, Martinezi *et al.* [MOSB16] described interactive visualizations analysing for different American countries and different states of the United States how the mortality related to injuries developed.

Al-Hajj *et al.* [AHFSP17] discuss child injury prevention and invited experienced stakeholders to a workshop to benefit from the spirit and dynamics of group discussion about the Canadian injury-related data. The decision for interventions, however, is difficult, since risk taking, risk perception and risk management are necessary parts of *normal* child development. Brussoni *et al.* [BBP*15] make an important distinction between *risks* that children voluntarily take and *hazards* that are not recognizable for children. If a play structure

breaks due to bad construction, this represents a hazard that needs to be avoided. As a consequence, standards were developed for playground design. We mention this discussion as an example for the need of a broad range of experts to consider all relevant perspectives including potentially harmful side effects.

Visual analytics support. Al-Hajj *et al.* [AHFSP17] designed an injury dashboard to monitor the injury situation in Canada. The dashboard gives an overview of all major aspects:

- A map view indicates the geographic distribution (the size of marks represents the frequency of injuries).
- Stacked bar charts indicate for girls and boys the amount of certain injury types, e.g. fractures or burns, further categorized in different age groups, e.g. 2–4 years.
- A temporal view shows how the frequency of different injury types developed over time.

The data may be flexibly filtered, e.g. with respect to the involved body part and the cause of the injury, e.g. transport-related and falls.

We briefly describe the evaluation of this system and encourage the reader to look in the original publication for more detail, since this evaluation covers many aspects of the system used in an integrated manner. The injury dashboard was used *cooperatively*: Visual analytics experts and domain experts, including trauma surgeons (surgeons specialized in the treatment of injuries), epidemiologists and policymakers jointly used the system to analyse data. The VA expert was needed to leverage the system's potential.

The session was audio- and video-captured and also the screen content was captured to analyse the interactions in detail. The analysis and interpretation was based on *distributed cognition*—an established theory to study cooperative cognitive activities, such as problem solving and decision-making [HHK00]. As a general result, the joint interaction with one visual representation was helpful to enable mutual understanding, including shared assumptions and beliefs. Patterns of cooperation could be identified that are potentially useful for other PH activities as well.

6.4. Pharmacoepidemiology

Drugs may control symptoms but they are often related to adverse effects. These effects have a high impact on the health care system, and therefore, pharmacoepidemiology, as discussed in Section 2.1, is an important branch of epidemiology research. This applies primarily if several drugs are taken together, if drugs are taken for a longer time and if patients are elderly with a reduced kidney function that is less capable of segregating drug components. Despite rigorous testing before drugs are admitted to regular use, rare but severe unwanted effects may arise and registration of such *adverse drug effects* is mandatory.

Adverse effects reporting. The Adverse Effects Reporting System of the Food and Drug Administration in the United States registered about 1.8 million ADEs in 2017 from which about 164 000 lead to death of the patient. It is not proven whether an ADE that is registered actually is *caused* by the drug. The system serves for detecting anomalies as a prerequisite for a detailed analysis of the cases and explanations of the ADEs. An ADE is stored with the following properties [MHD*14]:

- indication, i.e. the reason why the drug was prescribed,
- co-occurring drugs,
- adverse reaction,
- laboratory results and
- outcome, i.e. the severity ranging from reactions that require intervention to life-threatening situations and death.

The entries are categorized into effects that are mentioned in the product documentation (expedited) and non-expedited ADEs to which approximately half of the entries belong. The data are stored along with basic demographic information.

Data mining. Classic approaches for identifying ADEs rely on data mining, primarily the search for association rules with a minimum support and lift (recall Section 4.3) and Bayesian classifiers. As an example, Chazard *et al.* [CFB*11] identified 236 ADE detection rules related to about 115.000 hospital stays in France. However, rare events with a frequency only minimally above the expected frequency are hardly detected. Moreover, confounders often are not identified. Mittelstedt *et al.* [MHD*14] mention drugs that are prescribed for diabetes patients and that are associated with myocardial infarct. Myocardial infarct, however, is also associated with the natural course of diabetes.

Visual queries. Monroe *et al.* [MLdO*13] discuss a visual query approach to identify temporal events in a drug-related database to study interactions between drugs. They support the visual specification of interval-based events, e.g. the search for overlapping intervals representing drugs that were given partially at the same time and the search for point and interval-based events, e.g. whether a symptom (point event) occurred during an interval, where a drug was given. Also, the absence of an event can be searched for.

Visualization support. Mittelstedt *et al.* [MHD*14] introduce visual analytics-based hypotheses generation with an overview of drugs and adverse effects as well as interactive features to drill down, e.g. to search for co-occurring drugs for the same indication. Moreover, they support hypothesis-driven testing by an advanced query interface that enables the selection of drugs, adverse effects and temporal intervals that characterize when the adverse effect happened in relation to treatment time.

For the overview, the frequency of ADE is mapped to the size of circular glyphs, directing attention to more frequent ADE. The ADEs are colour-coded with respect to the severity. Concentric circles are generated to represent how often a drug leads to ADEs that require hospitalization, or even lead to life-threatening situations or death (see Figure 14). A temperature metaphor was used for the colour scale with 'hot' values (red, orange) representing the most severe events. Similar colour scales are frequently used, e.g. to emphasize districts with increased frequency of diseases [CRN*08].

The automatic component also employs a significance analysis based on the odds ratio (recall Section 2.2). Thus, the n most significant ADEs are emphasized. Temporal overviews are provided to indicate seasonal patterns of ADEs. For an overview of drug reaction pairs, two-level treemaps are employed with drugs at the first level and reactions on the second. This enables a comparison of the

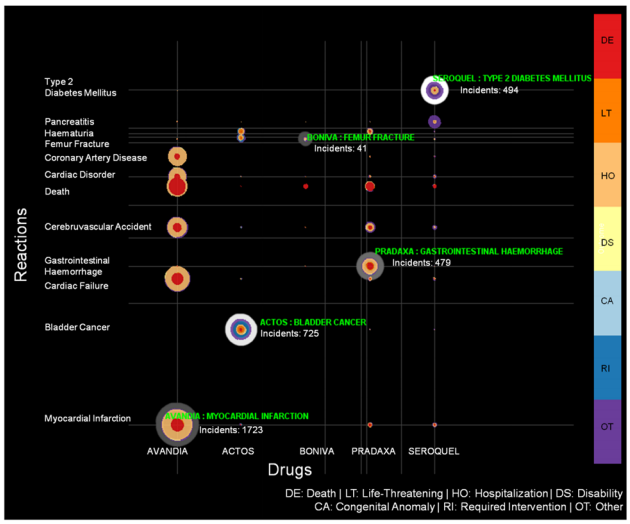

Figure 14: *Adverse drug effects are mapped in a plane with drugs (x-axis) and reactions (y-axis). The ADEs are categorized with respect to severity (red denotes death of the patient). The size of the glyphs represents the frequency (from: [MHD*14]).*

relative frequency of ADEs for a considerable number of drugs. Other visual analytics-based solutions for analysing ADEs were introduced by Buaceanu *et al.* [BAC*09] and Marcilly *et al.* [MHLA11].

In a similar way, epidemiologists analyse adverse effects related to medical implants. Appropriate visualization adds to a statistical analysis aiming at better patient information and prevention [FCBC15]. The temporal development, the spatial distribution of the related hospitals and the demographics of the patients are used for filtering and displaying results.

6.5. Surveillance of air quality

'With the rapid development of industrial society, air pollution has become a major issue in the modern world that has attracted increasing attention from the public and governments because of its impact on human health and societal development' [ZYL*17]. Numerous studies investigated *how* air pollution actually affects human health [DBR*97, Ped97] and thus demonstrate causal effects. Air pollution is considered responsible for 3.2 million deaths in 2010 worldwide [LVF*12]. Moreover, many respiratory diseases are considered to be caused at least partially by air pollution. Thus, monitoring air pollution, e.g. levels of respirable dust and NO_2, is regularly performed to identify changes in air pollution patterns and the influence of vehicular traffic, factories, terrain attributes and meteorological aspects on air pollution. This allows analysts to detect sources of excessive air pollution and thus supports decision-making about measures to improve air quality. Since air quality problems were particularly severe in fast-growing Asian cities, such as Hong Kong, major efforts were carried out there first [QCX*07, ZYL*17].

Monitoring environmental data was early considered an important challenge for visual analytics [TK09]. Although both air and water quality are essential for PH activities, visual analytics solutions focus on air quality surveillance. As an exception, Accorsi *et al.* [ASL*14] presented a system to analyse water quality.

Data. Based on current sensor technology, the necessary data are available in high-quality and high spatial resolution. Air quality monitoring stations may provide continuous streams of data. However, the analysis is carried out at a certain temporal granularity. Often, hourly measurements are employed [DMW*16]. Data processing is required in particular to remove obviously wrong or missing sensor values [DMW*16]. Air quality data typically relate to six scalar values representing fine dust ($PO_{2.5}$, PO_{10}) as well as CO, NO_2, O_3, SO_2). The surveillance of air quality requires the visualization and analysis of spatio-temporal air quality datasets [QCX*07, ZYL*17].

Requirements and tasks. The spatial character of air quality data requires map-based visualizations, i.e. any display should have an explicit link to the coordinates of a single station, or the group of stations to which it is related. The temporal character requires displays that convey the linear time component (time series data). In addition, the periodic character of air quality data is essential, e.g. daily, weekly and seasonal periods. Air quality data are moderately high-dimensional. Thus, techniques, such as parallel coordinates, may be used to show the data of selected stations. Analytical components may help to identify strong correlations, or to group stations with similar quality or to analyse time-series data in order to select and emphasize important intervals. Emissions due to vehicular traffic and industry should be integrated to support the identification of the *causes* of elevated pollution [DMW*16].

Visual analysis of the air pollution problem in Hong Kong. As a first example, we briefly describe the design and employment of a system developed in Hong Kong [QCX*07].

In contrast to later publications that did not explicitly consider weather visualization, the air quality data were displayed along with scalar weather data. Parallel coordinates were used to enable an understanding of correlations between scalar weather data (temperature, wind speed) and pollution. A special polar display was developed to display wind speed and direction. The spatial character of the data, however, is not displayed.

Multi-scale visual analysis of air pollution data. While air quality is often analysed on a local or regional level (recall [QCX*07]), it is also interesting to analyse large-scale patterns, e.g. on a country or even international level. Zoomable user interfaces, focus-and-context visualizations and displays that automatically adapt to the amount of data to be displayed, are viable solutions. Du *et al.* [DMW*16] introduced the AirVis system that is based on data from all 1111 Chinese monitoring stations and enables in-depth analysis of the country level. Calendar-based views are employed to show temporal data. For a station selected in the map, these views are presented as tooltips (see Figure 15).

Visual analytics of air quality data. The system presented by Zhou *et al.* [ZYL*17] was motivated by a project in Chinese cities. The underlying data represent six typical pollutants recorded at 946 stations in 190 cities covering a temporal range from July 2014 to May 2015. The data representing the amount of pollutants were averaged for both 8 and 24 h. Their visual analytics system comprised MCWs:

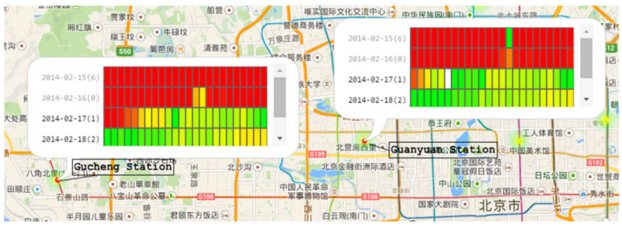

Figure 15: *A map-based view with tooltips for two selected stations. The tooltips contain a scrollable view that initially colour codes air pollution for four successive days on an hourly basis (from: [DMW*16]).*

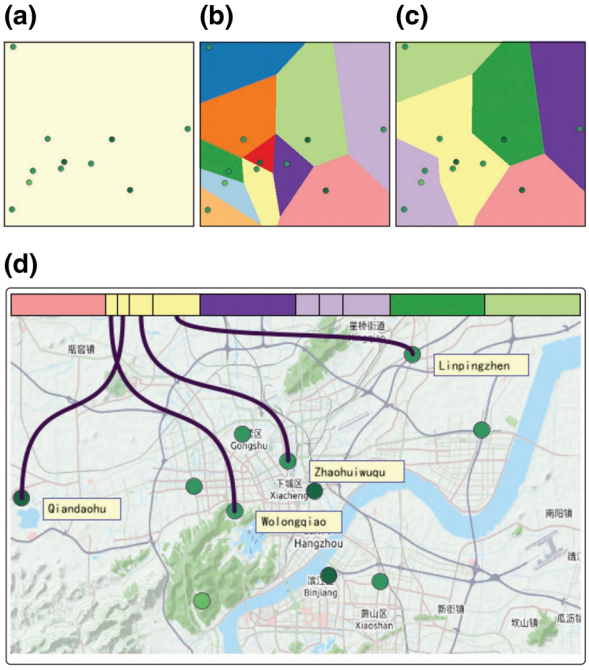

Figure 16: *Visual analytics of air quality: The monitoring stations are projected to 2D according to the similarity of the concentration of pollutants (a). The stations are clustered hierarchically according to the similarity of the concentrations (b, c). The correspondence to the geographic position is shown on a map view (d) (from: [ZYL*17]).*

- a map view with the monitoring stations,
- a view where the air quality is clustered and
- a *story line* that depicts changes of the air quality indicators per monitoring station over time.

A hierarchical clustering enables the analyst to inspect clusters at different levels. To explore the hierarchy, a treemap view is provided. The air pollution data are categorized in lower, moderate and higher values of a pollutant. The temporal course of the information is aggregated to daily, weekly, monthly and yearly patterns (see Figures 16 and 17). The visualization component supports a wide range of analysis tasks, including an overview of temporal developments and a fine-grained analysis.

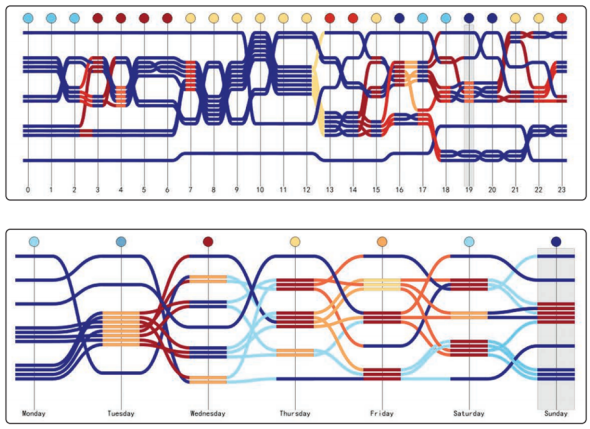

Figure 17: *The development of the air quality indicators over time in every cluster is shown on different temporal scales, e.g. daily (top row) and weekly (bottom row). Colour encodes the different clusters (from: [ZYL*17]).*

Oil and gas deployment. The air quality in the vicinity of oil and gas deployment sites is a special problem [CKSB11]. Hsu *et al.* [HCD*18] describe a system that enables the collection and display of time-referenced air quality data along with (self-reported) physical and psychosomatic symptoms as well as personal stories, involving images. Personal stories provided by affected citizens represent a less structured data type. The involvement of experts and affected citizens is a speciality carefully considered in the system design. The data comprise peaks of contamination, e.g. with fine dust, and are analysed with respect to peak duration (how long values exceed a threshold) and frequency. Visualization techniques include a map display as well as parallel coordinates to reveal relations from air quality sensors and symptoms. Summary statistics are provided at the zip-code level. This type of aggregation ensures the privacy of the citizens.

Discussion. The surveillance of air quality is related to *urban computing*, where meteorological data, traffic density or water quality are analysed. Zheng *et al.* [ZWC*16] provide a survey on visual analytics in urban computing. The influence of weather attributes on pollution gave rise to integrated visualizations [QCX*07]. Since air quality data have a temporal component, the design of VA solutions benefits from general design considerations of temporal data [AMST11]. Several authors, e.g. Wood *et al.* [WDSC07], report that they are inspired by general design principles for map-based visualization.

Typically, the risk of individuals is computed based on their home address. Pollution, however, may be quite different at their working place or other essential places of their activity space. This limits the expressiveness of small neighbourhood statistics [JGK10]. Predictions of air quality are typically related to pollutant concentrations, land use, and traffic. Since the quality of such data is often limited, satellite imaging as a remote sensing technique is discussed as an alternative [JGK10]. From a PH point of view, it is important to correlate measures of air pollution with the incidence of respiratory diseases to provide orientation for political measures, such as the definition of thresholds for critically elevated values. While current

visual analytics systems aim at PH experts, the sensitive issue of air quality also warrants the development of (simplified) displays for the general public.

6.6. Understanding pathways leading to asthma

An important aspect of preventive medicine is to understand how chronic diseases arise and develop. Visual analytics support was developed for Asthma research. Asthma is a widespread and potentially severe disease. Substantial research is carried out to better understand pathways that may lead to an asthma disease as well as stages of the disease [BBK*15, BDD14, JGK10]. This research involves complex data, such as molecular and phenotype information as well as industrial pollution. Jerret *et al.* [JGK10] present an example of map displays linking major industrial zones with adjacent asthma rates where a strong association was found. Visual analytics solutions may help to identify and characterize these and other subpopulations with modified asthma risk. Bhavnani *et al.* [BDD14] focused on the relation between cytokine levels and patient characteristics, e.g. respiratory function values. Bi-partite graph-based visualizations with two node types: patients and cytokines were generated. Hierarchical clustering on the patient data was performed to reveal patients with similar respiratory function values. The graph-based visualization was also used to quantitatively analyse the graph, e.g. with respect to degree assortativity. Heatmaps were employed to reveal the cytokine expression level for all patients. The heatmap was enhanced with the dendrogram of the hierarchical clustering. The limitation of this paper is the rather low number of patients (83) and cytokines. Thus, this technique may not scale well to larger data. The use of graph-based visualizations is typical for the analysis of biomedical data, e.g. in epidemiology. Bi-partite graphs are also used to understand disease–gene and disease–protein associations [BDD14]. Also, the combination of graph visualization and clustering, i.e. the visual encoding of cluster membership, is potentially useful.

6.7. Cancer epidemiology

In Section 2.1, we described cancer epidemiology as an essential field due to the high incidence and mortality of cancer. Cancer epidemiology deals primarily with risk factors or combinations of factors that increase the risk for acquiring a certain type of cancer. An essential aspect is the spatio-temporal analysis, e.g. whether a certain type of cancer is uniformly distributed in space, time and space-time or whether there are certain hot spots. A more fine-grained analysis also investigates the stages at which the tumour disease is diagnosed, i.e. whether there is a higher frequency for late-stage colorectal cancer in a certain region [DS07]. Such an analysis could give rise to campaigns for early detection measures.

Bieh-Zimmermann *et al.* [BZKF13] discuss the need for scalable visualizations to support the exploratory analysis of cancer registry data. Such data are typically analysed only with hypothesis-driven statistics methods. Simple diagrams, such as histograms, bar charts and pie charts are used to convey major facts. Bieh-Zimmermann *et al.* [BZKF13] provide an initial prototype for a subset of cancer registry data that employ parallel sets to reveal relations, e.g. between different age groups and frequent locations of tumours. Interaction facilities are provided for a fine-grained analysis. Given the multitude of data in cancer registers, however, a more comprehensive visual analytics system would be needed.

Spatio-temporal analysis. The commercial software Biomedware (https://www.biomedware.com/) provides support for analysing cancer registry data. Space-time clustering may be performed and the results are visualized along with various map types to identify regions and temporal intervals with significantly increased incidence of a certain cancer type. As an example, Nordsborg *et al.* [NME*14] provide a visual analysis of breast cancer data from the Danish cancer registry. In a similar way, Nordsborg *et al.* [NME*13] provide an analysis of more than 3,000 cases of Non-Hodgkin lymphoma (NHL). The spatial clustering of NHL serves to identify environmental risk factors for this disease. SaTScan (https://www.satscan.org/tutorials.html) is another commercial software for analysing diseases clusters that was largely used for analysing cancer registry data. The SaTScan website lists numerous publications where the software was used for analysing geo-referenced health statistics, e.g. for colorectal cancer [DS07]. SaTScan is also recommended at the website of the US National Cancer Institute (https://giv.cancer.gov) for spatial statistics and disease clustering.

Iqbal *et al.* [IHN*16] analysed a large amount of patient visit data to better understand diseases that are associated with cancer. Data from the National Health Insurance Claims database were employed for this research. The data were particularly large—it relates to 782 million patient visits, representing 20 million unique patients. Compared to epidemiological studies, however, fewer dimensions were available. In addition to age and gender, the procedural codes for diseases were stored. For each patient, a disease–disease association was assumed if the patient has both diseases within a 36 month period. Such associations were analysed for 100 male and 100 female age groups. Associations that were rare in relation to the overall population were discarded.

Animated display. For nine types of cancer, Iqbal *et al.* [IHN*16] found that some chronic diseases, such as diabetes, are associated with a moderately increased risk, e.g. for breast and colon cancer. They displayed these associations for the entire population and for persons of certain age groups. While most time-varying PH data are visualized with time-line-based techniques, Iqbal *et al.* employ animation and refer to their system as CAMA (Cancer Association Map Animation). The animation conveys how the prevalence of diseases changes depending on age. Inspired by Hans Roslings dynamic bubble charts, they employ similar visualizations where circular glyphs change their position in the chart to represent the dynamic character. The glyphs represent the co-morbidities of chronic diseases and cancer. Due to the large differences in the prevalence depending on age, a log-log scale is applied to the scatterplots.

Diseases are classified into 17 categories, e.g. skin diseases, respiratory diseases, which corresponds to the standardized diagnosis according to ICD-9. The categories are mapped to colour. This is not ideal since 17 colours cannot be discriminated pre-attentively. The visualization may be filtered for a subset of categories, individual diseases or association strength. The implementation is based on an open-source project [AACD10].

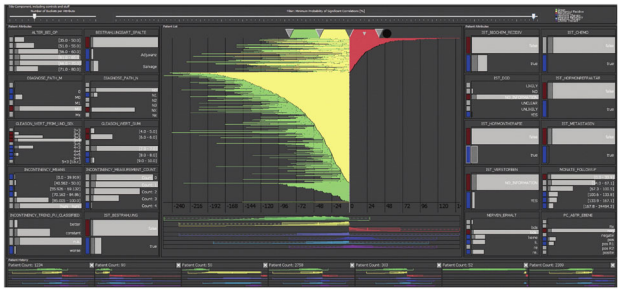

Figure 18: *The interface elements on the left enable the definition of a cohort, e.g. based on age, diagnostic values or treatment. The central view provides an overview over a cohort of prostate cancer patients based on their well-being with red denoting metastatic disease (from: [BSM*15]).*

No comprehensive evaluation or a comparison with other time-based visualizations is given. Thus, it is not clear whether this visualization is helpful for the knowledge discovery process of PH experts. It seems more likely that such visualizations serve the interest of the general public. Iqbal *et al.* [IHN*16] discuss arising research questions, e.g. whether the high insulin resistance associated with diabetes causes some types of cancer.

6.8. Cohort analysis of prostate cancer

In this subsection, we describe a research effort that is at the interface between epidemiology and clinical medicine. The underlying data are from a hospital—thus, it is data related to patients and the data exhibit a selection bias. We discuss this effort because the unusual high number of patients lead to a design that may inspire PH research. In contrast to the work described in the previous subsection, it is focused on one type of tumour with a particularly high incidence in males, namely prostate cancer. Bernard *et al.* [BSM*15] performed a comprehensive cohort analysis of 16 000 prostate cancer patients to improve guidelines for prostate cancer treatment.

Radical treatment, which may cause severe side effects, such as urinary dysfunction, is often not necessary. The subjective patient well-being is registered at every follow-up diagnosis. Filtering comprises variables, such as tumour stage and tumour grading as well as applied treatments. Correlations between the variables are automatically computed whenever the filtering changes. The patients' well-being is characterized as a categorical variable with three possible values, visually represented by green, yellow and red with red denoting a bad state (see Figure 18). Emphasis is put on cohorts' statistical properties, such as mean and standard deviation. A history function enables to go back to previously defined cohorts and related visualizations and analysis results.

While many features could be translated to similar medical research problems, some aspects are tailored towards prostate cancer. As an example, the course of the prostate-specific antigen (PSA) was considered as the most important information and thus a temporal overview visualization is generated that correlates this value with the patient well-being and optionally also with treatments, e.g. the start and end of hormone therapy or radiation treatment. The visual summary of the state of multiple patients in the *Patient Bundle View* is the most carefully discussed aspect. Four aggregation levels are provided and used for semantic zooming. The information related to the cohort is input to a Kaplan–Meier estimation of survival. Bernard *et al.* [BSB*15] may serve as an orientation with respect to the depth of the requirements and task analysis as well as the discussion of iterative prototyping.

Table 3 summarizes major applications in medical research for which visual analytics solutions were developed

7. Visual Analytics of Population-Based Cohort Study Data

We now discuss a special problem in more depth, namely the analysis of population-based studies (recall Section 3). Population-based studies involve healthy volunteers, no patients, as the research described in Section 6. The available data are much more comprehensive compared to clinical data, since the participants get a wide range of examinations and interviews lasting for one or two full days. The population-based character is the major difference to the cohort study described in Section 6.8.

The design, conductance and analysis of such studies is a major task for PH academics. These specialized epidemiologists should also be enabled to reveal subtle effects of lifestyle and environmental factors on human health. The wealth of data acquired in such studies enables a range of retrospective research goals that may not be completely anticipated when the cohort study was designed.

Since population-based data contain randomly selected participants, many diseases and injuries are rare. Thus, the focus is on

- frequent disorders, such as adipositas or fatty liver, known to be risk factors for a number of diseases, and
- frequent diseases, such as diabetes and coronary heart disease.

Population-based studies aim at an understanding of complex interactions between lifestyle, genetics and the outbreak of diseases.

7.1. Visual analytics and radiomics

Modern cohort studies involve medical image data, such as ultrasound or MRI (recall Section 3.1) to characterize the presence or absence of pathological abnormalities. As an example, the SHiP data contain the diagnosis of fatty liver based on ultrasound data. Despite efforts to standardize the process, manual diagnosis is to some extent subjective. Thus, automatic solutions are also considered to provide more reliable information.

Based on a segmentation of organs or other relevant structures, *imaging biomarkers* may be derived to characterize the morphology, e.g. the size, volume, circumference, compactness or non-sphericity of anatomical structures. A whole branch of research in radiology relates to the diagnostic value of features derived from image data. Inspired by genomics data, this branch is referred to as *radiomics*. Radiomics features are particularly interesting if they serve to identify diseases at a preclinical stage (recall Section 3.1). Most applications aim at the characterization of tumour diseases but, of course, a broader range of research questions is possible, e.g. related to neurodegenerative diseases or backpain. Meuschke *et al.* [MGW*18],

Table 3: *Major visual analytics systems for epidemiological research.*

Application area	Key features	Key publications
Cognitive ageing	Advanced data management, display of fibre tracts and related statistics	[AOH*14]
Food-borne diseases	Analysis of arsenic in different food categories	[JCIA10, SIC*11]
Injury prevention	Spatio-temporal analysis of injuries by type and age group	[AHFSP17]
Adverse drug effects	Overview of indication, frequency and outcome	[MHD*14]
Air quality	Spatio-temporal analysis of pollutants along with weather data	[QCX*07, ZYL*17]
Asthma research	Relation between cytokine levels and asthma, bipartite graphs	[BDD14, JGK10]
Cancer epidemiology	Animated bubble charts, coloured circular glyphs	[IHN*16, BZKF13, NME*13, NME*14]
Prostate cancer	Cohort definition, display of cohort at four levels	[BSM*15]

for example, describe how parameters derived from the shape of an aneurysm may be used to predict whether this aneurysm will rupture. This is an example of the fruitful integration of visual analytics and radiomics as discussed first by Bannach *et al.* [BBJ*17]. Radiomics features can be used for content-based image retrieval and cohort construction (recall Section 4.2).

Image analysis. Most radiomics-related research is carried out by radiologists and relies on manual segmentation. This involves a limited accuracy and reproducibility. For large- or mid-sized cohort study data, it is also not feasible, since too many datasets need to be processed in a tedious manner. Thus, there is a need for (semi-)automatic processes to derive radiomics features.

Automatic segmentation is challenging due to the large variety of anatomical and pathological variants. MRI and ultrasound data, the prevailing modalities in cohort studies, exhibit more artefacts than CT data. The choice of image acquisition parameters is a trade-off between patient comfort, image quality and costs (for scanning thousands of participants). As a consequence, the image quality might not be ideal. On the other hand, epidemiology data come with a lot of demographic data, e.g. gender, weight, height and age, which may be employed to decrease learning costs for a machine learning image analysis [TGR*15].

Toennies *et al.* [TGR*15] describe a general strategy for image analysis in epidemiology and emphasize that cohort data are acquired for *open research*. Since research questions related to the anatomy and function of organs may arise years after the data were acquired, they suggest a modular approach with at least partially reusable methods. These methods were applied for automatic kidney segmentation from MRI data of the SHiP data (recall Section 3.1) [GTL*12]. Some anatomic structures are too complex for immediate quantitative analysis. In this case, they may be decomposed to subshapes with reduced complexity and variability [TRE14]. As an example, the spine may be decomposed into the spinal canal, individual vertebrae and sections such as the lumbar spine [RET13]. The problem of large-scale image analysis of whole databases is still not solved. Further progress is expected by combining machine learning and visual analytics research [ZM16].

Applications. Klemm *et al.* [KLR*13] used the SHiP data to analyse the shape of the lumbar spinal canal to investigate whether it is associated with lower backpain—a hypothesis of the epidemiologists. After the lumbar spine was extracted and transformed to a three-dimensional (3D) surface model, its centreline was generated as a representative for the shape of the lumbar spine model. Afterwards, these centrelines are grouped into clusters with an agglomerative hierarchical clustering that was originally developed to cluster streamlines from blood flow simulations [OLK*14]. To visualize these clusters, a ribbon-based visualization was designed (see Figure 19). In the evaluation, an epidemiologist emphasized the importance of a high segmentation quality to derive imaging biomarkers.

Later, Klemm *et al.* [KOL*14] extended their framework to a web-based exploration tool. Based on an analysis workflow, they developed a framework that facilitates the generation of hypotheses and their subsequent statistical analysis. All variables in the cohort are listed, and the expert can drag and drop certain variables onto the main canvas, which leads to a representation of the mean lumbar spine model of the patients in the selected group. Additional refinement or selection of new variables results in a visualization showing correlations. Brushing and linking options support the analysis. With this framework, epidemiologists were able to explore shape information of the lumbar spine and its influence on diseases.

In this work, a subset of the SHiP (the 2.240 female participants) with 134 variables from which 21 are metric and 113 categorical was employed. In addition, nine parameters were derived from the spinal canal centreline.

7.2. Identification of strong correlations with disorders

Due to a large number of variables, epidemiologists benefit from an automatic analysis related to potential associations between lifestyle-related variables and disorders, such as increased breast density which is known to be a risk factor for breast cancer. Klemm *et al.* [KLG*16] presented the 3D regression heatmap to analyse correlations between variables. Their idea is to let the experts input simple regression formulas, e.g. $Cancer \sim X + Y$ to explore the correlations. This calculates all combinations of pairwise variables for a correlation of cancer by using the R^2 metric. For the visualization, a heatmap was employed. In case the expert types $Z \sim X + Y$, a regression cube was generated showing for every slice a 2D heatmap of correlations. The downside of their approach was the computation time: a dataset of 100 features needs roughly 14 h to compute $Z \sim X + Y$. Nonetheless, the experts stated that the approach was helpful to gain an overview and to find non-obvious correlations.

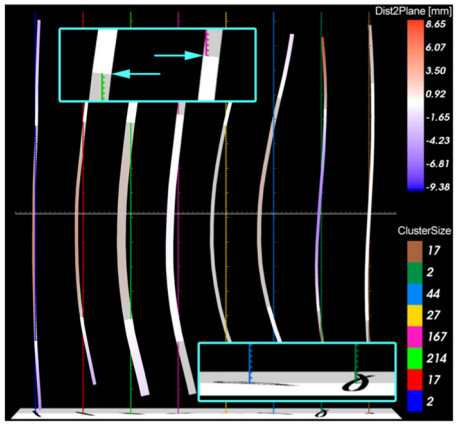

Figure 19: *Clustered centrelines of the spinal canal. Cluster size is mapped to the width of the ribbons and colour encodes distance to the midsagittal plane. Selection of a cluster leads to the display of related image data (from: [KLR*13]).*

7.3. Data quality

Visual analytics may help to identify and characterize quality problems in cohort study data, such as outliers, missing values and double counts [JCIA10]. Severe outliers, for example, are pre-attentively identified in an appropriate visualization and, as a consequence, outlier removal may be considered. Johnson *et al.* argue that visual analytics applications should 'clean up the data or make the analyst aware of the shortcomings in the data' [JCIA10]. Few attempts were made to assess and improve the quality of PH data. Shneiderman and Plaisant [SP19] recently provided a discussion of visual event analytics with a number of examples from health care, including their use for detecting data quality problems. As a specific example, an appropriate visualization clearly revealed that some patients were recorded to be admitted to the hospital much more often than being dismissed—a typical example where data relevant for billing are correctly registered and less relevant data contain errors. There are general strategies to clean data using visual analytics [GAM*14], but no specific solutions for PH. The exception is related to missingness, which we discuss in the following.

Missingness is an essential quality problem that occurs in all data sources that we discussed in this paper [Don06]. In the evaluation of a visual analytics solution for cohort study data, PH academics encouraged 'techniques for detection and handling of missing data', e.g. the presence of incomplete data needs to be clearly communicated [SMvB*10]. Missingness may occur in one cycle of a longitudinal study, i.e. the values for one participant are not complete, or between cycles, where participants do not show up in a later stage (*drop out*). *Missingness maps* [HKB*11] may serve to identify patterns, i.e. situations where missingness is not completely at random (see Figure 20). There are various strategies to cope with missingness [SWC*09]:

- *Complete case analysis*, where only complete datasets are analysed,
- *Single imputation*, where missing values are replaced with a median or average value and

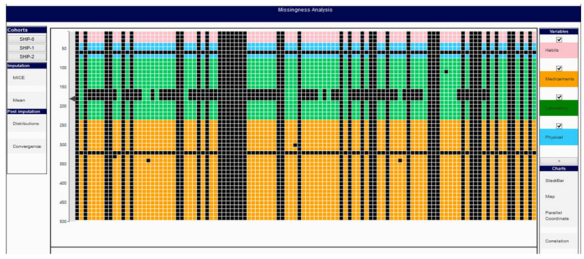

Figure 20: *Overview of missing data from an epidemiologic study: Rows represent variables and columns show participants. Missing values are indicated in black. Completely black columns represent drop-out participants (from: [ANI*17]).*

- *Multiple imputations*, where dependencies between variables are considered and multiple replacements are computed.

The first two strategies are straightforward to realize but not appropriate for typical PH data. If only complete cases are considered, the number of cases often shrinks drastically. The resulting subset may be no longer representative, if the missingness is not completely at random, i.e. the likelihood of a missing value for one variable depends on the value of another variable. Single imputation preserves all datasets and thus also the representative character. However, the median or average is often not a good guess for the missing value. If all missing values are replaced with the (same) average value, the variability of the distribution gets reduced.

Multiple imputations are based on a regression analysis: for a variable v_1 that is affected by missingness, the (linear) correlation to all other variables is computed. To save computational effort, only variables with a high correlation (e.g. one of the N highest values or above a threshold) are used for the *prediction matrix* that is employed for imputing the missing value. Imputation is performed several times, leading also to an estimate of the uncertainty involved. In addition to the size of the predictor matrix, the number of iterations influences the computational effort. The default value for this number is often five [ANI*17]. Alemzadeh *et al.* [ANI*17] showed the converging behaviour of the imputation with an increasing number of iterations and confirmed that five iterations often are a good choice.

The use of multiple imputations is appropriate when the missingness is likely to be not completely at random. It is supported in major statistics tools, e.g. the multiple imputations with chained equations (MICEs) package in R which is frequently used in epidemiology [BGO10]. MICE was used to prepare the data of the UK Biobank for an analysis of the predictive value of self-reported health data [GI15] and for the analysis of the SHiP data with respect to hepatic steatosis [ANI*17, ANI*19]. Visual analytics plays an essential role for the identification and the handling of missing data. After imputation, distributions of imputed values can be compared to distributions of the available values to validate the result.

8. Evaluation

Even if the development of a visual analytics system is based on an extensive requirements analysis, involving observations and

Table 4: *Use of visualization techniques in public health applications.*

Technique	Applications	Publications
Parallel coordinates	Cognitive ageing, air quality and weather conditions, air quality and symptoms	[AOH*14, QCX*07, HCD*18],
Parallel sets	Cancer registry data	[BZKF13],
Treemaps	Air quality data, adverse drug effects	[MHD*14, ZYL*17]
Calendar views	Incidence of diseases, air quality	[TSS05, DMW*16]
Age-pyramids	Age and gender effects in the incidence of diseases	[CWCN11]
Choropleth maps	Infectious disease analysis, combined analysis of human and animal health, pandemy visualization	[AAA*16, GLG*12, MLR*11]
Heatmaps	Malaria map, infectious disease analysis, asthma research	[AAA*16, BDD14]
Dotplots	Combined animal and human health	[MTJ*07, MHR*09, GLG*12]
Flow maps	Outbreak detection	[Guo07]
Bi-partite graphs	Asthma research	[BDD14]

Table 5: *Use of analytics components of visual analytics solutions.*

Technique	Applications	Publications
Spatial clustering	Cognitive ageing, air quality data	[AOH*14, QCX*07, HCD*18]
Space-time clustering	Cancer epidemiology (breast cancer, NHL)	[NME*13, NME*14]
Subspace clustering	Cohort study data, identification of subpopulations	[AHN*17a]
Hierarchical clustering	Cohort study data (backpain), asthma research	[KLR*13, BDD14]
Regression analysis	Cohort study data, identification of risk factors	[KLG*16]
Association rules	Cohort study data, identification of risk factors	[HSKO14, NSVK14a]
Multiple imputation	Analysis of missing values in cohort study data	[ANI*17]

interviews with all stakeholders, the prototype needs to be evaluated before any valid claims can be made regarding the benefit for the target users. Evaluation is considered by Shneiderman *et al.* [SPH13] as one of seven challenges for visual analytics in health care. 'Field deployment methods [. . .] in the real context of use and the impact of this context on the user experience' should be performed [ZS17]. Systems should be evaluated with respect to their ability to prevent errors, i.e. misleading information display, and with respect to their ability to support evidence-based explanations.

Evaluations of PH activities require the actual use of the systems by PH experts doing 'real' work with representative data. A typical lab experiment with a short timeframe is not sufficient. Informal evaluations with a few experts and different instruments, such as thinking aloud, video analysis and interviews are more promising. Insight-based evaluations may be carried out to understand the potential of a VA solution for knowledge discovery [SND05].

In the following, we discuss selected evaluation strategies and their use in systems developed for PH. The actual state of evaluations lacks behind the documented knowledge, e.g. in information visualization [LBI*11] and medical visualization [PRI18]. Many publications do not document any user feedback. Instead, in a 'case study', the hypothetical use of the system is explained, e.g. in the ID-VIEWER (recall Section 5.3 and [AAA*16]), in the PANVIZ system (recall Section 5.5 and [MLR*11]) and in the subpopulation discovery from cohort study data (recall Section 5 and [AHN*17a]).

Other publications gather feedback based on a presentation of the prototype or a video that illustrates how the prototype is intended to work. Such a video presentation was used, e.g. by Masoodian *et al.* [MLK16] and the discussion with four physicians, including two epidemiologists revealed some insights and ideas for extensions, e.g. related to a better support of cooperations.

Klemm *et al.* [KOL*14, KLG*16] were inspired by the seven evaluation scenarios discussed by Lam *et al.* [LBI*11] and chose the *visual data analysis and reasoning* scenario to evaluate their systems. Klemm *et al.* [KOL*14], for example, discussed how two epidemiologists employed a system for analysing cohort study data for hypothesis generation and testing with an in-depth discussion of understanding associations, use of clustering and detailed inspection of clustering results. Gesteland *et al.* [GLG*12] employed a *technology acceptance model* [VMDD03] in their evaluation of the EPICANVAS. According to this model, they questioned the *intent to use*.

The injury prevention dashboard was evaluated among others with *distributed cognition*—an evaluation technique that is intended to solve cooperative problem solving. The most extensive evaluation was found in the discussion of the EPINOME system (recall Section 5.5 and [LRS12]). Ten outbreak scenarios from which the 27 participants got a random selection of four were employed. The participants were enabled to use the system including the assignment of interventions, such as prophylax of high-risk contacts. The participants were selected to be as representative as possible for the target user group in terms of age, responsibility, experience and gender. The underlying data are realistic: it discriminates the reported cases and the actual (higher) number of cases. The simulation also considers paradox effects, i.e. systematic contact tracing increases the number of reported cases but reduces the actual number of cases since some cases can be avoided by warning or vaccinating contact

persons. The users appreciated that they can analyse what-if scenarios in detail. In summary, the users were asked among others whether the system provides new insights and decision support and whether the system helps them to do their job more efficiently. The large majority of the users agreed or strongly agreed to all statements. An even deeper analysis would be more precise in characterizing the nature of insights and the specific decision support.

Zakker and Sedig [ZS17] performed an informal evaluation with a demonstration part, an exploration part, and a feedback interview was performed. Seven PH experts with different background and experience were recruited from health research centres. Although the actually used visualizations, prepared with TABLEAU, appear rather simple from a visualization scientists' point of view, users uttered that information overload and clutter needs to be avoided. They prefer simple charts and some were even skeptical that a choropleth map may be interpreted wrongly.

In summary, different evaluation methods were used to understand the (potential) benefit of visual analytics systems in PH, including gathering informal feedback, the use of questionnaires, the analysis of the reasoning processes, distributed cognition and technology acceptance models. The evaluations focus on the overall impression and usefulness of whole visual analytics systems. Although useful, such evaluations do not allow to justify individual design decisions, related to the use of visualization, interaction and analysis techniques.

9. Research Agenda

In the following, we describe a research agenda largely driven by the gaps in the literature and the needs of PH experts.

Cooperative visual analytics. On the one hand, cooperative solutions are particularly important in PH due to the diversity of involved experts. On the other hand, we found only two systems [AHFSP17, MLK16] that explicitly addressed cooperation modes. Masoodian *et al.* employed targeted support with shared and private views. Both remote cooperation, e.g. between field workers gathering data and a PH office, and co-located cooperation is relevant. Al-Hajj *et al.* [AHFSP17] discussed the cooperation between domain scientists and visual analytics experts, but without any special hardware to support the cooperation. General thought on cooperative visual analytics [TIC09, IFP*12] may serve as orientation.

Visual analytics for the evaluation of PH interventions. An essential aspect of PH is the evaluation of measures aiming at improved prevention. This is crucial, e.g. to justify the considerable financial resources necessary to implement measures, such as vaccination programmes. Comprehensive data are acquired for evaluation purposes. Since the evaluation is strongly based on statistical data (evidence-based PH), there is an opportunity for visual analytics research. We have not found visual analytics systems supporting this evaluation process.

Support data management and integration. Various publications dealing with the problem-solving processes in PH scenarios emphasize the need to *integrate* data from different sources. We found very few publications that describe their data management, e.g. [AOH*14]. Due to the spatial and temporal character of the data, simple databases are not sufficient. Also the size of the data may require special solutions to ensure that analytical approaches can be performed fast enough. Support for *progressive analytics* may be useful to enable the use of regression analysis and clustering on large and high-dimensional data [SPG14].

Visual analytics of social media data. Social media data provide a multitude of timely information related to health problems. However, this information is informal and multi-lingual making an analysis challenging [Dre12]. Considerable research efforts in linguistics, topic modelling and text mining were carried out to extract information useful for PH. This research is mainly motivated by the expensive, slow and tedious process to collect data in a systematic manner [Dre12]. Social media analytics is particularly promising for discovering trends in mental diseases, where the occurrence of words such as 'bored', 'tired' and 'exhausted' are indicative. There is increasing use of social media analytics for physical diseases. As an example, Culotta [Cul14] found that aspects of the personality, attitude and nutrition add to the traditional demographics variables to identify vulnerable subpopulations. So far, the analysis of social media data for PH does not include visual analytics. However, this is a research opportunity for visual analytics.

Guidance. Most research described in this survey, including all systems to which the first author has contributed, is not regularly used by their intended target audience. They are too complex and rely too much on advanced visualization and analysis techniques that PH experts do not fully understand. To bring this area forward, it is essential that research ideas emerge from the *actual use* of such systems. Al-Hajj *et al.* [AHFSP17] discuss this issue with respect to their *injury dashboard* and highlight the importance of suggestions for appropriate visualization techniques (derived from the data to be displayed), context-sensitive help on demand, instructive videos and structured tutorials. While these methods may be sufficient for this particular system, most other systems probably need a major re-design from a user-centred point of view. The recent trend towards *guided visual analytics* [CGM*17] is also promising for PH applications. The *statistics wizard* developed for epidemiologists and introduced by Thew *et al.* [TSP*09] may serve as orientation.

Improved evaluation strategies. This aspect is related to the previous one: Once more user-centred systems are developed, the need to assess their use in-depth increases. Since PH experts are often concerned about the validity of their data and conclusions, it is essential to understand to what extent they trust the findings they acquire from using a visual analytics system. Moreover, repeated testing is recommended to understand how the use of a system changes over time and whether it reaches a *stable* state. Repeated testing leads to the idea of *long-term case studies*, where users employ the system themselves, report in a diary how they use a system and get interviewed—a type of evaluation advocated by Shneiderman and Plaisant [SP06]. Compared to the state in some other application areas where the knowledge discovery process was characterized in detail [SS05], the evaluations related to visual analytics in PH are not particularly elaborate. As an example, we did not find any long-term evaluation. Also, eye-tracking-based evaluations that have a great potential to reveal how users actually employed a system were not used in visual analytics for PH. The survey of Blaschek *et al.* [BKR*17] may serve as source of inspiration.

Molecular epidemiology involves an understanding of the spatial distribution of gene variants in pathogens, e.g. related to HIV or other infectious diseases. While there is already software support with basic visualization for analysing this data (see Carroll *et al.* [CAD*14]), we found no visual analytics solutions in this area.

Identification of non-linear regression. We discussed regression analysis as a means to identify and characterize associations between lifestyle, exposition to risk factors and diseases. In all VA systems used for PH, this analysis is restricted to *linear regression* (for scalar data) and *logistic regression* (for categorical data), which poses a severe limitation. Many strong associations are distinctly non-linear and would not be detected. Examples for non-linear relations relate to blood pressure and sleeping duration, where a range of normal values exist and both very low and very high values are associated with increased risk. Such a risk distribution is referred to as *U distribution* [FF11] and may be characterized with quadratic regression. Care is necessary to avoid overfitting when higher order polynomial associations are searched for.

Also partial regression, which analyses correlations, which relate only to a portion of the range of a variable, may yield essential findings. In other applications, partial regression models were interactively constructed and successfully used [MP13]. This paper may serve as inspiration for enhancing regression analysis in PH applications with a focus on a user-steered process.

Data preparation and data quality. There is only few research on how visual analytics may help to identify, assess or counteract quality problems in PH data, although this problem is clearly relevant (recall Section 7.3). Credibility and validity of the data is a major concern for PH experts. More research is needed to formalize *a priori* knowledge about the validity and plausibility of data to employ this knowledge for analysis and visualization of (potential) quality problems. Often, it may not be possible to reliably correct data. In this case, at least the resulting uncertainty should be quantified and visualized (recall Section 4.4.4).

Causality analysis and visualization. Visual analytics solutions are able to identify correlations. The ultimate goal of medical research is, however, to understand causal effects. We found no publications that discussed causality inference and visualization with respect to PH. In other areas, causality visualization is discussed [DMAF15]. The refinement of such techniques, e.g. to cohort study data, is an important area for further research.

Although not being a research issue in a narrower sense, there is a great need for open-source software to better cope with the interoperability issues, see, e.g. Carroll *et al.* [CAD*14] for a discussion.

10. Concluding Remarks

PH employs large, heterogeneous, partially incomplete data that are often geo-referenced and time-dependent. The complex data are extremely challenging for our cognitive abilities. Visual analytics solutions may support knowledge discovery and problem solving if the needs of the different stakeholders, such as epidemiologists and environmental health specialists, are adequately addressed. A wide variety of visual analytics techniques and strategies are incorporated in PH applications, including graph-based visualization and graph-theoretic measures, (subspace) clustering, dimension reduction, coordinated views, regression-based visualizations as well as various time-based visualizations. Table 4 gives an overview of visualization techniques in different applications. Not always was the selection of these visualization techniques carefully justified. Thus, we summarize the observed use of techniques without being able to give recommendations. Since most systems contain multiple views, histograms, scatterplots and timelines for temporal data, we have not mentioned these techniques. In a similar manner, we summarize the use of analytical techniques in Table 5. It turns out that different variants of clustering are widely used. They primarily serve to analyse subpopulations. Clustering results are displayed along with a variety of visualization techniques, typically by colour-coding the membership of items. The limited completeness and reliability of the underlying data is rarely considered. Most systems take the loaded data for granted and apply analytic and visualization techniques directly. While incompleteness is obvious and may give rise to the use of imputation strategies, unreliability is more difficult to quantify which would be a prerequisite for an uncertainty-aware visual analytics process.

Many systems support the spatio-temporal analysis of disease-related data. They are designed to afford comparisons at different temporal and spatial scales and are typically realized as MCW frameworks. Outbreak surveillance systems often incorporate simulation to enable predictions and thus to directly support decision-making with respect to possible interventions.

Despite the strong potential and in-depth research activities from both visual analytics and PH experts, simple and often static visualizations dominate in routine practice [ZS17]. Visual analytics solutions have to represent the temporal and spatial character for most tasks performed by PH experts, e.g. analysing how health indicators have developed. A tight coupling of visual analytics and statistics is of utmost importance. While visual analytics solutions often help to detect patterns and correlations, ultimately they are supposed to favour an *understanding* of the underlying mechanisms, e.g. biological and physiological processes that *explain* the findings.

Acknowledgements

We thank Shiva Alemzadeh, Lena Cibulski, Tommy Hielscher, Paul Klemm and Uli Niemann for major contributions to visual analytics systems for epidemiology, Myra Spiliopoulou for advice on medical data mining, Klaus Toennies for advice on image analysis for epidemiology and Monique Meuschke for proof reading the manuscript. Moreover, we thank our collaborators at the University of Greifswald Katrin Hegenscheid, Till Ittermann and Henry Völzke. This work was partially funded by the Carl-Zeiss foundation.

References

[AAA*16] Ali M., Ahsan Z., Amin M., Latif S., Ayyaz A., Ayyaz M.: Id-viewer: A visual analytics architecture for infectious diseases surveillance and response management in Pakistan. *Public Health 134*, Supplement C (2016), 72–85.

[AACD10] Al-Aziz J., Christou N., Dinov I.: SOCR motion charts: An efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data. *Journal of Statistical Education 18*, 3 (2010).

[ABKS99] Ankerst M., Breunig M. M., Kriegel H.-P., Sander J.: OPTICS: Ordering points to identify the clustering structure. In *ACM Sigmod Record* (1999), vol. 28, pp. 49–60.

[AF94] Allen J. F., Ferguson G.: Actions and events in interval temporal logic. *Journal of Logic and Computation 4*, 5 (1994), 531–579.

[AHFSP17] Al-Hajj S., Fisher B., Smith J., Pike I.: Collaborative visual analytics: A health analytics approach to injury prevention. *International Journal of Environmental Research and Public Health 14* (2017), 1056.

[AHN*17a] Alemzadeh S., Hielscher T., Niemann U., Cibulski L., Ittermann T., Völzke H., Spiliopoulou M., Preim B.: Subpopulation discovery and validation in epidemiological data. In *Proceedings of EuroVis Workshop on Visual Analytics (EuroVA)* (2017).

[AHN*17b] Alemzadeh S., Hielscher T., Niemann U., Cibulski L., Ittermann T., Völzke H., Spiliopoulou M., Preim B.: Visual subpopulation discovery and validation in cohort study data. *CoRR abs/1711.09377* (2017). URL: http://arxiv.org/abs/1711.09377

[AKMS07] Assent I., Krieger R., Müller E., Seidl T.: VISA: Visual subspace clustering analysis. *SIGKDD Explorations 9*, 2 (2007), 5–12.

[Amo20] Amory W. C.-E.: The untilled field of public health. *Modern Medicine 2* (1920), 183–191.

[AMST11] Aigner W., Miksch S., Schumann H., Tominski C.: *Visualization of Time-Oriented Data*. Springer Science & Business Media, Berlin, Germany, 2011.

[ANI*17] Alemzadeh S., Niemann U., Ittermann T., Völzke H., Schneider D., Spiliopoulou M., Preim B.: Visual analytics of missing data in epidemiological cohort studies. In *Proceedings of VCBM* (2017), pp. 43–52.

[ANI*19] Alemzadeh S., Niemann U., Ittermann T., Völzke H., Schneider D., Spiliopoulou M., Bühler K., Preim B.: Visual analysis of missing values in longitudinal cohort study data. *Computer Graphic Forum* (2019).

[AOH*14] Angelelli P., Oeltze S., Haasz J., Turkay C., Hodneland E., Lundervold A., Lundervold A. J., Preim B., Hauser H.: Interactive visual analysis of heterogeneous cohort-study data. *IEEE CG&A 34*, 5 (2014), 70–82.

[ASL*14] Accorsi P., Sallaberry A., Lalande N., Bringay S., Le Ber F., Poncelet P., Fabrègue M., Cernesson F., Braud A., Teisseire M.: HydroQual: Visual analysis of river water quality. In *Proceedings of IEEE Visual Analytics Science and Technology* (2014), pp. 123–132.

[BAC*09] Băceanu A., Atasiei I., Chazard E., Leroy N., PSIP Consortium: The expert explorer: A tool for hospital data visualization and adverse drug event rules validation. *Studies in Health Technology and Informatics 148* (2009), 85–94.

[BAHJ08] Beale L. L., Abellan J. J., Hodgson S. S., Jarup L. L.: Methodologic issues and approaches to spatial epidemiology. *Environmental Health Perspectives 116*, 8 (2008), 1105–1110.

[BBJ*17] Bannach A., Bernard J., Jung F., Kohlhammer J., May T., Scheckenbach K., Wesarg S.: Visual analytics for radiomics: Combining medical imaging with patient data for clinical research. In *Proceedings of IEEE Workshop on Visual Analytics in Healthcare (VAHC)* (2017), pp. 84–91.

[BBK*15] Basole R. C., Braunstein M. L., Kumar V., Park H., Kahng M., Chau D. H. P., Tamersoy A., Hirsh D. A., Serban N., Bost J., Lesnick B., Schissel B. L., Thompson M.: Understanding variations in pediatric asthma care processes in the emergency department using visual analytics. *Journal of the American Medical Informatics Association 22*, 2 (2015), 318–323.

[BBP*15] Brussoni M., Brunelle S., Pike I., Sandseter E. B. H., Herrington S., Turner H., Belair S., Logan L., Fuselli P., Ball D. J.: Can child injury prevention include healthy risk promotion? *Injury Prevention 21*, 5 (2015), 344–347.

[BDD14] Bhavnani S. K., Drake J., Divekar R.: The role of visual analytics in asthma phenotyping and biomarker discovery. In *Heterogeneity in Asthma*. A. R. Brasie (Ed.). Springer, Berlin, Germany (2014), pp. 289–305.

[Ber66] Bertin J.: *Sémiologie Graphique: Diagrammes, Réseaux, Cartographie*. Mouton, Paris, 1966.

[BGL99] Brownson R. C., Gurney J. G., Land G. H.: Evidence-based decision making in public health. *Journal of Public Health Management and Practice 5*, 5 (1999).

[BGO10] Buuren S. V., Groothuis-Oudshoorn K.: mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software 45*, 3 (2010), 1–68.

[BKR*17] Blascheck T., Kurzhals K., Raschke M., Burch M., Weiskopf D., Ertl T.: Visualization of eye tracking data: A taxonomy and survey. *Computer Graphics Forum 36*, 8 (2017), 260–284.

[BKWea15] Bamberg F., Kauczor H.-U., Weckbach S., Schlett C. L., Forsting M., Ladd S. C., Greiser K. H., Weber M. A., Schulz-Menger J., Niendorf T., Pischon T., Caspers S., Amunts K., Berger K., Bülow R., Hosten N., Hegenscheid K., Kröncke T., Linseisen J., Günther M., Hirsch J. G., Köhn A., Hendel T., Wichmann H. E., Schmidt B., Jöckel K. H., Hoffmann W., Kaaks R., Reiser M. F., Völzke H.; German National Cohort MRI Study Investigators: Whole-body MR imaging in the German National Cohort: Rationale, design, and technical background. *Radiology 277*, 1 (2015), 206–220.

[BSB*15] Bernard J., Sessler D., Bannach A., May T., Kohlhammer J.: A visual active learning system for the assessment of patient well-being in prostate cancer research. In *Proceedings of IEEE Workshop on Visual Analytics in Healthcare (VAHC)* (2015), pp. 1–8.

[BSM*15] Bernard J., Sessler D., May T., Schlomm T., Pehrke D., Kohlhammer J.: A visual-interactive system for prostate cancer cohort analysis. *IEEE CG&A 35*, 3 (2015), 44–55.

[BWMM15] Bryan C., Wu X., Mniszewski S., Ma K.-L.: Integrating predictive analytics into a spatiotemporal epidemic simulation. In *Proceedings of IEEE Visual Analytics Science and Technology* (2015), pp. 17–24.

[BZKF13] Bieh-Zimmert O., Koschtial C., Felden C.: Representing multidimensional cancer registry data. In *Proceedings of Knowledge Management and Knowledge Technologies* (2013), ACM, p. 35.

[CAD*14] Carroll L. N., Au A. P., Detwiler L. T., Fu T.-c., Painterd I. S., Abernethy N. F.: Visualization and analytics tools for infectious disease epidemiology: A systematic review. *Journal of Biomedical Informatics 51*, Supplement C (2014), 287–98.

[CCN09] Castronovo D. A., Chui K. K., Naumova E. N.: Dynamic maps: A visual-analytic methodology for exploring spatiotemporal disease patterns. *Environmental Health 8*, 1 (2009), 61–69.

[CFB*11] Chazard E., Ficheur G., Bernonville S., Luyckx M., Beuscart R.: Data mining to generate adverse drug events detection rules. *IEEE Transactions on Information Technology in Biomedicine 15*, 6 (2011), 823–830.

[CGM*17] Ceneda D., Gschwandtner T., May T., Miksch S., Schulz H.-J., Streit M., Tominski C.: Characterizing guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 111–120.

[CKSB11] Colborn T., Kwiatkowski C., Schultz K., Bachran M.: Natural gas operations from a public health perspective. *Human and Ecological Risk Assessment: An International Journal 17*, 5 (2011), 1039–1056.

[Con14] Consortium G. N. C. G.: The German national cohort: Aims, study design and organization. *European Journal of Epidemiology 29*, 5 (2014), 371–82.

[CRN*08] Chen J., Roth R. E., Naito A. T., Lengerich E. J., MacEachren A. M.: Geovisual analytics to enhance spatial scan statistic interpretation: An analysis of us cervical cancer mortality. *International Journal of Health Geographics 7*, 1 (2008), 57.

[Cul14] Culotta A.: Estimating county health statistics with Twitter. In *CHI-Conference-2A* (2014), pp. 1335–1344.

[CWCN11] Chui K. K. H., Wenger J. B., Cohen S. A., Naumova E. N.: Visual analytics for epidemiologists: Understanding the interactions between age, time, and disease with multi-panel graphs. *PLoS One 6*, 2 (2011), e14683.

[DBR*97] Devalia J., Bayram H., Rusznak C., Calderon M., Sapsford R., Abdelaziz M., Wang1 J., Davies R.: Mechanisms of pollution-induced airway disease: *In vitro* studies in the upper and lower airways. *Allergy 52* (1997), 45–51.

[DH56] Doll R., Hill A. B.: Lung cancer and other causes of death in relation to smoking. *British Medical Journal 2*, 5001 (1956), 1071–1081.

[DMAF15] Dang T. N., Murray P., Aurisano J., Forbes A. G.: ReactionFlow: An interactive visualization tool for causality analysis in biological pathways. *BMC Proceedings 9*, 6 (2015), S6.

[DMPM15] Dunne C., Muller M., Perra N., Martino M.: VoroGraph: Visualization tools for epidemic analysis. In *Proceedings of ACM Conference Extended Abstracts on Human Factors in Computing Systems* (2015), pp. 255–258.

[DMW*16] Du Y., Ma C., Wu C., Xu X., Guo Y., Zhou Y., Li J.: A visual analytics approach for station-based air quality data. *Sensors 17*, 1 (2016), 30.

[Don06] Donders A. R.: Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology 59*, 10 (2006), 1087–1091.

[Dre12] Dredze M.: How social media will change public health. *IEEE Intelligent Systems 27*, 4 (2012), 81–84.

[DRK*94] DiBiase D., Reeves C., Krygier J., MacEachren A. M., von Wyss M., Sloan J. L., Detweiler M. C.: Multivariate display of geographic data: Applications in earth system science. In *Modern Cartography Series*. A. M. Maceachren and F. Taylor (Eds.). Academic Press (1994), vol. 2, pp. 287–312.

[DS07] DeChello L. M., Sheehan T. J.: Spatial analysis of colorectal cancer incidence and proportion of late-stage in Massachusetts residents: 1995–1998. *International Journal of Health Geographics 6*, 1 (2007), 20.

[EKS*96] Ester M., Kriegel H.-P., Sander J., Xu X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (1996), vol. 96, pp. 226–231.

[Eub02] Eubank S.: Scalable, efficient epidemiological simulation. In *Proceedings of the ACM Symposium on Applied Computing* (2002), pp. 139–45.

[EW04] Elliott P., Wartenberg D.: Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives 112*, 9 (2004), 998–1006.

[FCBC15] Ficheur G., Careira L. F., Beuscart R., Chazard E.: EpiHosp: A web-based visualization tool enabling the exploratory analysis of complications of implantable medical devices from a nationwide hospital database. In *Proceedings of Medical Informatics Europe* (2015), pp. 409–413.

[Few06] Few S.: *Information dashboard design*. O'reilly Sebastopol, CA, 2006.

[FF11] Fletcher R. H., Fletcher S. W.: *Clinical Epidemiology*. Lippincott Williams and Wilkins, Philadelphia, 2011.

[FG05] Fabrikant S. I., Goldsberry K.: Thematic relevance and perceptual salience of dynamic geovisualization displays. In *Proceedings of ICA/ACI International Cartographic Conference* (2005).

[Fis93] Fisher P. F.: Visualizing uncertainty in soil maps by animation. *Cartographica: The International Journal for Geographic Information and Geovisualization 30*, 2–3 (1993), 20–27.

[FST*18] Firth J., Stubbs B., Teasdale S. B., Ward P. B., Veronese N., Shivappa N., Hebert J. R., Berk M., Yung A. R., Sarris J.: Diet as a hot topic in psychiatry: A population-scale study of nutritional intake and inflammatory potential in severe mental illness. *World Psychiatry 17*, 3 (2018), 365–367.

[GAK*11] Gschwandtner T., Aigner W., Kaiser K., Miksch S., Seyfang A.: CareCruiser: Exploring and visualizing plans, events, and effects interactively. In *Proceedings of PacificVis* (2011), pp. 43–50.

[GAM*14] Gschwandtner T., Aigner W., Miksch S., Gärtner J., Kriglstein S., Pohl M., Suchy N.: Timecleanser: A visual analytics approach for data cleansing of time-oriented data. In *Proceedings of Knowledge Management and Data-driven Business* (2014), pp. 18:1–18:8.

[GHL*07] Guerra C. A., Hay S. I., Lucioparedes L. S., Gikandi P. W., Tatem A. J., Noor A. M., Snow R. W.: Assembling a global database of malaria parasite prevalence for the malaria atlas project. *Malaria Journal 6*, 1 (2007), 17.

[GI15] Ganna A., Ingelsson E.: 5 year mortality predictors in 498,103 UK Biobank participants: A prospective population-based study. *The Lancet 386*, 9993 (2015), 533–540.

[GLG*12] Gesteland P. H., Livnat Y., Galli N., Samore M. H., Gundlapalli A. V.: The EpiCanvas infectious disease weather map: An interactive visual exploration of temporal and spatial correlations. *Journal of the American Medical Informatics Association 19*, 6 (2012), 954–959.

[Gon00] Gonna J.: *The Global Infectious Disease Threat and Its Implications for the United States*. Tech. Rep., National Intelligence Council, Washington, DC, 2000.

[GTL*12] Gloger O., Tönnies K. D., Liebscher V., Kugelmann B., Laqua R., Völzke H.: Prior shape level set segmentation on multistep generated probability maps of MR datasets for fully automatic kidney parenchyma volumetry. *IEEE Transactions on Medical Imaging 31*, 2 (2012), 312–325.

[Guo07] Guo D.: Visual analytics of spatial interaction patterns for pandemic decision support. *International Journal of Geographical Information Science 21*, 8 (2007), 859–877.

[GWP14] Gotz D., Wang F., Perer A.: A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics 48* (2014), 148–159.

[Har03] Harrower M.: Tips for designing effective animated maps. *Cartographic Perspectives*, *44* (2003), 63–65.

[HB03] Harrower M., Brewer C. A.: Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal 40*, 1 (2003), 27–37.

[HCD*18] Hsu Y.-C., Cross J., Dille P., Nourbakhsh I., Leiter L., Grode R.: Visualization tool for environmental sensing and public health data. In *Proceedings of ACM SIGACCESS Conference on Computers and Accessibility* (2018), pp. 99–104.

[HDS*15] Hripcsak G., Duke J. D., Shah N. H., Reich C. G., Huser V., Schuemie M. J., Suchard M. A., Park R. W., Wong I. C. K., Rijnbeek P. R., van der Lei J., Pratt N., Norén G. N., Li Y. C., Stang P. E., Madigan D., Ryan P. B.: Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. *Studies in Health Technology and Informatics 216* (2015), 574–578.

[HHK00] Hollan J., Hutchins E., Kirsh D.: Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction 7*, 2 (2000), 174–196.

[HKB*11] Honaker J., King G., Blackwell M.: Amelia ii: A program for missing data. *Journal of statistical software 45*, 7 (2011), 1–47.

[HM*87] Hennekens C. H., Mayrent S. L., Buring J. E.: *Epidemiology in Medicine*. Little Brown, Boston, MA, 1987.

[HM96] Howard D., MacEachren A. M.: Interface design for geographic visualization: Tools for representing reliability. *Cartography and Geographic Information Systems 23*, 2 (1996), 59–77.

[HNP*18] Hielscher T., Niemann U., Preim B., Völzke H., Ittermann T., Spiliopoulou M.: A framework for expert-driven subpopulation discovery and evaluation using subspace clustering for epidemiological data. *Expert Systems with Applications 113* (2018), 147–160.

[HRD*16] Hripcsak G., Ryan P. B., Duke J. D., Shah N. H., Park R. W., Huser V., Suchard M. A., Schuemie M. J., DeFalco F. J., Perotte A., Banda J. M., Reich C. G., Schilling L. M., Matheny M. E., Meeker D., Pratt N., Madigan D.: Characterizing treatment pathways at scale using the OHDSI network. In *Proceedings of the National Academy of Sciences 113*, 27 (2016), 7329–7336.

[HSKO14] Hrovat G., Stiglic G., Kokol P., Ojsteršek M.: Contrasting temporal trend discovery for large healthcare databases. *Computer Methods and Programs in Biomedicine 113*, 1 (2014), 251–257.

[HvdKHea06] Heeringa J., van der Kuip D. A., Hofman A., Kors J. A., van Herpen G., Stricker B. H., Stijnen T., Lip G. Y, Witteman J. C.: Prevalence, incidence and lifetime risk of atrial fibrillation: The rotterdam study. *European Heart Journal 27*, 8 (2006), 949–953.

[IBM*17] Ikram M. A., Brusselle G. G. O., Murad S. D., van Duijn C. M., Franco O. H., Goedegebure A., Klaver C. C. W., Nijsten T. E. C., Peeters R. P., Stricker B. H., Tiemeier H., Uitterlinden A. G., Vernooij M. W., Hofman A.: The Rotterdam Study: 2018 update on objectives, design and main results. *European Journal of Epidemiology 32*, 9 (2017), 807–850.

[IFP*12] Isenberg P., Fisher D., Paul S. A., Morris M. R., Inkpen K., Czerwinski M.: Co-located collaborative visual analytics around a tabletop display. *IEEE Transactions on Visualization and Computer Graphics 18*, 5 (2012), 689–702.

[IHN*16] Iqbal U., Hsu C.-K., Nguyen P. A. A., Clinciu D. L., Lu R., Syed-Abdul S., Yang H.-C., Wang Y.-C., Huang C.-Y., Huang C.-W., Chang Y.-C., Hsu M.-H., Jian W.-S., Li Y.-C. J.: Cancer-disease associations: A visualization and animation through medical big data. *Computer Methods and Programs in Biomedicine 127*, Supplement C (2016), 44–51.

[JAK*17] Joshi A., Amadi C., Katz B., Kulkarni S., Nash D.: A hcentered platform for HIV infection reduction in New York: Development and usage analysis of the ending the epidemic (ETE) dashboard. *JMIR Public Health and Surveillance 3*, 4 (2017), e95.

[JCIA10] Johnson M. O., Cohly H. H., Isokpehi R. D., Awofolu O. R.: The case for visual analytics of arsenic concentrations in foods. *International Journal of Environmental Research and Public Health 7*, 5 (2010), 1970–1983.

[JGK10] Jerrett M., Gale S., Kontgis C.: Spatial modeling in environmental and public health research. *International Journal of Environmental Research in Public Health 7*, 16 (2010), 1302–1329.

[JHLea01] John U., Hensel E., Lüdemann J., Piek M., Sauer S., Adam C., Born G., Alte D., Greiser E., Haertel U., Hense H. W., Haerting J., Willich S., Kessler C.: Study of Health in Pomerania (SHIP): A health examination survey in an east German region: Objectives and design. *Sozial- und Präventivmedizin 46*, 3 (2001), 186–194.

[KEV*18] Kwon B. C., Eysenbach B., Verma J., Ng K., deFilippi C., Stewart W. F., Perer A.: Clustervision: Visual supervision of unsupervised clustering. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 142–151.

[KKKW03] Kailing K., Kriegel H.-P., Kroeger P., Wanka S.: Ranking interesting subspaces for clustering high dimensional data. In *Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery* (2003), pp. 241–252.

[KLG*16] Klemm P., Lawonn K., Glaßer S., Niemann U., Hegenscheid K., Völzke H., Preim B.: 3D regression heat map analysis of population study data. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 81–90.

[KLR*13] Klemm P., Lawonn K., Rak M., Preim B., Toennies K., Hegenscheid K., Völzke H., Oeltze S.: Visualization and analysis of lumbar spine canal variability in cohort study data. In *Proceedings of VMV* (2013), pp. 121–128.

[KOL*14] Klemm P., Oeltze-Jafra S., Lawonn K., Hegenscheid K., Völzke H., Preim B.: Interactive visual analysis of image-centric cohort study data. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1673–1682.

[KPS15] Krause J., Perer A., Stavropoulos H.: Supporting iterative cohort construction with visual temporal queries. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 91–100.

[LBI*11] Lam H., Bertini E., Isenberg P., Plaisant C., Carpendale S.: Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics 18*, 9 (2011), 1520–1536.

[LHM*07] Lindley S. J., Handley J. F., McEvoy D., Peet E., Theuray N.: The role of spatial risk assessment in the context of planning for adaptation in UK urban areas. *Built Environment 33*, 1 (2007), 46–69.

[LM14] Luz S., Masoodian M.: Readability of a background map layer under a semi-transparent foreground layer. In *Proceedings of ACM Conference on Advanced Visual Interfaces* (2014), pp. 161–168.

[LRS12] Livnat Y., Rhyne T., Samore M. H.: Epinome: A visual-analytics workbench for epidemiology data. *IEEE CG&A 32*, 2 (2012), 89–95.

[LVF*12] Lim S. S., Vos T., Flaxman A. D., Danaei G., Shibuya K., Adair-Rohani H., Amann M., Anderson H. R., Andrews K. G., Aryee M., Atkinson C., Bacchus L. J., Bahalim A. N., Balakrishnan K., Balmes J., Barker-Collo S., Baxter A., Bell M. L., Blore J. D., Blyth F., Bonner C., Borges G., Bourne R., Boussinesq M., Brauer M., Brooks P., Bruce N. G., Brunekreef B., Bryan-Hancock C., Bucello C., Buchbinder R., Bull F., Burnett R. T., Byers T. E., Calabria B., Carapetis J., Carnahan E., Chafe Z., Charlson F., Chen H., Chen J. S., Cheng A. T., Child J. C., Cohen A., Colson K. E., Cowie B. C., Darby S., Darling S., Davis A., Degenhardt L., Dentener F., Des Jarlais D. C., Devries K., Dherani M., Ding E. L., Dorsey E. R., Driscoll T., Edmond K., Ali S. E., Engell R. E., Erwin P. J., Fahimi S., Falder G., Farzadfar F., Ferrari A., Finucane M. M., Flaxman S., Fowkes F. G., Freedman G., Freeman M. K., Gakidou E., Ghosh S., Giovannucci E., Gmel G., Graham K., Grainger R., Grant B., Gunnell D., Gutierrez H. R., Hall W., Hoek H. W., Hogan A., Hosgood H. D. 3rd, Hoy D., Hu H., Hubbell B. J., Hutchings S. J., Ibeanusi S. E., Jacklyn G. L., Jasrasaria R., Jonas J. B., Kan H., Kanis J. A., Kassebaum N., Kawakami N., Khang Y. H., Khatibzadeh S., Khoo J. P., Kok C., Laden F., Lalloo R., Lan Q., Lathlean T., Leasher J. L., Leigh J., Li Y., Lin J. K., Lipshultz S. E., London S., Lozano R., Lu Y., Mak J.,

Malekzadeh R., Mallinger L., Marcenes W., March L., Marks R., Martin R., McGale P., McGrath J., Mehta S., Mensah G. A., Merriman T. R., Micha R., Michaud C., Mishra V., Mohd Hanafiah K., Mokdad A. A., Morawska L., Mozaffarian D., Murphy T., Naghavi M., Neal B., Nelson P. K., Nolla J. M., Norman R., Olives C., Omer S. B., Orchard J., Osborne R., Ostro B., Page A., Pandey K. D., Parry C. D., Passmore E., Patra J., Pearce N., Pelizzari P. M., Petzold M., Phillips M. R., Pope D., Pope C. A. 3rd, Powles J., Rao M., Razavi H., Rehfuess E. A., Rehm J. T., Ritz B., Rivara F. P., Roberts T., Robinson C., Rodriguez-Portales J. A., Romieu I., Room R., Rosenfeld L. C., Roy A., Rushton L., Salomon J. A., Sampson U., Sanchez-Riera L., Sanman E., Sapkota A., Seedat S., Shi P., Shield K., Shivakoti R., Singh G. M., Sleet D. A., Smith E., Smith K. R., Stapelberg N. J., Steenland K., Stöckl H., Stovner L. J., Straif K., Straney L., Thurston G. D., Tran J. H., Van Dingenen R., van Donkelaar A., Veerman J. L., Vijayakumar L., Weintraub R., Weissman M. M., White R. A., Whiteford H., Wiersma S. T., Wilkinson J. D., Williams H. C., Williams W., Wilson N., Woolf A. D., Yip P., Zielinski J. M., Lopez A. D., Murray C. J., Ezzati M., AlMazroa M. A., Memish Z. A.: A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet 380*, 9859 (2012), 2224–2260.

[MGP*04] MacEachren A. M., Gahegan M., Pike W., Brewer I., Cai G., Lengerich E., Hardisty F.: Geovisualization for knowledge construction and decision support. *IEEE CG&A 24*, 1 (2004), 13–17.

[MGW*18] Meuschke M., Günther T., Wickenhöfer R., Gross M., Preim B., Lawonn K.: Management of cerebral aneurysm descriptors based on an automatic ostium extraction. *IEEE CG&A 38*, 3 (2018), 58–72.

[MHD*14] Mittelstädt S., Hao M. C., Dayal U., Hsu M., Terdiman J., Keim D. A.: Advanced visual analytics interfaces for adverse drug event detection. In *Proceedings of Advanced Visual Interfaces* (2014), pp. 237–244.

[MHLA11] Marcilly R., Hackl W. O., Luyckx M., Ammenwerth E.: Scorecards: A new method to prevent adverse drug events? Preliminary results from a clinical field study. *Health Information Systems 166* (2011), 234–245.

[MHR*09] Maciejewski R., Hafen R., Rudolph S., Tebbetts G., Cleveland W. S., Grannis S. J., Ebert D. S.: Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. *IEEE CG&A 29*, 3 (2009), 18–28.

[MLdO*13] Monroe M., Lan R., del Olmo J. M., Shneiderman B., Plaisant C., Millstein J.: The challenges of specifying intervals and absences in temporal queries: A graphical language approach. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)* (2013), pp. 2349–2358.

[MLK16] Masoodian M., Luz S., Kavenga D.: Nu-view: A visualization system for collaborative co-located analysis of geospatial disease data. In *Proceedings of the Australasian Computer Science Week Multiconference* (2016), p. 48.

[MLR*11] Maciejewski R., Livengood P., Rudolph S., Collins T. F., Ebert D. S., Brigantic R. T., Corley C. D., Muller G. A., Sanders S. W.: A pandemic influenza modeling and visualization tool. *Journal of Visual Languages and Computing 22*, 4 (2011), 268–278.

[Mon06] Monmonier M.: Cartography: Uncertainty, interventions, and dynamic display. *Progress in Human Geography 30*, 3 (2006), 373–381.

[MOSB16] Martinez R., Ordunez P., Soliz P. N., Ballesteros M. F.: Data visualisation in surveillance for injury prevention and control: Conceptual bases and case studies. *Injury Prevention 22*, Suppl 1 (2016), i27–i33.

[MP13] Mühlbacher T., Piringer H.: A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 1962–1971.

[MRH*05] MacEachren A. M., Robinson A., Hopper S., Gardner S., Murray R., Gahegan M., Hetzler E.: Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science 32*, 3 (2005), 139–160.

[MRH*08] Maciejewski R., Rudolph S., Hafen R., Abusalah A. M., Yakout M., Ouzzani M., Cleveland W. S., Grannis S. J., Wade M., Ebert D. S.: Understanding syndromic hotspots - A visual analytics approach. In *Proceedings of IEEE Visual Analytics Science and Technology* (2008), pp. 35–42.

[MRO*12] MacEachren A. M., Roth R. E., O'Brien J., Li B., Swingley D., Gahegan M.: Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2496–2505.

[MTJ*07] Maciejewski R., Tyner B., Jang Y., Zheng C., Nehme R. V., Ebert D. S., Cleveland W. S., Ouzzani M., Grannis S. J., Glickman L. T.: LAHVA: Linked animal-human health visual analytics. In *Proceedings of IEEE Visual Analytics Science and Technology* (2007), pp. 27–34.

[MV13] Marathe M., Vullikanti A. K. S.: Computational epidemiology. *Communications of the ACM 56*, 7 (2013), 88–96.

[NME*13] Nordsborg R. B., Meliker J. R., Ersbøll A. K., Jacquez G. M., Raaschou-Nielsen O.: Space-time clustering of non-Hodgkin lymphoma using residential histories in a Danish case-control study. *PLoS One 8*, 4 (2013), e60800.

[NME*14] Nordsborg R. B., Meliker J. R., Ersbøll A. K., Jacquez G. M., Poulsen A. H., Raaschou-Nielsen O.: Space-time clusters of breast cancer using residential histories: A Danish case–control study. *BMC Cancer 14*, 1 (2014), 255.

[NSVK14a] Niemann U., Spiliopoulou M., Völzke H., Kühn J.: Interactive medical miner: Interactively exploring subpopulations

in epidemiological datasets. In *Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases* (2014), pp. 460–463.

[NSVK14b] Niemann U., Spiliopoulou M., Völzke H., Kühn J.-P.: Subpopulation discovery in epidemiological data with subspace clustering. *Foundations of Computing and Decision Sciences 39*, 4 (2014), 271–300.

[OLK*14] Oeltze S., Lehmann D. J., Kuhn A., Janiga G., Theisel H., Preim B.: Blood flow clustering and applications in virtual stenting of intracranial aneurysms. *IEEE Transactions on Visualization and Computer Graphics 20*, 5 (2014), 686–701.

[OS14] Ola O., Sedig K.: The challenge of big data in public health: An opportunity for visual analytics. *Online Journal of Public Health Informatics 5*, 3 (2014), 223.

[Pea12] Pearce N.: Classification of epidemiological study designs. *International Journal of Epidemiology 41*, 2 (2012), 393–397.

[Ped97] Peden D.: Mechanisms of pollution-induced airway disease: In vivo studies. *Allergy 52* (1997), 37–44.

[PKH*16] Preim B., Klemm P., Hauser H., Hegenscheid K., Oeltze S., Tönnies K. D., Völzke H.: Visual analytics of image-centric cohort studies in epidemiology. In *Visualization in Medicine and Life Sciences III, Towards Making an Impact*. L. Linsen, B. Hamann and H.-C. Hege (Eds.). Springer, Berlin, Germany (2016), pp. 221–248.

[PMBea13] Petersen S. E., Matthews P. M., Bamberg F., Bluemke D. A., Francis J. M., Friedrich M. G., Leeson P., Nagel E., Plein S., Rademakers F. E., Young A. A., Garratt S., Peakman T. C., Sellors J., Collins R., Neubauer S.: Imaging in population science: Cardiovascular magnetic resonance in 100,000 participants of UK Biobank - Rationale, challenges and approaches. *Journal of Cardiovascular Magnetic Resonance 28* (2013), 15–46.

[PMR*96] Plaisant C., Milash B., Rose A., Widoff S., Shneiderman B.: LifeLines: Visualizing personal histories. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)* (1996), pp. 221–227.

[PP95] Pagendarm H. G., Post F. H.: Comparative visualization-approaches and examples. In *Proceedings of Visualization in Scientific Computing* (1995), pp. 95–108.

[PRI18] Preim B., Ropinski T., Isenberg P.: A critical analysis of the evaluation practice in medical visualization. In *Proceedings of Eurographics Workshop on Visual Computing for Biology and Medicine* (2018), pp. 45–56.

[QCX*07] Qu H., Chan W.-Y., Xu A., Chung K.-L., Lau K.-H., Guo P.: Visual analysis of the air pollution problem in Hong Kong. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1408–1415.

[RAM*11] Rind A., Aigner W., Miksch S., Wiltner S., Pohl M., Turic T., Drexler F.: Visual exploration of time-oriented patient data for chronic diseases: Design study and evaluation. In *Proceedings of Conference of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society* (2011), pp. 301–320.

[RET13] Rak M., Engel K., Tönnies K. D.: Closed-form hierarchical finite element models for part-based object detection. In *Proceedings of VMV* (2013), pp. 137–44.

[RFF*08] Robertson G. G., Fernandez R., Fisher D., Lee B., Stasko J. T.: Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1325–1332.

[RMR11] Robinson A., MacEachren A., Roth R.: Designing a web-based learning portal for geographic visualization and analysis in public health. *Health Informatics Journal 17*, 3 (2011), 191–208.

[Rob07] Roberts J. C.: State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of Coordinated and Multiple Views in Exploratory Visualization* (2007), pp. 61–71.

[Ros15] Rosen G.: *A History of Public Health*. JHU Press, Baltimore, MD, 2015.

[RTM*07] Revere D., Turner A. M., Madhavan A., Rambo N., Bugni P. F., Kimball A., Fuller S. S.: Understanding the information needs of public health practitioners: A literature review to inform design of an interactive digital knowledge management system. *Journal of Biomedical Informatics 40*, 4 (2007), 410–421.

[RWA*13] Rind A., Wang T. D., Aigner W., Miksch S., Wongsuphasawat K., Plaisant C., Shneiderman B.: Interactive information visualization to explore and query electronic health records. *Foundations and Trends in Human-Computer Interaction 5*, 3 (2013), 207–298.

[SBCvdS04] Saffer J. D., Burnett V. L., Chen G., van der Spek P.: Visual analytics in the pharmaceutical industry. *IEEE CG&A 24*, 5 (2004), 10–15.

[SGAea15] Sudlow C., Gallacher J., Allen N., Beral V., Burton P., Danesh J., Downey P., Elliott P., Green J., Landray M., Liu B., Matthews P., Ong G., Pell J., Silman A., Young A., Sprosen T., Peakman T., Collins R.: UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine 12*, 3 (2015), e1001779.

[Shn03] Shneiderman B.: The eyes have it: A task by data type taxonomy for information visualizations. In *The Craft of Information Visualization*. B. Bederson and B. Shneiderman (Eds.). Elsevier, Amsterdam, 2003, pp. 364–371.

[SIC*11] Sims J. N., Isokpehi R. D., Cooper G. A., Bass M. P., Brown S. D., St John A. L., Gulig P. A., Cohly H. H.: Visual analytics of surveillance data on foodborne vibriosis, United States, 1973–2010. *Environmental Health Insights 5* (2011), 71.

[SMvB*10] Steenwijk M. D., Milles J., van Buchem M. A., Reiber J. H. C., Botha C. P.: Integrated visual analysis for heterogeneous

datasets in cohort studies. In *Proceedings of IEEE Workshop on Visual Analytics in Healthcare* (2010).

[SND05] Saraiya P., North C., Duca K.: An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics 11*, 4 (2005), 443–456.

[SP06] Shneiderman B., Plaisant C.: Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (2006), pp. 1–7.

[SP19] Shneiderman B., Plaisant C.: Interactive visual event analytics: Opportunities and challenges. *IEEE Computer 52*, 1 (2019), 27–35.

[SPDO12] Sedig K., Parsons P., Dittmer M., Ola O.: Beyond information access: Support for complex cognitive activities in public health informatics tools. *Online Journal of Public Health Informatics 4*, 3 (2012).

[SPG14] Stolper C. D., Perer A., Gotz D.: Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1653–1662.

[SPH13] Shneiderman B., Plaisant C., Hesse B. W.: Improving healthcare with interactive visualization. *IEEE Computer 46*, 5 (2013), 58–66.

[SS05] Seo J., Shneiderman B.: A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization 4*, 2 (2005), 96–113.

[STH02] Stolte C., Tang D., Hanrahan P.: Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 52–65.

[SWC*09] Sterne J. A., White I. R., Carlin J. B., Spratt M., Royston P., Kenward M. G., Wood A. M., Carpenter J. R.: Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal 338* (2009), b2393.

[TC05] Thomas J. J., Cook K.: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society, Los Alamitos, CA, 2005.

[TGR*15] Töennies K. D., Gloger O., Rak M., Winkler C., Klemm P., Preim B., Völzke H.: Image analysis in epidemiological applications. *it - Information Technology 57*, 1 (2015), 22–29.

[TIC09] Tobiasz M., Isenberg P., Carpendale S.: Lark: Coordinating co-located collaboration with information visualization. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1065–1072.

[TK09] Thomas J., Kielman J.: Challenges for visual analytics. *Information Visualization 8*, 4 (2009), 309–314.

[TLLH13] Turkay C., Lundervold A., Lundervold A. J., Hauser H.: Hypothesis generation by interactive visual exploration of heterogeneous medical data. In *Proceedings of Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (2013), pp. 1–12.

[TRE14] Tönnies K., Rak M., Engel K.: Deformable part models for object detection in medical images. *Biomedical Engineering Online 13*, 1 (2014), S1.

[TRL*17] Tong C., Roberts R., Laramee R. S., Berridge D., Thayer D.: Cartographic treemaps for visualization of public healthcare data. In *Proceedings of Computer Graphics and Visual Computing (CGVC)* (2017).

[TSP*09] Thew S., Sutcliffe A., Procter R., de Bruijn O., McNaught J., Venters C. C., Buchan I.: Requirements engineering for E-science: Experiences in epidemiology. *IEEE Software 26*, 1 (2009), 80–87.

[TSS05] Tominski C., Schulze-Wollgast P., Schumann H.: 3D information visualization for time dependent data on maps. In *Proceedings of Information Visualisation* (2005), pp. 175–181.

[VASea11] Völzke H., Alte D., Schmidt C. O., Radke D., Lorbeer R., Friedrich N., Aumann N., Lau K., Piontek M., Born G., Havemann C., Ittermann T., Schipf S., Haring R., Baumeister S. E., Wallaschofski H., Nauck M., Frick S., Arnold A., Jünger M., Mayerle J., Kraft M., Lerch M. M., Dörr M., Reffelmann T., Empen K., Felix S. B., Obst A., Koch B., Gläser S., Ewert R., Fietze I., Penzel T., Dören M., Rathmann W., Haerting J., Hannemann M., Röpcke J., Schminke U., Jürgens C., Tost F., Rettig R., Kors J. A., Ungerer S, Hegenscheid K., Kühn J. P., Kühn J., Hosten N., Puls R., Henke J., Gloger O., Teumer A., Homuth G., Völker U., Schwahn C., Holtfreter B., Polzer I., Kohlmann T., Grabe H. J., Rosskopf D., Kroemer H. K., Kocher T., Biffar R., John U., Hoffmann W.: Cohort profile: The study of health in Pomerania. *International Journal of Epidemiology 40*, 2 (2011), 294–307.

[VMDD03] Venkatesh V., Morris M. G., Davis G. B., Davis F. D.: User acceptance of information technology: Toward a unified view. *MIS Quarterly* (2003), 425–478.

[WDSC07] Wood J., Dykes J., Slingsby A., Clarke K.: Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1176–1183.

[WFR*05] Walhovd K. B., Fjell A. M., Reinvang I., Lundervold A., Dale A. M., Eilertsen D. E., Quinn B. T., Salat D., Makris N., Fischl B.: Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiology of Aging 26*, 9 (2005), 1261–170.

[Win20] Winslow C.-E. A.: The untilled field of public health. *Modern Medicine 2* (1920), 183–191.

[WLB*17] Widanagamaachchi W., Livnat Y., Bremer P., Duvall S. L., Pascucci V.: Interactive visualization and exploration of patient progression in a hospital setting. In *Proceedings of American Medical Informatics Association Annual Symposium* (2017).

[YELL10] Ystad M., Eichele T., Lundervold A. J., Lundervold A.: Subcortical functional connectivity and verbal episodic memory in healthy elderly—Resting state fmri study. *NeuroImage 52*, 1 (2010), 379–388.

[ZG02] Zhang J., Goodchild M. F.: *Uncertainty in Geographical Information*. CRC Press, London, 2002.

[ZGP15] Zhang Z., Gotz D., Perer A.: Iterative cohort analysis and exploration. *Information Visualization 14*, 4 (2015), 289–307.

[ZM16] Zhang S., Metaxas D.: Large-Scale medical image analytics: Recent methodologies, applications and future directions. *Medical Image Analysis 33*, Supplement C (2016), 98–101.

[ZS17] Zakkar M., Sedig K.: Interactive visualization of public health indicators to support policymaking: An exploratory study. *Online Journal of Public Health Informatics 9*, 2 (2017), e190.

[ZWC*16] Zheng Y., Wu W., Chen Y., Qu H., Ni L. M.: Visual analytics in urban computing: An overview. *IEEE Transactions on Big Data 2*, 3 (2016), 276–296.

[ZYL*17] Zhou Z., Ye Z., Liu Y., Liu F., Tao Y., Su W.: Visual analytics for spatial clusters of air-quality data. *IEEE CG&A 37*, 5 (2017), 98–105.