DOI: 10.1111/cgf.14424

COMPUTER GRAPHICS *forum*
Volume 0 (2021), number 0 pp. 1–37

EHR STAR: The State-Of-the-Art in Interactive EHR Visualization

Q. Wang and R.S. Laramee

School of Computer Science, University of Nottingham, UK

Abstract

Since the inception of electronic health records (EHR) and population health records (PopHR), the volume of archived digital health records is growing rapidly. Large volumes of heterogeneous health records require advanced visualization and visual analytics systems to uncover valuable insight buried in complex databases. As a vibrant sub-field of information visualization and visual analytics, many interactive EHR and PopHR visualization (EHR Vis) systems have been proposed, developed, and evaluated by clinicians to support effective clinical analysis and decision making. We present the state-of-the-art (STAR) of EHR Vis literature and open access healthcare data sources and provide an up-to-date overview on this important topic. We identify trends and challenges in the field, introduce novel literature and data classifications, and incorporate a popular medical terminology standard called the Unified Medical Language System (UMLS). We provide a curated list of electronic and population healthcare data sources and open access datasets as a resource for potential researchers, in order to address one of the main challenges in this field. We classify the literature based on multidisciplinary research themes stemming from reoccurring topics. The survey provides a valuable overview of EHR Vis revealing both mature areas and potential future multidisciplinary research directions.

Keywords: visualization, interaction, interaction techniques, information visualization, electronic health records, visual analytics

1. Introduction and Motivation

Several healthcare data institutes strive to exploit software-based technology to study and ultimately improve a nation's collective health [Bus04, Act09, FJV*09, Lar14, Eur17]. Due to the large volume of heterogeneous electronic healthcare data, researchers incorporate techniques such as Machine Learning (ML), Event Sequence Simplification (ESS) and Natural Language Processing (NLP) with interactive visualization and visual analytics, in order extract useful information enabling healthcare providers to obtain a more comprehensive understanding of the underlying patterns and behaviour related to health [SNLOB17]. Those insights provide a useful context in assisting in the optimization of diagnostic and treatment process for both individuals and cohorts of patients, the evaluation of quality and effectiveness of healthcare services, and the prevention of a future public health crisis [FP10].

We present a state-of-the-art (STAR) report of research literature focusing on visualization of EHRs and PopHRs to address these ongoing trends. The contributions we provide to the field in this EHR STAR include:

- An up-to-date overview of recent EHR Vis literature featuring a concise overview of important terminology and recent research in the field, with 213 related references and 19 tables,
- Novel classifications of 51 EHR Vis literature based on six reoccurring research themes and the Unified Medical Language System (UMLS),
- A survey of 34 high quality open access healthcare data sources and datasets,
- An EHR STAR that appeals to researchers from visualization, visual analytics, healthcare, biomedical and related disciplines,
- An overview of future challenges and open research directions in the field.

We have developed an online EHR STAR literature browser for the readers: https://ehr.wangqiru.com. It features all of the EHR papers and datasets along with several filtering and sorting options based on author, year, technique and search terms. We believe it offers a valuable resource for those interested in this topic.

1.1. Survey challenges

This section describes the challenges in the field of EHR Vis and in conducting a survey of related literature. We face a number of challenges stemming from the related literature search.

Diverse literature sources: As literature is spread across conferences and journals from different communities, researchers struggle to keep up with the latest published work. This also increases the time and effort required to identify solved and unsolved problems.

Multidisciplinary research themes: A well-defined classification and scope to organize relevant literature is challenging due to multidisciplinary research themes. As the complexity of research grows, cross-disciplinary collaborations are fostered, and the literature on EHR Vis often spans multiple themes. Different combinations of research expertise produce papers that may be difficult to classify. A typical EHR Vis project might involve visualization, Natural Language Processing (NLP) and Machine Learning (ML).

Inconsistent Medical terminology: The choice of medical terminology standard varies between authors, this increases the work required to classify literature and the difficulty to provide a concise overview of recent research in the field. We address the challenge directly by adopting a medical terminology standard, UMLS, in Section 2.2, and presenting a list of standardized terminology and definitions used in the related literature in Section 1.4.

We also face a number of challenges stemming from digital healthcare data.

Healthcare data acquisition: It is generally challenging to find open and accessible healthcare datasets for conducting research in the field, due to the sensitive nature of the data [MIT16]. There are a number of ways of acquiring electronic healthcare datasets:

1. **Cooperation with relevant health institutes:** This can be the ideal situation but not every researcher has the opportunity to work closely with a relevant institute and obtain access to electronic healthcare data.
2. **Open Access datasets:** There are a number of open access datasets available online. In order to address this challenge directly, we classify and describe them in Section 6. However, the challenge with such datasets is that the access may be restrictive. EHRs may be redacted and lack some data dimensions that are important for EHR Vis research. Based on our investigation, some datasets are old and outdated.
3. **Proprietary datasets:** A license to access proprietary datasets can be expensive. We provide some example license costs in Section 6.5.4 where we describe some proprietary datasets.

Data protection: Electronic healthcare data contains highly sensitive information that requires extra precaution during analysis. Researchers and institutes must comply with the laws and regulations such as HITECH [Act09] and GDPR [Eur16]. This increases the difficulty in data acquisition for research.

Data heterogeneity: Electronic healthcare data is heterogeneous, it may include free text, scalar, ordinal, images and categorical attributes in one record [PL20].

Scalability: The size of an electronic healthcare dataset is often huge. The rate of data growth exceeds the capacity of algorithms and software developed to visualize it [AKDA15].

High-dimensionality: Closely related to heterogeneity, healthcare datasets are high-dimensional and complex [GS14, TRL*17]. The ability to visualize large datasets with many attributes effectively remains a challenging problem [AKDA15].

We address some of these challenges directly in this STAR in Section 6, which includes a survey of open access electronic healthcare data sources, with a dedicated list of Data References. We also present related future challenges in the field in Section 7.

1.2. Literature search methodology

We started our literature search primarily on papers from the following conferences and journals:

- **VIS:** IEEE VIS conferences
- **EuroVis:** EuroVis conferences
- **TVCG:** We have carefully selected papers on EHR Vis from the *IEEE Transactions on Visualization and Computer Graphics* journal
- **VAHC:** Literature published in the *IEEE Workshop on Visual Analytics in Healthcare* is also reviewed, since VAHC primarily focuses on applying interactive visualization techniques for healthcare data

After the initial search and looking into the references, we found more literature from venues listed in Table 1.

We first conduct a breadth-first search. Table 2 shows the list of keyword combinations we use for our breadth-first literature search. We use IEEE Xplore [IEE], The ACM Digital Library [Thea], Google Scholar [Gooc], Vispubdata [IHK*17], Semantic Scholar [The19], Mendeley [Els] and Research Gate [Res] as digital libraries and tools for searching. Previous surveys serve as a good starting point for finding papers on topics of interest. Cross-referencing the extensive Survey of Surveys by McNabb and Laramee [ML17], we find another two related surveys on EHR Vis [RWA*13, WBH15].

We then conduct a depth-first search on the results obtained from the breadth-first search. We review each paper to find other relevant research including:

- The previous related work section and its references
- Mendeley's [Els] "related documents" functions
- The "cited by" function provided by Google Scholar [Gooc] and Semantic Scholar [The19] to discover forward-looking related papers

1.3. Survey scope

In this section, we describe the scope of the survey. Due to the large volume of publications related to EHR Vis, we apply constraints to narrow down the list of literature. We describe those constraints below in this section.

Table 1: *Conferences and Journals (both Visualization and Non-Visualization venues) used for discovering literature and the number of papers found.*

Source (Visualization Venues)	Years	No. of Papers
IEEE Transactions on Visualization and Computer Graphics	2009-2020	16
IEEE Workshop on Visual Analytics in Healthcare	2011, 2014, 2015, 2017	4
EGUK Computer Graphics & Visual Computing	2017, 2018	3
IEEE Workshop on Visualization of Electronic Health Records	2014	2
The Annual EuroVis Conference and Computer Graphics Forum	2015, 2016, 2019	3
IEEE Conference on Visual Analytics Science and Technology	2006	1
IEEE Pacific Visualization Symposium	2011	1
IEEE Computer Graphics and Applications	2015	1
Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications	2016	1
The Visual Computer	2021	1
Total	**2006-2021**	**33**

Source (Non-Visualization Venues)	Years	No. of Papers
ACM Human Factors in Computing Systems	2004, 2010, 2011	3
American Medical Informatics Association Annual Symposium	1998 & 2011	2
Methods of Information in Medicine	2001	1
Conference on Advanced Visual Interfaces	2004	1
Journal of Universal Computer Science	2005	1
IEEE Transactions on Information Technology in Biomedicine	2007	1
Information	2009	1
Ergonomics and Health Aspects of Work with Computers	2011	1
BMC Public Health	2012	1
Government Information Quarterly	2012	1
Computer Methods and Programs in Biomedicine	2013	1
Online Journal of Public Health Informatics	2016	1
Journal of the American Medical Informatics Association	2018	1
Bioinformatics	2019	1
ACM Transactions on Computing for Healthcare	2020	1
Total	**1998-2020**	**18**

Table 2: *Keyword combinations used for discovering EHR Vis literature.*

Search Keywords	Additional Keywords
Visualization	electronic health record, electronic medical record, EHR, EMR personal health record, population health record, PHR, PopHR clinical decision support healthcare, health care, clinical, medical medicine, treatment, surgery, hospital

In Scope: In this STAR, we focus on EHR and PopHR Vis as defined in Section 1.4.

We include peer-reviewed literature focusing on real-world scenarios and empirical applications of EHR Vis. We emphasize research with healthcare data collected through clinical practice and that which provides clinical decision support.

Novel techniques are also included. We include Event Sequence Simplification (ESS), a widely adopted technique to provide succinct visual layouts [MLL*13] hidden in EHR data-related processes. We include papers on EHR Vis with geospatial visualization, as a geographical dimension might be relevant in a PopHR dataset. Geospatial visualization partially overlaps with this survey. We include research describing EHR Vis with Natural Language Processing (NLP) techniques. Friedman and Hripcsak recognize text visualization with NLP as one of the most commonly used tools to extract information from EHR data and for studying clinical and research questions [FH99]. We also include papers describing EHR Vis systems developed with Machine Learning (ML) and data mining techniques, as they have gained traction in their applications in assisting clinical research [WZM*16].

We focus on papers published in the previous 10 years. We refer to these papers as *focus papers*. Older papers such as LifeLines [PMR*96], LifeLines2 [PMS*98b] and PatternFinder [FKSS06], contribute significantly to the field, with mature implementations deployed in clinical practices. We still include them as *context papers* and in the meta-data such as the classification Table 3, without a detailed description. By considering the publication year, we are able to investigate the fields that are less mature and provide more accurate future research directions.

Out of Scope: We introduce the following criteria to constrain the scope of this STAR.

Non peer-reviewed publications: We exclude papers that are not peer-reviewed.

Resource-oriented: We exclude papers focusing on the visualization of related resource-oriented EHR data. We define resource-oriented EHR data as the data that focuses on the management of clinical practices, such as hospital bed occupancy rates and inter/intra-hospital patient transfer times. These studies generally do not focus on the clinical decision support directly.

Off-topic: We exclude papers that focus on the use of EHR in the study of disease relations and pathogen outbreaks.

Basic visual designs: In order to focus on novel and interactive visualization techniques, we exclude papers that describe EHR Vis with very basic, static visual designs such as a pie chart, line chart, bar chart or bubble chart. Including classic, static visual designs does not advance the state of the art.

Off-the-shelf solutions: We exclude papers that use off-the-shelf solutions to generate images. In generally, they do not propose a novel visualization technique. We also exclude papers that demonstrate visual designs but do not provide a custom-built solution.

1.4. Background and terminology

Healthcare-related terminology is one of the challenges in the literature. We address this challenge by studying some of the must popular terms used in the literature. Here we provide and classify the terminology used in this STAR.

EHR: To the best of our knowledge, there is no standard definition of an Electronic Health Record (EHR) even since its inception in the 1960s [MIT16]. Iakovidis defines EHR as digitized healthcare information on individual patients that is accessible, secure and highly usable for supporting the analysis of healthcare, education and research [Iak98]. Gunter and Terry define EHR as, *"A longitudinal collection of electronic health information about individual patients and populations"* [GT05], p.1]. The U.S. National Cancer Institute defines EHR as, *"An electronic (digital) collection of medical information about a person that is stored on a computer"* [Nata]. The U.S. Centers for Medicare and Medicaid Services defines EHR as, *"An electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the key administrative clinical data relevant to that persons care under a particular provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports"* [Thee]. The World Health Organization (WHO) defines EHR as, *"Health records residing in an electronic system specifically designed for data collection, storage, and manipulation, and to provide safe access to complete data about patients"* [WP17], p.16].

In this STAR we define EHR as a longitudinal collection of comprehensive patient medical information in machine readable formats, that is maintained and shared by healthcare providers, and stored securely in an electronic system.

EMR: EHR and Electronic Medical Record (EMR) are sometimes used interchangeably to represent digitized health records used to improve quality of care and estimate costs [ZAR*11, Eva16, GBMG17, STBR18]. Unlike EHR, an EMR is stored and used internally without inter-organization sharing [HBAS17]. For purposes of this STAR, we group EMR terminology and literature into the EHR category.

PHR: To the best of our knowledge, a definition of Personal Health Record (PHR) was first proposed in the early 2000s with Tang et al. [TAB*06] stating that a PHR differs from an EHR by its accessibility. A PHR is managed by the data owner and is authorized for sharing with healthcare providers when necessary [TAB*06]. The U.S. National Cancer Institute defines PHR as, *"A collection of information about a person's health that allows the person to manage and track his or her own health information"* [Natb]. The NHS classifies a medical record as a PHR if it is secure, usable and available online whilst being managed by the person who the record represents [NHSb].

PopHR: Population Health Record (PopHR) is first defined by Friedman and Parrish as, *"A repository of statistics, measures, and indicators regarding the state of and influences on the health of a defined population, in computer processable form, stored and transmitted securely, and accessible by multiple authorized users"* in 2010 [FP10], p.360]. A PopHR dataset focuses on the health of a population, without storing identifiable information on individual members of the population. We make a distinction between EHR and PopHR in this survey. The research focusing on PopHR is summarized in Section 4.6.

EHR Visualization: We consider the visualization of EHR and PopHR for clinical decision support, as a sub-field of information visualization and visual analytics (EHR Vis), with the following definitions.

2. Literature Classification

This section describes our literature classification method. We derive classification dimensions based on:

- Recurring multidisciplinary research themes derived from our literature search, described in Section 2.1.
- The Unified Medical Language System (UMLS), introduced in Section 2.2, as the medical terminology standard for classifying literature.

Table 3: ***UMLS table:*** *Classification table of the reviewed literature. We extract keywords used in each paper in order to retrieve the UMLS code and terminology via the UMLS Browser [Bod04]. Keywords are only indicated where they differ from the UMLS term. Papers are grouped by UMLS Code on the y-axis and by the number of EHR documents visualized on the x-axis. Green highlights context papers included in this STAR.*

			Number of Electronic Health Records Visualized					
UMLS Code	**UMLS Term**	**Keywords**	**1**	**2 - 100**	**101 - 1,000**	**1,001 - 5,000**	**5,001 - 100,000**	**>100,000**
C0003125	Anorexia nervosa			[HMA*05]				
C0004238	Atrial fibrillation			[FUS*15]				
C0007222	Cardiovascular diseases	Cardiovascular disease					[JCG*20] (7,537 patients with 64,269 EHRs)	
C0009378	Colonoscopy	Colonoscopy, biopsy, appendiceal-orifice			[TPC*18]			
C0010337	Care of intensive care unit patient	Critical care	[BSM04]	[HPU01] [SSBC19]		[MDM*14a] (1,178 EHRs shown)	[GJG*19] (46,251 patients)	
C0011847	Diabetes		[RMA*10]		[GSCE11]			
C0011854	Diabetes dellitus, insulin-dependent	Type 1 diabetes	[ZCD19]		[KAS*20]			[KCK*19] (1.4 million patients)
C0014544	Epilepsy				[WLLP21]			
C0018802	Congestive heart failure						[WG12] (6,328 EHRs)	
C0020179	Huntington disease	Huntington's disease			[KAS*20]			
C0020443	Hypercholesterolemia							[KCK*19]
C0020538	Hypertensive disease	Hypertensive						[KCK*19]
C0021400	Influenza		[AM12] [RGT*13]					
C0021711	Neonatal intensive care		[KPT*14]	[HPU01] [KCJM16]				
C0023981	Longitudinal Studies	Longitudinal cohort study				[ANI*19]		
C0024117	Chronic obstructive airway disease	Chronic obstructive pulmonary disease					[GXZ*18] (5,804 patients)	
C0030567	Parkinson disease	Parkinson's disease			[KAS*20]			
C0030677	Patient care management		[GSG*04]			[MDM*14a]		
C0030704	Patient transfer						[WGP*11] (7,041 patients)	
C0031330	Pharmacology	Pharmacovigilance			[MLL*13]			
C0031437	Phenotype			[GHC*16] [GGC*17]		[GND*18]		
C0032285	Pneumonia		[AM12]					
C0034065	Pulmonary embolism		[BSM04]					
C0035242	Respiratory tract diseases	Respiratory diseases					[JCG*20]	
C0038454	Cerebrovascular accident	Stroke			[LPK*16]			
C0040034	Thrombocytopenia				[WPS*09]			
C0085207	Gestational diabetes	Gestational diabetes mellitus		[FUS*15]				
C0262926	Medical history		[ZAR*11]	[SSBC19]		[GS14] (2,899 patients)	[FKSS06] (950 patients, over 26,000 EHRs)	[JFB*16] (439,547 patients, 833,710 EHRs) [KCK*19]
C0441472	Clinical action		[GAK*11]					
C0599880	Treatment plan		[BSM04] [FN11] [GAK*11] [ZCD19]				[JCG*20]	
C0600139	Prostate carcinoma	Prostate cancer				[BSKR19] (almost 2,000 patients)	[BSB*15] (about 16,000 patients) [BSM*15] (about 16,000 patients)	
C0679831	Patient history	Patient's history	[PMS*98b] [BAK07] [DJC17]				[BSB*15] [BSM*15]	
C0684249	Carcinoma of lung	Lung cancer	[BWH14]					
C0872379	Disease subtype	Disease subtyping				[GND*18]		
C1659543	Breast Density				[KLG*15]			
C2711227	Steatohepatitis	Hepatic Steatosis			[KLG*15]			
C3242284	Population health				[KLG*15]	[TML*17] [TRL*17] [TML18] [ML19]		[SNK*12] [OS16] (12 million EHRs)
C5204342	Clinical history	Patient clinical history	[GOT*19]	[FUS*15]				

Table 4: ***Terminology table:*** *Terminology used in each focus and context paper included in this STAR, order by year of publication. The x-axis indicates the terminology used in each paper, and their subject category is described in Section 3. This table indicates a mixture of terms are used throughout the literature. We clarify the terminology in Section 1.4. Green highlights context papers.*

				Is published in	
Literature	**EHR**	**EMR**	**Other Terms**	**Vis community**	**Year**
Plaisant et al. [PMS*98b]			Computerized patient records		1998
Horn et al. [HPU01]	🟥			✓	2001
Bade et al. [BSM04]	🟥				2004
Goren-Bar et al. [GSG*04]			Time-oriented clinical data	✓	2004
Hinum et al. [HMA*05]			Medical data		2005
Fails et al. [FKSS06]			Personal medical history	✓	2006
Bui et al. [BAK07]		🟥		✓	2007
Wang et al. [WPS*09]	🟥			✓	2009
Rind et al. [RMA*10]			Medical data		2010
Faiola and Newlon [FN11]		🟥			2011
Gotz et al. [GSCE11]	🟥			✓	2011
Gschwandtner et al. [GAK*11]			Patient record	✓	2011
Wongsuphasawat et al. [WGP*11]	🟥			✓	2011
Zhang et al. [ZAR*11]	🟥			✓	2011
Alonso and McCormick [AM12]			Public health data		2012
Sopan et al. [SNK*12]			Public health data		2012
Wongsuphasawat and Gotz [WG12]		🟥		✓	2012
Monroe et al. [MLL*13]	🟥			✓	2013
Ramírez-Ramírez et al. [RGT*13]			Public health	✓	2013
Borland et al. [BWH14]			Population health	✓	2014
Gotz and Stavropoulos [GS14]		🟥		✓	2014
Kamaleswaran et al. [KPT*14]		🟥		✓	2014
Malik et al. [MDM*14a]	🟥			✓	2014
Bernard et al. [BSB*15]	🟥			✓	2015
Bernard et al. [BSM*15]	🟥			✓	2015
Federico et al. [FUS*15]	🟥			✓	2015
Klemm et al. [KLG*15]			Population health	✓	2015
Glueck et al. [GHC*16]	🟥			✓	2016
Jiang et al. [JFB*16]	🟥			✓	2016
Kamaleswaran et al. [KCJM16]		🟥		✓	2016
Loorak et al. [LPK*16]	🟥			✓	2016
Ola and Sedig [OS16]		🟥			2016
Dabek et al. [DJC17]	🟥			✓	2017
Glueck et al. [GGC*17]	🟥			✓	2017
Tong et al. [TML*17]			Public healthcare data	✓	2017
Tong et al. [TRL*17]			Public healthcare data	✓	2017
Glueck et al. [GND*18]	🟥			✓	2018
Guo et al. [GXZ*18]	🟥			✓	2018
Tong et al. [TML18]			Public healthcare data	✓	2018
Trivedi et al. [TPC*18]	🟥				2018
Alemzadeh et al. [ANI*19]			Longitudinal cohort study	✓	2019
Bernard et al. [BSKR19]	🟥			✓	2019
Glicksberg et al. [GOT*19]	🟥				2019
Guo et al. [GJG*19]	🟥			✓	2019
Kwon et al. [KCK*19]		🟥		✓	2019
McNabb and Laramee [ML19]			Population health data	✓	2019
Sultanum et al. [SSBC19]		🟥		✓	2019
Zhang et al. [ZCD19]	🟥			✓	2019
Jin et al. [JCG*20]	🟥				2020
Kwon et al. [KAS*20]		🟥		✓	2020
Wang et al. [WLLP21]	🟥			✓	2021
Total unique papers: 51	25	10	16	40	

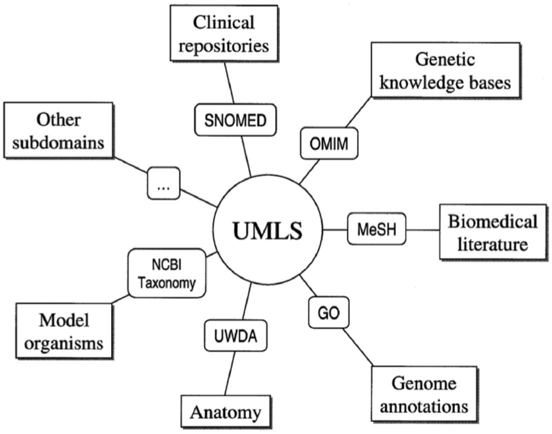


Figure 1: *The various subdomains integrated in the UMLS Terminology. Image courtesy of [Bod04].*

2.1. Multidisciplinary research themes

EHRs are often large-scale and may contain noisy data [CXR18]. This means an automated process can be implemented in order to achieve both efficiency and accuracy in the pre-processing and visualization stages. From the related literature, we have identified several major research themes in processing and visualizing EHRs. We provide a brief description of these themes here and review the related literature in detail in Section 4.

- Machine Learning (ML)
- Natural Language Processing (NLP)
- Event Sequence Simplification (ESS)
- Geospatial Visualization (GEO)
- Visual Analytics with Clustering
- Visual Analytics with Comparison

Table 8 shows an overview of our literature classification based on multidisciplinary research themes. We describe Table 8 in Section 4 on the visualization of EHR data.

2.2. Adopting a medical terminology standard

Gesulaga et al. identify one of the primary barriers to the adoption and deployment of EHR Vis systems in a clinical environment as stemming from resistance from clinical professionals due to the lack of expertise in computer systems including visualization [GBMG17]. By adopting a medical terminology standard, we hope to bridge the gap between two communities, thus reaching a wider audience beyond information visualization and visual analytics, and take advantage of the extensive work invested into the standardized terminology development.

UMLS was introduced by the US National Library of Medicine in 2004. It incorporates a growing list of 2.5 million medical concepts and 12 million relations among these concepts from multiple dictionaries in order to provide a terminology standardization. A schematic of the integrated dictionaries is shown in Figure 1. Dictionaries often use different lexical items to describe identical or similar terms. An integrated standard will make these resources interoperable, machine-readable and help dismantle the barrier to multidisciplinary research [Bod04].

In order to classify each paper, we first extract their keywords to obtain their corresponding code and terminology from the UMLS. Table 3 shows the overview classification of research papers found from our literature search. The x-axis is mapped to the number of EHRs visualized in the corresponding paper. The y-axis is mapped to the corresponding UMLS Code and terminology found along with the keywords appearing in each paper. We can observe from Table 3 the lack of convergence or consolidation with respect to the health conditions addressed in the EHR Vis literature. This is most likely due to the relative immaturity of the field. We also do not observe many research groups working together as a wider team-effort to tackle challenges in the field. And finally we can observe that not many papers are dealing with the really large EHR and PopHR datasets with over 100,000 records.

3. Related Work

This section introduces related work with a special emphasis on previous related surveys. Papers with a focus on visualization or visual analytics of EHR data are described in Section 3.1. We present previous PopHR survey papers in Section 3.2.

Our STAR differs from previous ones by including a novel, up-to-date overview using a medical terminology standard described in Section 2.2, with 29 more recent publications on EHR visualization. Table 5–7 clearly indicate both the overlap and divergence between this STAR and previous surveys. In addition, we introduce a survey of 34 open healthcare data sources in Section 6 to address the challenge of healthcare data access.

3.1. Related work with an EHR focus

In this section, we divide related work with an EHR focus into two sub-categories, related work with an EHR Vis focus and related work without an EHR Vis focus but rather on analysis. We also investigate both the overlap and divergence of the literature presented here with previous surveys, as shown in Table 5–7 for focus papers, context papers, and out of scope papers respectively.

Related Work with an EHR Vis Focus

The IEEE Workshop on Visual Analytics in Healthcare (VAHC) started in 2010 and has been hosted six times at the IEEE VIS conference and four times at the American Medical Informatics Association (AMIA) Annual Symposium. EHR Vis has great potential for influencing the clinical decision-making process and conducting research on epidemiology [Eva16]. The quantity of literature has grown since an early survey published in 2013 by Rind et al. [RWA*13]. There are a number of older related surveys published since then, we present them in this section.

Roque et al. compare six information visualization systems designed for providing overviews of EHR data [RST10]. Systems are classified based on the users, goals, and tasks. Four of these systems

Table 5: ***Focus papers:*** *Y-axis, common Focus papers from previous survey papers, ordered by the year of publication. X-axis,* Ⓔ*indicates an EHR focused survey and* Ⓟ*indicates a PopHR focused survey. We can see that some of previously published EHR Vis papers are common to multiple surveys.*

	Related Work							
Literature	Ⓔ Rind et al. [RWA*13]	Ⓟ Carroll et al. [CAD*14]	Ⓔ West et al. [WBH15]	Ⓔ Gotz and Borland [GB16]	Ⓔ Onukwugha et al. [OPS16]	Ⓔ Rind et al. [RFG*17]	Ⓟ Preim and Lawonn [PL20]	Year
Gotz et al. [GSCE11]	🟥		🟥					2011
wongsuphasawat et al. [WGP*11]	🟥		🟥			🟥		2011
Alonso and Mccormick [AM12]		🟥						2012
Sopan et al. [SNK*12]		🟥						2012
wongsuphasawat and gotz [WG12]	🟥		🟥	🟥		🟥		2012
Monroe et al. [MLL*13]	🟥		🟥		🟥			2013
Ramírez-ramírez et al. [RRGT*13]		🟥						2013
Borland et al. [BWH14]				🟥				2014
Malik et al. [MDM*14a]					🟥			2014
Bernard et al. [BSB*15]							🟥	2015
Bernard et al. [BSM*15]							🟥	2015
Federico et al. [FUS*15]						🟥		2015
	4	3	4	2	3	3	2	
Total unique papers: 12 | Total appearances: 21								

Table 6: ***Context papers:*** *Y-axis, overlapping context papers from previous survey papers, ordered by the year of publication. X-axis,* Ⓔ*indicates an EHR focused survey and* Ⓟ*indicates a PopHR focused survey. We can observe that the 2013 survey by Rind et al. [RWA*13] has some thematic overlap with this one.*

	Related Work						
Literature	Ⓔ Roque et al. [RST10]	Ⓔ Rind et al. [RWA*13]	Ⓔ Simpao et al. [SAGR14]	Ⓔ West et al. [WBH15]	Ⓔ Onukwugha et al. [OPS16]	Ⓔ Rind et al. [RFG*17]	Year
Plaisant et al. [PMS*98a]	🟥	🟥		🟥	🟥		1998
Horn et al. [HPU01]		🟥					2001
Bade et al. [BSM04]		🟥					2004
Goren-bar et al. [GBSGA*04]	🟥	🟥		🟥			2004
Hinum et al. [HMA*05]		🟥				🟥	2005
Fails et al. [FKSS06]		🟥					2006
Bui et al. [BAK07]	🟥						2007
Wang et al. [WPS*09]	🟥	🟥	🟥	🟥			2009
Rind et al. [RMA*10]		🟥				🟥	2010
Faiola and Newlon [FN11]		🟥					2011
	4	9	1	3	1	2	
Total unique papers: 10 | Total appearances: 20							

are included in our survey as context papers (Table 6) and two are excluded with reasons indicated in Table 7.

Rind et al. [RWA*13] review 14 information visualization systems for exploring and querying EHR documents, as shown in Table 5.2 in their work. The survey identifies four major challenges in the field, and highlights the potential that information visualization has on supporting medical tasks. Some 14 systems are compared by (1) supported data types (categorical and numerical), (2) multi-variate support, (3) subject cardinality (support for one patient versus multiple patient records), and (4) supported medical scenarios. Two systems are included in our survey as focus papers (Table 5), nine are included as context papers (Table 6), and seven are papers considered out of scope with reasoning indicated in Table 7.

Simpao et al. [SAGR14] discuss applications of visual analytics in healthcare since the HITECH Act in 2009. The authors review eight visual analytics tools for EHR and categorize their application

Table 7: ***Out of scope papers:*** *Y-axis, out of scope papers from previous survey papers, ordered by the year of publication, with the exclusion criteria described in Section 1.3: (S) Scientific Visualization. (N) Not peer-reviewed. (RO) Resource-oriented system. (OT) Off-topic. (B) Basic visual designs. (OS) Off-the-shelf solution. X-axis,* Ⓔ*indicates an EHR focused survey and* Ⓟ*indicates a PopHR focused survey.*

Literature	Related Work								Exclusion Criteria	Year
	Ⓔ	Ⓔ	Ⓟ	Ⓔ	Ⓔ	Ⓔ	Ⓔ	Ⓟ		
	Roque et al. [RST10]	Rind et al. [RWA*13]	Carroll et al. [CAD*14]	Simpao et al. [SAGR14]	West et al. [WBH15]	Onukwugha et al. [OPS16]	Rind et al. [RFG*17]	Preim and Lawonn [PL20]		
Kosara and Miksch[KM01]	🟥								B	2001
Atkinson and Unwin[AU02]			🟥						OS	2002
Chittaro et al.[CCT03]		🟥							OT	2003
Brodbeck et al.[BGD05]		🟥							B	2005
Aigner and Miksch[AM06]	🟥	🟥					🟥		B, OS	2006
Blanton et al.[BMM*06]			🟥						OT	2006
Da Silva et al.[DGG*07]			🟥						B, OS	2007
Guo[Guo07]								🟥	B, OT	2007
Hu et al.[HZC*07]			🟥						B, OT	2007
Pieczkiewicz et al.[PFH07]		🟥							B	2007
Gao et al.[GMA*08]			🟥						B, OS	2008
Hallett[Hal08]	🟥								B	2008
Heitgerd et al.[HDE*08]			🟥						B	2008
Reinhardt et al.[REA*08]			🟥						B	2008
Yi et al.[YHH*08]			🟥						OS	2008
Bashyam et al.[BHW*09]					🟥				B, S	2009
Connors et al.[CKS*09]		🟥							OT	2009
Wongsuphasawat and Shneiderman[WS09]		🟥							OT	2009
Goldsmith et al.[GTC*10]				🟥					B	2010
Klimov et al.[KST10]		🟥			🟥	🟥			RO	2010
Kumasaka et al.[KNK10]				🟥					OT	2010
Naumova[Nau10]				🟥					OT	2010
Steenwijk et al.[SMB*10]		🟥							S	2010
Willison[Wil10]					🟥				B, N, OS	2010
Driscoll et al.[DGM*11]			🟥						B, OT	2011
Hripcsak et al.[HAP11]					🟥				B, OS	2011
Lewis et al.[LFL*11]			🟥						B	2011
Maciejewski et al.[MLR*11]								🟥	B, OT	2011
Gesteland et al.[GLG*12]								🟥	OT	2012
Joshi and Szolovits[JS12]					🟥				B, OS	2012
Livnat et al.[LRS12]								🟥	B, OT	2012
Mane et al.[MBS*12]				🟥					B	2012
Perer and Sun[PS12]				🟥					B, OS	2012
Stubbs et al.[SKD12]					🟥				B	2012
Rajwan et al.[RBL*13]				🟥					B, OS	2013
Freifeld et al.[FMRB14]			🟥						B, OS	2014
Gálvez et al.[GAS*14]				🟥					B	2014
Simpao et al.[SAD*14]						🟥			B	2014
Dunne et al.[DMPM15]								🟥	B, OT	2015
Masoodian et al.[MLK16]								🟥	B, OT	2016
Caballero et al.[CCDW17]		🟥			🟥	🟥			RO	2017
Abukhodair et al.[AKOS18]		🟥			🟥	🟥			RO	2018
	3	10	11	7	8	4	1	6		
	Total unique papers: 42 | Total appearances: 50									

into different scenarios: (1) using mathematical and algorithmic based processing techniques such as text mining and NLP to derive insight from data, (2) predefined data models to input EHR and output predictive risk assessment results for stratifying patients, (3) enhancing EHR systems with more sophisticated rules-based functions, (4) analysing continuous data streams in the nontraditional healthcare environment, such as data transmitted from wearable monitors, (5) aimed at cost-cutting and revenue-generating, such as automated billing and auditing, optimizing resource allocation. From these eight EHR Vis tools, one is included in our survey as a context paper (Table 6), and seven are considered out of scope (Table 7).

West et al. [WBH15] publish a systematic review of 18 papers, by highlighting crucial metrics to evaluate EHR systems. Those metrics include (1) visualization techniques applied to utilize the screen space efficiently while preserving as much data as possible, (2) interactive user options to identify abnormalities within the data, (3) visualization of the entire dataset even if there are missing values or inaccurate data entries, (4) visualization of temporal data including event sequences and real-time data streams, and (5) training time required for users and software. Some 13 EHR systems are described in these 18 papers. We include four as focus papers (Table 5), three as context papers (Table 6) and exclude six papers (Table 7).

Onukwugha et al. [OPS16] publish a survey of EHR Vis for cancer analysis. The authors describe five cancer-related EHR Vis systems followed by two EHR systems in detail with case studies visualizing a prostate cancer archive and a health insurance claim dataset. The authors focus on EHR systems from three perspectives, (1) the ability to identify and rectify errors in data, (2) visualization techniques and interactive options provided to support data analysis, and (3) cogent visualizations generated to present findings to decision-makers. From these seven EHR Vis systems, we include four as focus papers (Table 5), one as a context paper (Table 6) and exclude two papers (Table 7).

Gotz and Borland [GB16] discuss challenges and opportunities for the interactive visualization of EHR, with four EHR Vis systems reviewed in detail. The authors provide a broad range of empirical applications incorporating EHR Vis, (1) Patient-centred point-of-care applications that provide support for clinicians on communication and analysis for a single patient. (2) Patient-facing applications, similar to patient-centred point-of-care applications, providing patient-oriented support via techniques such as storytelling. (3) Population management applications supporting institutional policymakers to allocate healthcare resources intelligently. (4) Health outcomes research that support discovery and insight that generalize across a population at large. We include two as focus papers (Table 5) in our survey and exclude two papers (Table 7).

Rind et al. [RFG*17] publish a survey of EHR Vis with a focus on time-oriented datasets. The authors identify technical challenges arising from the temporal dimension of EHR datasets, as (1) the interpretation of discrete and continuous temporal dimensions, (2) the scalability from a single patient to a cohort of patients and (3) data-processing techniques to address uncertainties caused by data quality. Detailed descriptions of four EHR systems are provided, we include two as focus papers (Table 5) and two as context papers (Table 6).

Related Work with EHR Focus Outside the Visualization Community

To date, we have not found any further related EHR Vis surveys beyond what we describe. However, we found other work related to EHR analysis outside the visualization community with a focus on EHR data.

MIT Critical Data published a related book, *Secondary Analysis of Electronic Health Records* [MIT16]. The first chapter identifies the objective of secondary analysis of EHR data as the utilization of EHR data to provide evidence for informing best practices in clinical care. EHR has comparative advantages in both cost-effectiveness and feasibility. The second chapter reviews three open access EHR databases (as one of them no longer provides open access, we only include two of these databases in Table 17 in Section 6.5 as focus data sources) in detail with compact descriptions of three additional databases with more restrictive access limitations (we exclude these three databases, as two have discontinued and one no longer provides open access).

Chapter three introduces opportunities and challenges in the secondary analysis of EHR. EHR creates novel opportunities for researchers and clinicians, large datasets and queries provide evidence to support hypotheses. The authors identify that scalability and data accessibility as two major challenges in the field, which overlap with our findings in Section 7 and Table 18. Other challenges identified are data protection, data interoperability, the cost of data infrastructure and the varied quality of research outputs. The rest of the book describes techniques in data pre-processing and analysis with example studies conducted using EHR databases reviewed in chapter two.

Shickel et al. [STBR18] survey six ML-EHR systems developed with Deep Learning techniques for predictive analytics using EHRs in detail. In addition, 25 systems are included for comparison and discussion. These systems are divided into two categories based on their applied machine learning techniques: Supervised and Unsupervised, as shown in Figure 3 in Shickel et al. [STBR18] survey. Another classification dimension is derived from the target task and subtasks of previous EHR systems.

Koleck et al. [KDBB19] systematically review 27 systems that adopt NLP algorithms for extracting structured data from free text EHRs. Table 3 in Koleck et al. [KDBB19] shows the classification by the clinical specialty. The survey scope is defined to include symptom science research that focuses on the description, evaluation, or use of an NLP algorithm or pipeline to process or analyse patient symptom terms. Reporting demographic information is essential for NLP-EHR studies, as symptom experience is known to vary by common demographic factors. Reporting such information helps avoid potential bias and improve the effectiveness of tailored interventions. Some 27 systems are evaluated, with eight critical indicators identified by the authors.

3.2. Related work with a PopHR vis focus

This section introduces related work with an emphasis on PopHR, which focuses on the visualization of the health of a population, rather than individuals.

Carroll et al. [CAD*14] publish a systematic review of 88 articles with a primary focus on infectious disease, needs of public health users, or usability of information visualizations. Each article is reviewed and classified into the following six categories with a focus on: (1) information needs and learning behaviour of public health professionals, (2) architecture of tools, (3) user preference with a focus on usability issues and barriers to adoption of tools, (4) features of tools, (5) usability and evaluation and (6) implementation and adoption. These categories are not mutually exclusive, in total 14 EHR systems are reviewed in detail, we include three (Table 5) as focus papers, none as context papers, and exclude 11 with reasons indicated in Table 7.

Preim and Lawonn review the existing visual analytics solutions for supporting Public Health (PH) [PL20] with structured data. The authors describe PH datasets as heterogeneous and high-dimensional, often containing temporal and spatial dimensions, therefore flexible visual analytics solutions will benefit the analysis process and provide support for PH decision-making. The survey classifies these solutions based on commonly used visualization and visual analytics techniques, as shown in Tables 4 and 5 in their work. The survey then expands into three particular areas of PH, (1) analysis and control of epidemics with 8 solutions, (2) visual analytics for epidemiological research with 14 solutions, and (3) visual analytics of population-based cohort study data. We include two (Table 5) as focus papers, none as context papers, and exclude six with reasons indicated in Table 7.

4. Visualization of EHR data

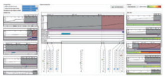
This section describes 41 focus papers on EHR Vis found from our literature search. We further categorize these papers based on six multidisciplinary research themes derived from our investigation, as shown in the Table 8. Each theme is described in this section in detail. We also provide an interactive EHR STAR Browser containing all literature described in this section. Note that each paper description follows the guidlines proveded by Laramee ([Lar11]).

Table 8: ***Overview of EHR Vis techniques:*** *Ordered by the publication year. The x-axis is mapped to the re-occurring research themes we extracted from the literature.* 🟥 *Red highlights the primary theme,* ⬛ *darkGrey highlights the secondary theme, and* 🟩 *Green highlights context papers.*

	Complementary Techniques							
Literature	ML	NLP	ESS	GEO	Clustering	Comparison	Others	Year
🟩 Plaisant et al. [PMS*98b]							🟥	1998
🟩 Horn et al. [HPU01]							🟥	2001
🟩 Bade et al. [BSM04]							🟥	2004
🟩 Goren-Bar et al. [GSG*04]							🟥	2004
🟩 Hinum et al. [HMA*05]							🟥	2005
🟩 Fails et al. [FKSS06]			🟥					2006
🟩 Bui et al. [BAK07]		🟥						2007
🟩 Wang et al. [WPS*09]			🟥					2009
🟩 Rind et al. [RMA*10]							🟥	2010
🟩 Faiola and Newlon [FN11]							🟥	2011
Gotz et al. [GSCE11]					🟥			2011
Gschwandtner et al. [GAK*11]						🟥		2011
Wongsuphasawat et al. [WGP*11]			🟥					2011
Zhang et al. [ZAR*11]		🟥						2011
Alonso and McCormick [AM12]				🟥				2012
Sopan et al. [SNK*12]				🟥				2012
Wongsuphasawat and Gotz [WG12]			🟥					2012
Monroe et al. [MLL*13]			🟥					2013
Ramírez-Ramírez et al. [RGT*13]				🟥				2013
Borland et al. [BWH14]						🟥		2014
Gotz and Stavropoulos [GS14]			🟥					2014
Kamaleswaran et al. [KPT*14]					🟥			2014
Malik et al. [MDM*14a]			🟥					2014
Bernard et al. [BSB*15]	🟥							2015
Bernard et al. [BSM*15]						🟥		2015
Federico et al. [FUS*15]						🟥		2015
Klemm et al. [KLG*15]				🟥				2015
Glueck et al. [GHC*16]						🟥		2016
Jiang et al. [JFB*16]		⬛		🟥				2016
Kamaleswaran et al. [KCJM16]					🟥			2016
Loorak et al. [LPK*16]			🟥			⬛		2016
Ola and Sedig [OS16]				🟥				2016
Dabek et al. [DJC17]	🟥							2017
Glueck et al. [GGC*17]		⬛				🟥		2017
Tong et al. [TML*17]				🟥				2017
Tong et al. [TRL*17]				🟥				2017
Glueck et al. [GND*18]	🟥					⬛		2018
Guo et al. [GXZ*18]	⬛	⬛	🟥					2018
Tong et al. [TML18]				🟥				2018
Trivedi et al. [TPC*18]	⬛	🟥						2018
Alemzadeh et al. [ANI*19]				🟥				2019
Bernard et al. [BSKR19]			🟥					2019
Glicksberg et al. [GOT*19]							🟥	2019
Guo et al. [GJG*19]	⬛		🟥					2019
Kwon et al. [KCK*19]	🟥							2019
McNabb and Laramee [ML19]				🟥				2019
Sultanum et al. [SSBC19]	⬛	🟥						2019
Zhang et al. [ZCD19]						🟥		2019
Jin et al. [JCG*20]	🟥		⬛					2020
Kwon et al. [KAS*20]	🟥							2020
Wang et al. [WLLP21]		⬛				🟥		2021
Total unique paper: 51	10	8	12	11	3	10	8	

4.1. Machine learning

This section introduces the literature that combines Machine Learning (ML) and EHR Vis. We follow the definition of ML by Alpaydin [Alp10] as the process of optimizing the performance of a predefined model, based on example data or past experience. The outcomes from the process are either predictive to provide guidance on the future or descriptive to acquire knowledge from the existing data. The application of ML techniques such as deep learning [ROC*18], neural networks [KCK*19], support vector machines [ZXG19, BRMF19] and topic models [BRMF19], have evolved recently to increase automation of processing EHR archives. From examining Table 8, we can observe that incorporating ML into EHR Vis is a relatively new trend and not very mature. Also, we believe that EHR Vis could benefit more with the help of ML techniques. Table 9 presents an overview of the EHR literature in this subsection indicating which ML techniques are used. We can observe that active learning is a recurring theme in the visualization literature.



UMLS Code	UMLS Term
C0600139	Prostate carcinoma
C0679831	Patient history

Bernard et al. contribute a visual active learning system [BSM*15] extending their prior work [BSB*15]. The system en-

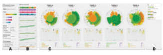
Table 9: *An overview table of ML topics discussed in the literature described in Section 4.1. Papers with ML as a secondary theme are highlighted in Green.*

Literature	ML Topics	UMLS Term	Year
Bernard et al. [BSB*15]	Active Learning REPTree	Prostate carcinoma	2015
Dabek et al. [DJC17]	Unspecified	Patient history	2017
Guo et al. [GXZ*18]	Clustering	Chronic obstructive airway disease	2019
Glueck et al. [GNDV*18]	Active Learning	Phenotype Disease subtype	2018
Trivedi et al. [TPC*18]	Support Vector Machine Bag-of-words	Colonoscopy	2018
Guo et al. [GJG*19]	Neural Networks	Care of intensive care unit patient	2019
Kwon et al. [KCK*19]	Recurrent Neural Networks	Diabetes mellitus, insulin-dependent Hypercholesterolemia Hypertensive disease Medical history	2019
Sultanum et al. [SSBC19]	Active Learning	Care of intensive care unit patient Medical history	2019
Jin et al. [JCG*20]	Recurrent Neural Networks	Cardiovascular diseases Respiratory tract diseases Treatment plan	2020
Kwon et al. [KAS*20]	Hidden Markov Models	Diabetes mellitus, insulin-dependent Huntington disease Parkinson disease	2020

ables physicians to evaluate the well-being status of prostate cancer patients by exploiting the patient's history as recorded in their respective EHRs. The phrase visual active learning system refers to a system that uses an active learning approach which requires physicians for feedback and corrections during the training of the model. The resulting visualization enables quick identification of possible diagnoses of individual patient's symptoms.

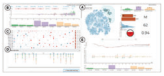
UMLS Code	UMLS Term
C0679831	Patient history

Dabek et al. propose a timeline-based framework for aggregating and summarizing EHRs [DJC17]. The main challenge they address is the heterogeneous nature of EHR data sources. The framework implements a patient timeline that conveys temporal events with nodes. Each node contains a textual summary generated automatically via machine learning. A separate panel is presented with a sunburst chart visualizing patient diagnoses and a horizon chart visualizing lab test results.



UMLS Code	UMLS Term
C0031437	Phenotype
C0872379	Disease subtype

Glueck et al. present PhenoLines, a visual analysis tool for the interpretation of disease subtypes that exploits the application of topic modelling applied to clinical data [GNDV*18]. Based on the Human Phenotype Ontology (HPO) extracting and mapping method introduced in the prior work [GHC*16, GGC*17], PhenoLines aims to support the filtering, comparison, simplification and interpretation of temporal evolution of phenotype probabilities within and between disease subtypes. Topic modeling is used to mine cross-sectional patient's comorbidity data from high dimensional EHRs. PhenoLines enables interactive analysis of the derived topic models, by encoding them in sunburst charts, as shown in Figure 2.



UMLS Code	UMLS Term
C0011854	Diabetes mellitus, insulin-dependent
C0020443	Hypercholesterolemia
C0020538	Hypertensive disease
C0262926	Medical history

Kwon et al. [KCK*19] first present RetainEx, a recurrent neural networks (RNN) approach that develops interactivity and interpretability for prediction tasks and incorporates the temporal dimension in patient history data. As the RNN uses a black-box approach, it is difficult to couple the predictions to a particular attribute used during training. The authors then introduce RetainVis, an interactive visual analytics tool for assisting the user in understanding the process of prediction. Histogram, scatterplot, matrix and glyph designs are used to present influential attributes leading to the prediction.

UMLS Code	UMLS Term
C0007222	Cardiovascular diseases
C0035242	Respiratory tract diseases
C0599880	Treatment plan

Jin et al. [JCG*20] introduce CarePre, an intelligent system that converts EHR data from a cohort of patients into sequences of events, and leverages machine learning techniques for the prediction of a patient's risk level during diagnosis. The system then recommends the most influential treatment plans. Based on the available

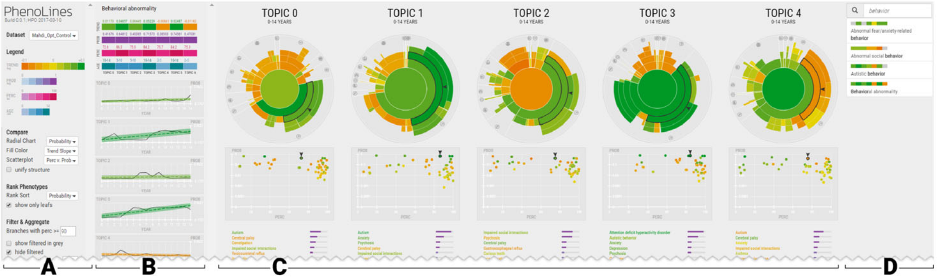
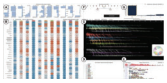


Figure 2: *PhenoLines [GNDV*18] includes (A) A settings panel for interactive functions such as sort, filter and aggregate, (B) A detail panel renders the phenotype in the selected topic with juxtaposed timeline charts, (C) The topics panel provides an overview of all topics extracted, and (D) A search panel. Image courtesy of Glueck et al. [GNDV*18].*

Table 10: *An overview table of NLP approaches adopted by the literature described in Section 4.2. Papers with NLP as a secondary theme are highlighted in Green.*

Literature	NLP Approaches	UMLS Term	Year
Zhang et al. [ZAR*11]	Unspecified	Medical history	2011
Jiang et al. [JFB*16]	Named Entity Recognition	Medical history	2016
Glueck et al. [GGC*17]	Natural Language Queries	Phenotype	2017
Trivedi et al. [TPC*18]	Automated Retrieval Console cTAKES	Colonoscopy	2018
Sultanum et al. [SSBC19]	cTAKES	Care of intensive care unit patient Medical history	2019
Wang et al. [JFB*16]	Natural Language Queries	Epilepsy	2021

EHR data, CarePre is also able to predict the likelihood of an outbreak for a set of potential diseases selected by the user. The MIMIC dataset [JPS*16] is used for thorough evaluations with seven physicians including two case studies. We include this open access dataset in Table 17.



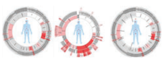
UMLS Code	UMLS Term
C0011854	Diabetes mellitus, insulin-dependent
C0020179	Huntington disease
C0030567	Parkinson disease

Kwon et al. [KAS*20] present DPVis, a multiple views visual analytics system that focuses on visual disease progression analysis in order to develop fully interpretable and interactive visualizations. Hidden Markov models (HMMs) are trained to infer the most probable state sequences based on the user-chosen attributes. DPVis incorporates multiple interactive visual designs including matrix, chord diagram and parallel beeswarm plots to support the exploration of disease progression and discover associations between patterns and variables.

4.2. Natural language processing

This section introduces EHR Vis papers incorporating Natural Language Processing (NLP) as a complementary technique. We follow the definition of NLP from Liddy as, *"a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications"* [Lid01]. As an active area of research, NLP has evolved since its inception in the 1940s.

As one of the most widely used analytical techniques in healthcare, NLP is capable of transforming unstructured text into a structured and machine-readable format [KDBB19]. Clinicians have very diverse ways of documenting patient records. This may require appropriate modifiers to capture words, phrases and their relationships in EHRs [SSBC19]. Table 10 shows a summary of the NLP techniques used in the EHR Vis literature. It is evident that incorporating NLP techniques is still in its early stages and has much room to grow.



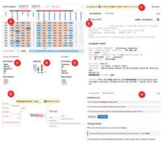
UMLS Code	UMLS Term
C0262926	Medical history

Zhang et al. develop AnamneVis in order to capture a complete picture of a patient's medical history [ZAR*11]. AnamneVis incorporates NLP algorithms to extract structured medical information from unstructured data sources such as doctor-patient dialogs and medical reports. The International Classification of Diseases (ICD) as the medical standard for mapping diseases and symptoms. The Five Ws concept [ZBA*13] is adopted for mapping the relations between extracted information. A sunburst diagram is used to visualize the data in two layouts, (1) a hierarchy-centric layout for the

Table 11: *An overview table of event types in the literature described in Section 4.3. Papers with ESS as a secondary theme are highlighted in Green.*

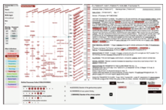
Literature	Event Types	UMLS Term	Year
Wongsuphasawat et al. [WGP*11]	Hospital discharge and transfer flows	Patient transfer	2011
Wongsuphasawat and Gotz [WG12]	Congestive heart failure	Congestive heart failure	2012
Monroe et al. [MLL*13]	Prescriptions	Pharmacology	2013
Gotz and Stavropoulos [GS14]	Diagnoses, lab tests, and medications	Medical history	2014
Malik et al. [MDM*15]	Respiratory and radiation		2015
Loorak et al. [LPK*16]	Stroke	Cerebrovascular accident	2016
Guo et al. [GXZ*18]	Diagnoses, procedures, hospital admission and discharge)	Chronic obstructive airway disease	2018
Bernard et al. [BSKR19]	Biological indicator for prostate cancer	Prostate carcinoma	2019
Guo et al. [GJG*19]	Hospital admission and discharge, death, prescriptions, infusions, lab tests	Care of intensive care unit patient	2019
Jin et al. [JCG*20]	Hospital admission, prescriptions, diagnoses, treatments	Cardiovascular diseases Respiratory tract diseases Treatment plan	2020

hierarchy information representing diagnosed ICD codes, and (2) a patient-centric layout for the past diagnoses and procedures taken. In addition, a sankey diagram is used to illustrate past medical diagnostic flow of the patient.



UMLS Code	UMLS Term
C0009378	Colonoscopy

Trivedi et al. introduce NLPReViz, a visual analytic and visualization tool that uses Support Vector Machine for training NLP models in real time [TPC*18]. Users are able to train, review and revise trained NLP models by rectifying the binary results from the previous execution. Re-trained models are used for next execution and provide a more accurate result. We classify NLPReVis as NLP, since it uses a combination of NLP and ML, but with more of a focus on NLP, this is reflected in Table 8.

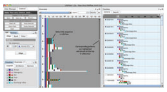


UMLS Code	UMLS Term
C0010337	Care of intensive care unit patient
C0262926	Medical history

Sultanum et al. present Doccurate, a system embodying a curation-based approach that automatically extracts relevant information from large clinical text datasets, to provide an accurate and sufficient overview for a patient [SSBC19]. After interviewing six domain experts, the authors conclude that preserving the original text in clinical notes is crucial for the visualization of EHRs. Doccurate provides automation in data processing and customization for visualization while preserving the link to the original data.

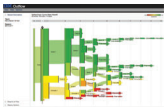
4.3. Event sequence simplification

This section includes EHR Vis literature with a focus on Event Sequence Simplification (ESS). We follow the definition of ESS as any technique used for reducing the visual complexity of event sequences in aggregated display overviews [MLL*13, MSD*16]. EHRs by nature are temporal events unfolding successively, ESS enables events to be trimmed down to their core elements, improving both data-processing and visualization of EHRs. The technique is adopted by EHR Vis systems such as LifeLines [PMR*96] and EventFlow [MLL*13]. Table 11 provides a summary of event types appearing in this sub-section. Events associated with hospitals are a recurring theme.



UMLS Code	UMLS Term
C0030704	Patient transfer

Wongsuphasawat et al. [WGP*11] introduce LifeFlow for providing an interactive visual overview of event sequence data. Following the approach used in LifeLines2 [WPS*09], the authors introduce an aggregation method that groups events into a tree-based hierarchical data structure. Nodes of the same type are rendered as a color-coded event bar, the height of an event bar is proportional to the number of records, and the gap between event bars represents the average time between events. Although the case study of LifeFlow focuses on the analysis of patient transfers between hospital departments, we still include the paper for its aggregation method. We believe the technique is also applicable to EHRs.



UMLS Code	UMLS Term
C0018802	Congestive heart failure

Wongsuphasawat et al. [WG12] introduce Outflow, a visualization of temporal event sequence data. Outflow uses a different approach that visualizes the aggregation results using a graph-based representation, which simplifies the comparison of alternative paths with the same state. Both papers are supported by user studies.

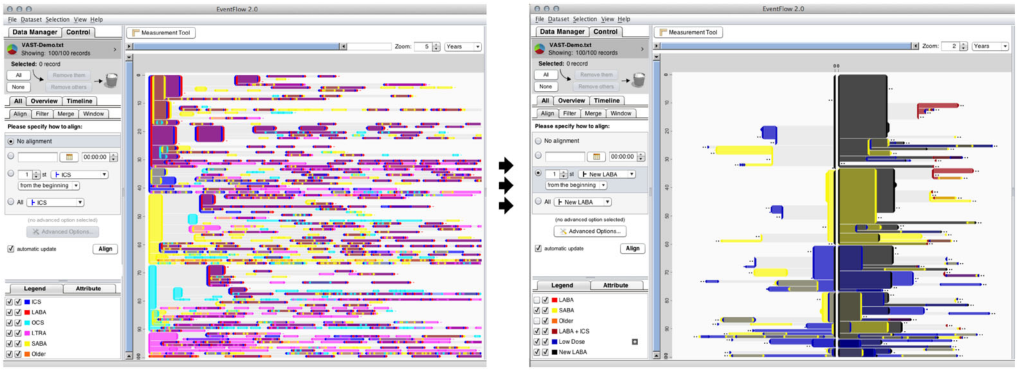
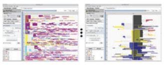
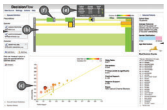


Figure 3: *EventFlow [MLL*13] visualizing the original Long-Acting β-Agonists dataset on the left, and the simplified dataset on the right. The number of visual elements is reduced from 2,700 to 492. Image courtesy of [MLL*13].*



UMLS Code	UMLS Term
C0031330	Pharmacology

Monroe et al. introduce a technique to simplify temporal event sequence data [MLL*13], following their previous work called EventFlow [MWP*12]. EventFlow transforms temporal events into an aggregated display to identify hidden trends in the data, this is particularly useful for EHRs as the scalability and the dimensionality of EHRs grow, the visual complexity also increases. An example is shown in Figure 3. The authors propose user-driven simplification, achieved via filtering-based selection: (1) filtering by record which allows the user to remove records through querying or clicking, (2) filtering by category which hides the selected categories and aggregate visual elements into fewer and larger displays, (3) filtering by time enables the user to define a time frame to reduce visual density, (4) filtering by attributes which enables the user to define threshold values. However, filtering-based simplification removes events from the original data. Transformation-based simplification is introduced to preserve the logical relations between events: (1) interval event merging is used to remove gaps or overlap between events, (2) category merging enables categories to be combined to reduce visual elements without removing events, (3) marker event insertion allowing the user to collapse multiple events into a single one.

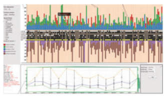


UMLS Code	UMLS Term
C0262926	Medical history

Gotz and Stavropoulos [GS14] introduce DecisionFlow for visualizing large numbers (thousands) of high-dimensional temporal event sequence data. Instead of visualizing the entire dataset from the beginning, DecisionFlow allows the user to construct a query with multiple constraints to retrieve the initial data. The result is then aggregated to generate milestones, and visualized for further analysis and interactions. The user is able to set and modify milestones interactively to achieve filtering and selection.

UMLS Code	UMLS Term
C0010337	Care of intensive care unit patient
C0030677	Patient care management

Malik et al. present CoCo [MDM*14a], a visual analytics tool for comparing groups (cohorts) of temporal event sequence data. Inspired by EventFlow [MLL*13] and Outflow [WG12], CoCo enables users to explore statistics about the underlying dataset as they interact with the simplified temporal event sequences. CoCo offers a combination of user-driven and automated methods to enable comparisons of cohort events. The authors evaluate the work [MDM*15] with two case studies.



UMLS Code	UMLS Term
C0038454	Cerebrovascular accident

Loorak et al. [LPK*16] present TimeSpan, a visualization tool designed to explore the temporal aspects of the stroke treatment process. The authors collaborate with a team of domain experts to derive and classify a list of basic tasks in the domian of stroke care analysis. Temporal events are visualized using a parallel coordinates with stacked bar charts extended with the Bertin-style matrices [Ber99], and are aligned based on their positive effect on the patient. A unique evaluation with a focus group session is also presented.

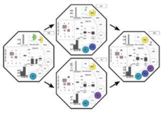
UMLS Code	UMLS Term
C0024117	Chronic obstructive airway disease

Guo et al. describe EventThread [GXZ*18], a visualization system for revealing the evolution of patterns across stages in event sequence data. EventThread uses Term Frequency - Inverse Document Frequency (TF-IDF) [RJ76], a common technique used to measure the importance of text segments in a document, to capture the primary sequential pattern in the data. Events are then grouped into threads by similarity, with interactive options provided to

Table 12: *An overview table of comparative designs adopted by the literature described in Section 4.4. The x-axis is mapped to the comparative design categorization by Gleicher et al. [GAW*11]. Papers with Comparison as a secondary theme are highlighted in Green.*

	Comparative Design				
Literature	Juxtaposition	Superposition	Explicit Encoding	UMLS Term	Year
Gschwandtner et al. [GAK*11]	🟥			Clinical action Treatment plan	2011
Bborland et al. [BWH14]	🟥	🟥	🟥	Carcinoma of lung	2015
Bernard et al. [BSM*15]	🟥			Prostate carcinoma Patient History	2015
Federico et al. [FUS*15]	🟥			Atrial fibrillation Gestational diabetes Clinical history	2015
Glueck et al. [GHC*16]	🟥			Phenotype	2016
Loorak et al. [LPK*16]		🟥		Cerebrovascular accident	2016
Glueck et al. [GGC*17]	🟥			Phenotype	2017
Glueck et al. [GNDV*18]	🟥	🟥		Phenotype	2018
Glicksberg et al. [GOT*19]	🟥			Clinical history	2019
Zhang et al. [ZCD19]	🟥		🟥	Diabetes mellitus, insulin-dependent Treatment plan	2019
Wang et al. [WLLP21]	🟥	🟥		Epilepsy	2021

facilitate further analysis. Guo et al. introduce EventThread 2 [GJG*19] to improve the system's ability to handle the temporal dimension by adopting neural network models. Both work involve collaborations with medical experts and case studies with EHR data.

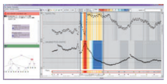


UMLS Code	UMLS Term
C0600139	Prostate carcinoma

Bernard et al. propose a technique for visualizing post-operative prostate cancer, that segments patient histories based on time and then aggregates the results by therapy states and biological conditions [BSKR19]. Instead of treating patient histories as event sequences, the segmented results are presented using a static dashboard, with extensive use of colors and glyphs for encoding variables, in order to visualize longitudinal changes in patient histories. The segmentation of patient histories is done by using a sliding window approach to traverse through the dataset. Evaluation is performed with groups of both expert and non-expert users.

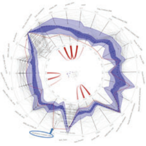
4.4. Visual analytics and comparison

This section describes research on visual analytics combined with analytical comparison of EHRs. We follow the three categories of comparative visual designs by Gleicher et al. [GAW*11], juxtaposition, superposition and explicit encodings. Table 12 summarizes the types of comparison techniques used in the EHR Vis literature. Juxtaposition, the simplest, is the most common choice by a wide margin.



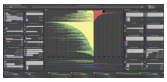
UMLS Code	UMLS Term
C0441472	Clinical action
C0599880	Treatment plan

Gschwandtner et al. present CareCruiser [GAK*11], an enhanced visual analysis system to explore the result of each applied clinical action and identifies sub-optimal treatment choices. CareCruiser supports the visualization of (1) hierarchical data which includes the structure of treatment plans and sub-plans, (2) temporal data referring to the execution sequence of treatment plans and sub-plans, and the patient's condition over time, (3) qualitative data which represents relevant characteristics of treatment plans and sub-plans. Aligning, filtering and focus+context are provided for investigation of the patient's condition and responses to treatments, as well as comparison between multiple patients.



UMLS Code	UMLS Term
C0684249	Carcinoma of lung

Borland et al. [BWH14] describe radial coordinates, a visualization technique based on parallel coordinates, a scatterplot and a chord diagram. The technique allows a more efficient utilization of the space by representing each variable using an axis, arranged radially around a scatterplot. Chords are used to represent relationships between variables. The design supports comparison of high and low prevalance values across all dimensions in the data. The radial style parallel coordinates visual design is applied to NHS data from the UK.



UMLS Code	UMLS Term
C0600139	Prostate carcinoma
C0679831	Patient History

Bernard et al. present an interactive visualization system for identifying, categorizing, and analysing EHRs of cohorts of prostate cancer patients [BSM*15]. The system supports the visualization

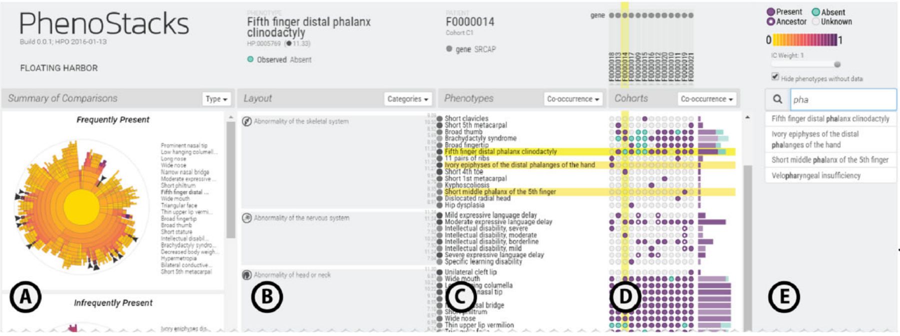
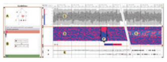

PhenoStacks
Build 0.0.1; HPO 2016-01-13
FLOATING HARBOR
Fifth finger distal phalanx clinodactyly
Observed Absent
F0000014
Cohort C1
gene SRCAP
gene
Present
Absent
Ancestor
Unknown
0
1
Hide phenotypes without data
Summary of Comparisons
Type
Frequently Present
Infrequently Present
Layout
Categories
Abnormality of the skeletal system
Abnormality of the nervous system
Abnormality of head or neck
Phenotypes
Co-occurrence
Cohorts
Co-occurrence
pha
Fifth finger distal phalanx clinodactyly
Ivory epiphyses of the distal phalanges of the hand
Short middle phalanx of the 5th finger
Velopharyngeal insufficiency
Short clavicles
Short 5th metacarpal
Broad thumb
Brachydactyly syndrome
Broad fingertip
Fifth finger distal phalanx clinodactyly
11 pairs of ribs
Ivory epiphyses of the distal phalanges of the hand
Short 4th toe
Short 1st metacarpal
Kyphoscoliosis
Short middle phalanx of the 5th finger
Dislocated radial head
Hip dysplasia
Mild expressive language delay
Moderate expressive language delay
Intellectual disability, severe
Intellectual disability, moderate
Intellectual disability, borderline
Intellectual disability, mild
Severe expressive language delay
Specific learning disability
Unilateral cleft lip
Wide mouth
Wide nose
Thin upper lip vermilion
A
B
C
D
E

Figure 4: *PhenoStacks [GGC*17] includes (A) The summary panel conveying phenotype patterns across patient cohorts in a sunburst chart, (B) The layout view enables the user to select phenotypes by collapsing, filtering and clustering, (C) The list view shows the phenotype names with a sorting function, (D) The observations Plot visualizes the actual and inferred phenotype observations in a matrix, and enables the user to explore and identify potential patterns, and (E) The search panel supports natural language queries for searching phenotypes. Image courtesy of [GGC*17].*

of multiple patients with (1) an overview that supports direct selection of patients, (2) dynamic queries against attributes to achieve filtering, and (3) a history panel that stores previous cohorts that can be retrieved easily for comparison. The system also offers a guided analysis of correlations between patients in the cohort.



UMLS Code	UMLS Term
C0004238	Atrial fibrillation
C0085207	Gestational diabetes
C5204342	Clinical history

Federico et al. introduce Gnaeus, a guideline-based knowledge-assisted visual analytics system for EHRs [FUS*15]. Gnaeus utilizes computer-interpretable clinical guidelines (CIGs), which are generated based on evidence-based clinical practice guidelines, to assist the analysis of EHR data. Selected parameters from the raw data are placed in parallel with clinical actions executed to visualize the outcome, with related CIGs on the side to provide recommendations. The system enables the user to compare administered treatment with evidence-based best practices.

UMLS Code	UMLS Term
C0031437	Phenotype

Glueck et al. introduce PhenoBlocks, a visual analytics tool that supports the comparison of phenotypes between patients [GHC*16]. PhenoBlocks introduces a differential hierarchy comparison algorithm for analysing phenotypes pairwise between patients, and uses a customized sunburst radial hierarchy layout [SZ00] for visualizing the results.

UMLS Code	UMLS Term
C0031437	Phenotype

Glueck et al. present PhenoStacks, a visualization system to support comparison of cross-sectional phenotype within and between patient cohorts [GGC*17]. The system adopts glyphs of Human Phenotype Ontology (HPO) developed in the prior work PhenoBlocks [GHC*16] and supports sorting and filtering by phenotype or patient attributes. Search is powered with natural language queries (See Figure 4). To reduce visual redundancy, the authors propose a topology simplification algorithm, a greedy depth-first approach, for eliminating duplicates in phenotype datasets.

UMLS Code	UMLS Term
C0011854	Diabetes mellitus, insulin-dependent
C0599880	Treatment plan

Zhang et al. describe IDMVis [ZCD19], a temporal event sequence visualization system developed for Type 1 diabetes treatment decision support. They provide a new method of hierarchical task abstraction for clinicians. Inspired by Temporal Folding, a technique for visualizing temporal event sequences [DSP*17], the authors propose a visual technique of dual sentinel event alignment and time scaling to further enhance the visualization for a large number of temporal event sequences. In addition to the single-event alignment that enables the alignment of trend lines based on a single designated event, the technique enables the alignment of trend lines between two user-chosen events with zooming.

UMLS Code	UMLS Term
C0014544	Epilepsy

Wang et al. present LetterVis, a visualization tool to support the analysis of clinic letters through five customized visual layouts with support from natural language queries [WLLP21]. A letter-space layout is derived from the physical layout of text on A4-size letters used by clinicians, exploiting implicit knowledge of the clinicians who compose the letters. This layout is used to depict the query results in (1) the global view that shows all the letters loaded in one superimposed letter-space, (2) a thumbnail view for individual letters, and (3) a focus view for the original content with query results highlighted. (4) A co-occurrence matrix is

Table 13: *An overview table of clustering dimensions used in the literature described in Section 4.5.*

Literature	Clustering Dimensions	UMLS Term	Year
Gotz et al. [GSCE11]	Medical decisions	Diabetes	2011
Kamaleswaran et al. [KPT*14]	Temporal	Neonatal intensive care	2014
Kamaleswaran et al. [KCJM16]	Temporal Respiratory physiologic signals	Neonatal intensive care	2016

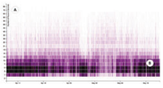
included for visualizing antiepileptic drug (AED) co-prescriptions. In the (5) drug chain view, where each AED is represented by a block in the chain, provides a visual representation of prescription progression.

4.5. Visual analytics with clustering and others

This section describes papers that use hierarchical clustering algorithms for EHR analytics. According to the survey by Xu and Wunschll [XW05], hierarchical clustering algorithms are widely used in the information visualization discipline. This conforms with our findings that all papers included in this section (Table 13) adopt hierarchical clustering algorithms to produce homogeneous subgroups based on similarities. EHRVis may benefit from applying other clustering techniques (e.g. Vector Quantization and Estimation via Mixture Densities) to assist in analysis.

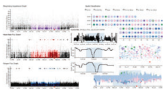
UMLS Code	UMLS Term
C0011847	Diabetes

Gotz et al. introduce DICON [GSCE11], a visualization tool that supports the exploration of similarity in cohorts of patients. Clusters are represented by dynamic icons and are generated using similarity and cluster analysis algorithms. The cluster refinement stage requires user guidance to evaluate cluster quality and apply refinements. Users can drag and drop, merge and split an individual patient or a cluster to refine clustering results.



UMLS Code	UMLS Term
C0021711	Neonatal intensive care

Kamaleswaran et al. uses a tri-event heatmap representation for displaying high frequency complex data [KPT*14], neonatal spells, collected in neonatal intensive care units. Their clustering includes a temporal factor and a non-linear similarity metric. The authors apply both density estimation and logarithmic clustering to normalize and discretize the non-parametric distribution during data pre-processing. The resulting visualization supports the exploration of frequency, duration, and severity of spells.



UMLS Code	UMLS Term
C0021711	Neonatal intensive care

Kamaleswaran et al. introduce a visualization technique called a Temporal Intensity Map (TIM) [KCJM16], a customized heatmap with the y-axis representing the critical distance interval determined by a density estimation function. The focus is on the visual analysis of event streams that reveal important infomation about frequency and duration of streaming events derived from real-time event stream algorithms. The authors further introduce a dashboard visual analysis system, PhysioEx, formed by a TIM, a sequence graph, a linear graph, and a streams graph for analysing neonatal data and predicting physiological behaviours of newborns.

Query by any CDM vocabulary concept
Export all data and images
Cohort Visualization and Filtering
Select Patient
Patient Clinical Report
Patient Encounter Timeline
Data Explorer

UMLS Code	UMLS Term
C5204342	Clinical history

Glicksberg et al. describe PatientExploreR [GOT*19], an interactive interface that facilitates the visualization and querying of EHRs. By incorporating the Observational Medical Outcomes Partnership (OMOP) common data model introduced by the Observational Health Data Sciences and Informatics [Obs20], PatientExploreR's advanced querying function allows physicians to search, filter and compare patients with combinations of items from multiple medical terminology standards such as the UMLS described in Section 2.2. When a patient is selected, an interactive timeline presents all clinical events with the ability to expand the details, along with basic visual designs. We include this paper for the advanced querying support coupled with the integration of OMOP common data model.

4.6. PopHR vis and Geospatial visualization

This section describes research on EHR Vis with a geospatial focus. Table 14 summarizes the geospatial landscape coupled by this sub-section of literature. PopHR Vis papers are also included in this section.

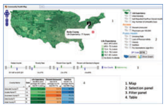
UMLS Code	UMLS Term
C0032285	Pneumonia
C0021400	Influenza

Alonso and McCormick describe Epidemiological Parameter Investigation from Population Observations Interface (EPIPOI), that automatically extracts three parameters describing trends, seasonality and anomalies, and a time series from large epidemiological datasets [AM12]. These three dimensions can be visualized using maps combined with time series data to reveal spatial patterns.

Table 14: *An overview table of geospatial regions covered in the literature described in Section 4.6.*

Literature	Geospatial Regions	UMLS Term	Year
Alonso and mccormick [AM12]	Brazil	Pneumonia Influenza	2012
Sopan et al. [SNK*12]	US	Population health	2012
Ramírez-ramírez et al. [RRGT*13]	Ontario, Canada	Influenza	2013
Klemm et al. [KLG*15]	Germany	Population health Breast Density Steatohepatitis	2013
Jiang et al. [JFB*16]	Indiana, US	Medical history	2016
Ola and Sedig [OS16]	Global	Population health	2016
Tong et al. [TML*17]	England, UK	Population health	2017
Tong et al. [TRL*17]	England, UK	Population health	2017
Tong et al. [TML18]	England, UK	Population health	2018
Alemzadeh et al. [ANI*19]	Germany	Longitudinal studies	2019
Mcnabb and Laramee [ML19]	Ireland UK US	Population health	2019

EPIPOI additionally supports wavelet analysis to reveal sinusoidal patterns of a time series with different frequencies, and Fourier Series to identify biologically relevant descriptors of seasonality.



UMLS Code	UMLS Term
C3242284	Population health

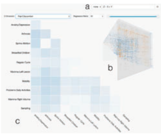
Sopan et al. introduce the Community Health Map [SNK*12] that interactively visualizes public healthcare datasets using a multivariate choropleth map. Selection enables users to visualize multiple datasets gathered from Hospital Referral Regions and administrative counties in the U.S. Filtering of income, poverty rate, age and education level are supported to enable the comparison of different socioeconomic classes.

UMLS Code	UMLS Term
C0021400	Influenza

Ramírez-ramírez et al. introduce SIMID [RRGT*13], a surveillance and spatio-temporal visualization tool for infectious diseases. Based on the existing data, SIMID simulates the spread of infectious disease using interactive animated maps. With customizable input parameters such as vaccination rate and mortality rate, SIMID is able to generate different mitigation strategies with variation and uncertainty that reflect the randomness in disease outbreak progression.

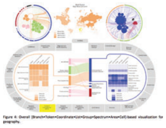
UMLS Code	UMLS Term
C0262926	Medical history

Jiang et al. introduce Health-Terrain [JFB*16] to support the visual exploration of large healthcare datasets. Based on UMLS described in Section 2.2, the authors extract related terms from unstructured clinical notes via NLP. The authors propose a spatial texture based approach to integrate geospace with other dimensions, that consists of (1) constructing random noise patterns with color variations to map different attributes, and (2) color-coding the offset contours of geographical regions to map the temporal dimension. The authors propose a visual design called a Spiral Theme Plot based on ThemeRiver [HHN00] and spiral pattern [WAM01], to help physicians discover patterns and trends in events. Health-Terrain is included in this section, since it is a combination of geospatial visualization and NLP with the main focus on the former.



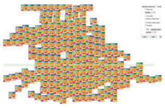
UMLS Code	UMLS Term
C3242284	Population health
C1659543	Breast Density
C2711227	Steatohepatitis

Klemm et al. [KLG*15] propose the 3D Regression Heat Map, a novel 3D visual encoding that offers an overview of a hepatic steatosis dataset (a subset of the SHIP dataset included in Table 17). The resulting 3D heat map enables the exploration of relationships between several user-defined independent features and a user-defined target disease. Each 3D heat map slice can be projected onto a 2D space for further analysis. The approach enables experts to verify their disease-specific hypotheses and derive new ones.



UMLS Code	UMLS Term
C3242284	Population health

Ola and Sedig [OS16] present a geospatial visual design for studying large healthcare datasets. The design combines several visualization techniques to support the exploration of the relationships between age group, risk, and cause of death at multiple levels of granularity.

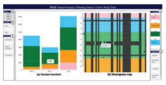


UMLS Code	UMLS Term
C3242284	Population health

Tong et al. present a hybrid visual layout called a cartographic treemap, to visualize high-dimensional healthcare data collected by the National Healthcare Service (NHS) in the U.K. [TML*17]. By combining the space-filling advantages of treemaps for the display of hierarchical, multivariate data together with geospatial information, cartographic treemaps support exploration, analysis and comparison of complex population healthcare data from Public Health England. They further extend the work it by adding a time variate, enabling the visualization of the temporal evolution trends hidden in EHR data [TRL*17].

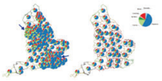
UMLS Code	UMLS Term
C3242284	Population health

Tong et al. extend their previous work with a cartographic layout algorithm that generates cartograms with topological features using NHS's population healthcare data [TML18]. The proposed algorithm preserve nearby node's topological features to increase the recognizability and reduce layout errors.



UMLS Code	UMLS Term
C0023981	Longitudinal Studies

VIVID is a web-based framework proposed by Alemzadeh et al. [ANI*19] to support the handling of the missing values in cohort study data. The framework includes various visual designs to enable the user to explore the missing values (stacked barchart and matrix) build imputation models (bean plot and bee swarm plot) and generate predictions for the missing values (chord diagram and parallel coordinates).



UMLS Code	UMLS Term
C3242284	Population health

McNabb and Laramee present a glyph placement algorithm to support multivariate geospatial visualization of a Public Health England dataset [ML19]. The authors identify four major challenges for representing geospatial data on existing choropleths: (1) Size perceivability: sizes of glyphs and areas on a map are not easily perceivable. (2) Visualization of multivariate geospatial data: geospatial designs such as choropleths, cartograms, symbol maps etc. generally fail to depict multivariate data. (3) Occlusion: glyphs on a map often overlap and are over-plotted. (4) Glyph placement: existing solutions to address occlusion often de-couple glyphs from their original geospatial regions they are intended to represent. The authors address these challenges by introducing a scale-aware map that supports dynamic modification to the level-of-detail shown via zooming and custom scaling options. The algorithm produces a map that is enhanced with glyphs which are dynamic, scale-aware and coupled to their geospatial contexts.

5. Evaluation

Evaluation of EHR and PopHR visual designs is very difficult due to their complex visual interfaces. An EHR Vis system is often characterized according to target user requirements. The resulting visual designs may not seem useful to evaluators [Mun09]. Furthermore, an isolated evaluative process is hardly sufficient to assess an EHR Vis system. *Grounded evaluation* [IZCC08], where visualization designers work closely with EHR experts to (1) understand predesign context, (2) conduct iterative prototyping and refinement, and 3) conduct late-stage acceptance testing, might be a solution to address the evaluation problem. We observe that grounded evaluation is being practiced in many projects (especially from the visualization community) included in this STAR.

In this section, we summarize the evaluation techniques adopted by each paper. The result is summarized in Table 15.

1. **Domain expert feedback (30 papers, 59%)** is the most popular evaluation technique used. This approach aims to understand the current health-related work practice or assess the value of the newly developed tool [IIJ*13]. While the technique is commonly used in the reviewed literature, we observe a trend of increasing involvement of domain experts since the year 2018, where domain experts participate in multiple stages of the software development life cycle, such as planning, requirement analysis and testing [TPC*18, BSKR19, SSBC19, JCG*20]. The close involvement informs the development process and enables rapid feature development and innovation.
2. **Interview (26 papers, 51%):** Where a set of guided questions along with open-ended questions are provided and answered usually in person, including both expert and novice users. Interviews can be performed multiple times throughout the software development life cycle. We observe an increase in adoption of interviews since 2014. Interviewees usually respond in depth during interviewing sessions [GHC*16, KCJM16, GGC*17, TPC*18, BSKR19, JCG*20] and provide personal interpretations beyond interaction and usability aspects [HHH16].
3. **Case study (17 papers, 33%):** A case study often provides the most in-depth evaluation result, as the participants are usually placed in a real-world situation after the provided training [KCK*19]. This enables the target audience to generate in-depth feedback based on their experience. We notice a long time period from about 2011-2018, where case studies do not generally appear in this literature. The reason for this could come from the author side or the reviewer side.
4. **Controlled user study (8 papers, 16%)** is a type of usability study, where a set of predefined tasks are performed by participants with a certain level of expertise (or novice participants after training) in a controlled environment. They may benefit from a large number of participants [FKSS06, WS09, GSCE11, MDM*14b, BSKR19]. We observe a decrease in popularity since 2014. Isenberg et al. suggest a controlled user study is typically time-consuming and resource-intensive to design, conduct and analyse [IIJ*13]. Controlled user studies are difficult to design for complex systems.

Table 15: ***Evaluation table:*** *An overview of evaluation techniques used in the literature, ordered by the popularity on the x-axis and the publication year on the y-axis. The x-axis represents the evaluation style with the number of participants shown in the individual cells.* ● *indicates an undisclosed number of participants. Green highlights context papers.*

	Evaluation				
Literature	**Domain Expert Feedback**	**Interview**	**Case Study**	**Controlled User Study**	**Year**
Plaisant et al. [PMS*98b]					1998
Horn et al. [HPU01]	2				2001
Bade et al. [BSM04]					2004
Goren-Bar et al. [GSG*04]	●				2004
Hinum et al. [HMA*05]	●	●			2005
Fails et al. [FKSS06]	8				2006
Bui et al. [BAK07]		●			2007
Wang et al. [WPS*09]		●	2		2009
Rind et al. [RMA*10]					2010
Faiola and Newlon [FN11]	16		●		2011
Gotz et al. [GSCE11]			2		2011
Gschwandtner et al. [GAK*11]	4	1			2011
Wongsuphasawat et al. [WGP*11]	●	1	2	10	2011
Zhang et al. [ZAR*11]	6				2011
Alonso and McCormick [AM12]					2012
Sopan et al. [SNK*12]	3	●		●	2012
Wongsuphasawat and Gotz [WG12]	3	●		●	2012
Monroe et al. [MLL*13]	●				2013
Ramírez-Ramírez et al. [RGT*13]					2013
Borland et al. [BWH14]					2014
Gotz and Stavropoulos [GS14]				12	2014
Kamaleswaran et al. [KPT*14]					2014
Malik et al. [MDM*14a]		4		10	2014
Bernard et al. [BSB*15]	●		●		2015
Bernard et al. [BSM*15]	6				2015
Federico et al. [FUS*15]		5			2015
Klemm et al. [KLG*15]	3	3	2		2015
Glueck et al. [GHC*16]	2	2			2016
Jiang et al. [JFB*16]		●			2016
Kamaleswaran et al. [KCJM16]	4	4			2016
Loorak et al. [LPK*16]	5	5			2016
Ola and Sedig [OS16]					2016
Dabek et al. [DJC17]		●			2017
Glueck et al. [GGC*17]		4	6		2017
Tong et al. [TML*17]	4		1		2017
Tong et al. [TRL*17]	2		1		2017
Glueck et al. [GND*18]	2	4			2018
Guo et al. [GXZ*18]	1	1	3		2018
Tong et al. [TML18]			1		2018
Trivedi et al. [TPC*18]	9	9		9	2018
Alemzadeh et al. [ANI*19]					2019
Bernard et al. [BSKR19]	10	14		14	2019
Glicksberg et al. [GOT*19]					2019
Guo et al. [GJG*19]	3	3	2		2019
Kwon et al. [KCK*19]	2		1		2019
McNabb and Laramee [ML19]			3		2019
Sultanum et al. [SSBC19]	6	12		6	2019
Zhang et al. [ZCD19]	6	6			2019
Jin et al. [JCG*20]	2	7	2		2020
Kwon et al. [KAS*20]	9	9	1		2020
Wang et al. [WLLP21]	2	3	3		2021
Total unique papers: 51	30	26	17	8	

6. Open Access Healthcare Data

Finding open access EHR data is very time-consuming and sometimes challenging, because VIS researchers are not often involved in EHR data collection and curation. This is usually performed by healthcare organizations. As a response to the challenges stemming from healthcare data visualization, we present a collection of open heath datasets, and our methodology for searching for open healthcare datasets, along with associated challenges, a carefully-defined scope and classification in this section. The result is a useful overview of healthcare data sources, with a curated list of publicly accessible healthcare datasets. The entire collection of data sources is accessible via our interactive EHR STAR Browser, available at https://ehr.wangqiru.com. We hope this section provides a helpful jump-start for potential researchers to develop visual healthcare data systems and form collaborations.

6.1. Healthcare data challenges

In this section, we discuss some major challenges faced in EHR data.

The *accessibility* of EHR data is one of the main barriers to researchers in general [MIT16]. We face several challenges searching for related data, which requires a considerable amount of time to search for. User registration and verification required by some data providers increases the manual labour. EHR data is more special due to its sensitive nature, and also comes in unstructured forms, e.g. clinic letters and hospital discharge letters. Converting the data into a structured form may lose valuable insight. Futhermore, an anonymization process is usually applied to EHR data by the respective data governance group.

Data quality is critical to EHR research, as much data is entered and computed manually, it is likely to contain incomplete and erroneous values. A special case is identified by Shneiderman and Plaisant [SP19] where a patient record was reported as being admitted 14 times but discharged only twice by a hospital. Verifying data quality requires a significant amount of time and effort. EHRs were not originally created with supporting research in mind [MIT16]. Overtime, the secondary use of EHR data in supporting healthcare research is emerging and widely accepted worldwide, this in turn improves the quality control measures for collecting them [KRN*19].

Data interoperability is challenging, given there is no standard definition of an EHR, healthcare providers often develop their own format to support the clinical work flow [CBC*17]. The lack of a standardized terminology, such as the UMLS, also contributes to this challenge.

These challenges remain unsolved. We see recent efforts in addressing these challenges, such as building a freely accessibility EHR database [GAG*00, JPS*16] and improving data validation and interoperability [FJV*09].

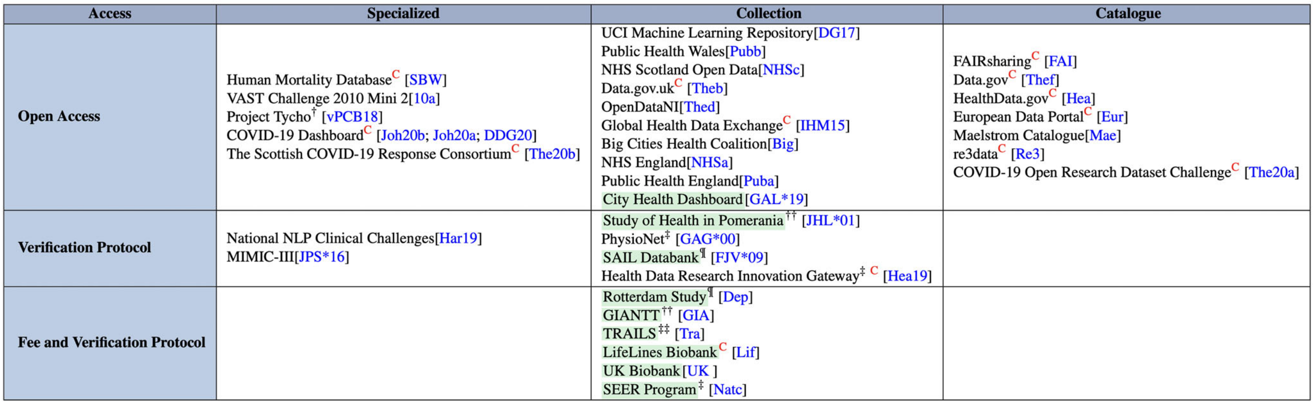
6.2. Healthcare Data search methodology

We focus on healthcare datasets that are openly accessible from a reputable data provider such as a non-profit organization, scientific

Table 16: *Keyword combinations used for discovering relevant healthcare data.*

Search Keyword Combinations		
open, free, public	electronic health record, electronic medical record, EHR, EMR personal health record, population health record, PHR, PopHR	data, dataset, database
	healthcare, health care, clinical, medical medicine, treatment, surgery, hospital	

Table 17: ***Data source table:*** *Data sources ordered by the year of establishment. See the detailed description of focus data sources in Section 6.5. Green highlights context data sources.* ^{C}*Contains COVID-19 data.* †*Registration required for open access.* ‡*Partially open access.* ‡‡*Free access for project collaborators, paid access for non-collaborators.* ¶*Free access for project collaborators, no access for non-collaborators.* ††*Data is not archived in English.*



Access	Specialized	Collection	Catalogue
Open Access	Human Mortality Database^{C} [SBW] VAST Challenge 2010 Mini 2[10a] Project Tycho† [vPCB18] COVID-19 Dashboard^{C} [Joh20b; Joh20a; DDG20] The Scottish COVID-19 Response Consortium^{C} [The20b]	UCI Machine Learning Repository[DG17] Public Health Wales[Pubb] NHS Scotland Open Data[NHSc] Data.gov.uk^{C} [Theb] OpenDataNI[Thed] Global Health Data Exchange^{C} [IHM15] Big Cities Health Coalition[Big] NHS England[NHSa] Public Health England[Puba] City Health Dashboard [GAL*19]	FAIRsharing^{C} [FAI] Data.gov^{C} [Thef] HealthData.gov^{C} [Hea] European Data Portal^{C} [Eur] Maelstrom Catalogue[Mae] re3data^{C} [Re3] COVID-19 Open Research Dataset Challenge^{C} [The20a]
Verification Protocol	National NLP Clinical Challenges[Har19] MIMIC-III[JPS*16]	Study of Health in Pomerania†† [JHL*01] PhysioNet‡ [GAG*00] SAIL Databank¶ [FJV*09] Health Data Research Innovation Gateway‡ ^{C} [Hea19]	
Fee and Verification Protocol		Rotterdam Study¶ [Dep] GIANTT†† [GIA] TRAILS‡‡ [Tra] LifeLines Biobank^{C} [Lif] UK Biobank [UK] SEER Program‡ [Natc]	

research or an initiative that provides trustworthy health related sources. We start by examining data sources mentioned in the related literature we found. Our search results are shown in Table 17. We check for conference associated events such as the annual IEEE Visualization Contest dating back to 2004 [VIS04], VAST challenges [VA 18] and National NLP Clinical Challenges (n2c2) [Har19] for relevant data. We also use keyword combinations listed in Table 16 with data search engines [Goob, Gooa] and well-known government data portals [Eur, Thef, Theb, Thed] to expand our results. We present 34 related healthcare datasets found in Table 17.

6.3. Healthcare data scope

The EHR data survey scope includes datasets that (1) offer free and open access to external researchers, (2) have greater than 500 records and 5 attributes in each record, (3) are published by credible providers, (4) have derived publications in peer-reviewed journals and (5) are archived in English for accessibility. To verify the eligibility, we examine each dataset, or the most popular datasets if multiple datasets are provided as a collection or catalogue. We refer to these as *focus data sources*.

Context Data and Out of Scope Healthcare Data Sources

During our search, we found some candidates that fulfil some but not all criteria. We still include them as *context data sources* in our data source overview Table 17.

We generally exclude datasets that require an access fee, with the exception of some candidates as context data sources. We generally exclude datasets that are accessible solely via project collaborations. We generally exclude datasets that are not archived in English. However, we do include some as context data sources (if they are high quality) in Table 17 for interested readers, and describe them in Section 6.5.4.

We exclude datasets that are not directly related to EHR. Here are some noteworthy examples.

Health IT Dashboard [Thec] provides datasets on the adoption, utilization and performance of information technology in healthcare facilities sponsored by the US government, these datasets are excluded. **The VAST Challenge 2010 Mini Challenge 3** [VAS10b] provides a dataset on genetic sequences for tracing the mutations of the Drafa virus. Each sequence of single molecules is coded as a single alphabet, therefore the dataset does not contain any actual EHR information and is excluded. **The VAST Challenge 2011 Mini Challenge 1** [VAS11a] provides data containing posts collected from social media platforms for the identification of an epidemic outbreak, these datasets are excluded due to the lack of an EHR dimension.

6.4. Healthcare data sources classification

We present a description of data sources in this section. Table 17 displays an overview of data sources we found.

Based on the *focus data source* and *context data source* introduced above, we classify data sources into three categories:

A *specialized source* refers to datasets focusing on a single specialty or area of specialization. The Human Mortality Database [SBW] provides multiple datasets specifically on all-cause mortality from over 50 countries or regions, therefore we classify it as a specialized focus data source.

A *collection source* provides access to multiple datasets from different specialties, such as the UCI Machine Learning Repository [DG17], which provides data on breast cancer, diabetes, hepatitis and other diseases.

A *catalogue source* does not host data on its own website but provides links to other webpages, The Registry of Research Data Repositories (r3data) [Re3] is a catalogue source that hosts over 2,000 scientific datasets, each comes with a comprehensive description and a link pointing to its homepage.

6.5. Open access healthcare data sources

Based on the classification, we briefly describe each open access healthcare data source in their corresponding section. We describe each data source using the Five Ws [ZBA*13]:

- Who the data provider is
- When the data was collected and published
- Where the data was collected
- Why the data was collected
- What the data contains

6.5.1. *Specialized healthcare data sources*

This section describes focus data sources that focus on a single health related specialty.

Human Mortality Database began as a collaborative project in 2000 [SBW], involving research teams in the Department of Demography at the University of California, Berkeley, USA and the Max Planck Institute for Demographic Research in Rostock, Germany. The database provides open access to detailed mortality and population data for over 50 countries and regions to promote relevant research. Depending on the geographical location, data archives may span over a century.

VAST Challenge 2010 Mini 2 [VAS10a], as a part of the IEEE Conference on Visual Analytics Science and Technology (VAST), provides open access to data such as hospital admittance and death records in several cities involved in a major fictitious epidemic outbreak in 2009.

Project Tycho was launched by the University of Pittsburgh in 2013 [vPCB18], incorporating a collection of death rate data from infectious diseases and their historical spread between 1888 - 2014. The initial archive focused on the history of diseases throughout the US. It has now expanded to include over 360 datasets on 92 infectious diseases at a global level in a standard format.

The COVID-19 Dashboard is an online interactive dashboard developed by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [Joh20b, Joh20a, DDG20], as a real time visualization for the number of COVID-19 cases, deaths and recovery rates around the world. The raw data is available for open access.

The Scottish COVID-19 Response Consortium (SCRC) is founded by the University of Glasgow, the consortium includes a group of epidemiologists, mathematicians, computer scientists for developing new models to help inform the control of COVID-19 in Scotland. It offers open access to COVID-19 related data provided by 15 healthboard areas of NHS Scotland [CARA*20].

6.5.2. *Collection healthcare data sources*

This section describes focus data sources that provide access to multiple datasets from different specialties.

UCI Machine Learning Repository was created by David Aha and fellow graduate students at University of California Irvine in 1987, as a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The repository contains over 110 health related datasets, including subjects such as breast cancer, diabetes, epilepsy and more.

The National Health Service (NHS) of the United Kingdom provides open access to various healthcare data collected through its operation, the data is made accessible via different portals including **Public Health Wales** (established in 1999) [Pubb], **NHS Scotland Open Data** (2009) [NHSc], **The Government Digital Service** (2011) [Theb], **OpenDataNI** (2012) [Thed], **Public Health England** (2017) [Puba] and **NHS England** (2017) [NHSa]. Example datasets hosted on these portals including mortality rate from cancer, liver, cardiovascular diseases and more.

Big Cities Health Coalition [Big] is a forum founded in 2014, serves as a platform for the leaders of 14 largest metropolitan health departments in the US, to exchange strategies and jointly address challenges related to promoting and protecting the health and safety of the people they serve. The forum provides open access to data including mortality from various causes, maternal and child health, HIV etc., covering over 62 million people from 2010-2016.

Global Health Data Exchange [IHM15] operated by the Institute for Health Metrics and Evaluation, provides a catalog of global health and demographic data. It currently hosts over 12 billion population health records collected from 195 countries. The mission of the exchange is to serve as *a critical resource for informed policymaking*. The exchange supports searching and filtering data by over 350 diseases, injuries and risk factors.

6.5.3. *Catalogue healthcare data sources*

This section describes catalogue data sources that do not host data on their website but provide links to other data sources.

FAIRsharing [FAI] started in 2007 as a community-driven registry providing descriptions of standards, databases and data policies. Datasets can be published on FAIRsharing to increase the visibility and foster collaboration. The registry not only hosts a catalogue of health related databases, but also provides access to proven standards and data policies to reduce the potential for unnecessary reinventions.

The **U.S. Government's Open Data** [Thef] and **HealthData.gov** [Hea] started offering links to datasets in 2011, to ensure compliance with relevant Open Data Policy and promote research and innovation. Public entities ranging from federal agencies to local government departments collected over 200,000 datasets, including popular healthcare data on cancer, diabetes and hypertension.

The **European Data Portal** [Eur] was established in 2012 aiming to serve as a point of access to public data published by institutions, agencies and other bodies across European countries. Over 10,000 health related datasets including HIV-related, norovirus and cancer are available.

Maelstrom Catalogue [Mae] is a catalogue of epidemiological research founded by McGill University in 2012. The catalogue later expanded to include population health studies, to promote collaborative research. It currently hosts links to over 200 well-known research projects.

re3data [Re3] is funded by the German Research Foundation in 2012, as a global registry of over 2,000 research data repositories from multiple academic disciplines. It aims to provide permanent storage and access to healthcare data for the scientific community.

The **COVID-19 Open Research Dataset Challenge** [The20a] is a challenge launched in 2020 by the Allen Institute for Artificial Intelligence on Kaggle, an online community of data scientists. The challenge offers over 59,000 academic journals for free, in order to attract researchers and develop novel solutions to study the ongoing evolution of COVID-19. Some 1,300 novel solutions have been submitted and many are accompanied by open access anonymized patient data, as a part of the submission requirements.

6.5.4. *Context healthcare data sources*

A context healthcare data source refers to a data source that does not fulfil all criteria listed in Section 6.3, but we include and describe some high quality sources here for interested readers.

UK Biobank [UK] recruited 500,000 participants aged between 40-69 years in the U.K. from 2006 - 2010, with extensive physical measurements and blood, urine and saliva samples collected in conjunction with wearable monitors and online assessments of personal well-being. Researchers are obliged to return their results and findings to benefit the research community. We include the UK Biobank as a *context data* only as it charges a one time access fee of £2,100 (reduced to £600 for researchers from developing countries or students).

LifeLines Biobank [SRP*08, Lif] archives 167,000 participants including all age groups in the Netherlands. The research collects physical and physiological measurements such as blood pressure, skin autofluorescence, and biomaterials such as blood and urine, from participants, along with regular online questionnaires on stress and quality of life. We include LifeLines Biobank as a context data source only as it charges a one time access fee of approximately € 7,800.

Tracking Adolescents' Individual Lives Survey (TRAILS) [Tra] is an ongoing research project that studies the psychological, social and physical development of over 2,500 adolescents in the Netherlands since the year 2000. The research is conducted in the form of questionnaires and interviews on topics such as cognitive functioning, academic performance, tests on fitness condition, and physical measurements such as baroreflex sensitivity. We include TRAILS as a context data source only as it charges a one time access fee of over € 3,000, however, the fee is waived if a collaboration is formed with the TRAILS research group.

Rotterdam Study [Dep] is another well-known population-based study ongoing in Ommoord, Rotterdam since 1990, with a focus on the risk factors of cardiovascular, neurological, ophthalmological and endocrine diseases in the elderly aged 55 years and over. Three cohorts (1990, 2000, 2006) included 14,926 participants and has resulted in over 2,000 scientific articles. We include the study as a context data source only as it charges an access fee and the access is only granted to collaborations formed with the study's principal investigators.

Secure Anonymised Information Linkage (SAIL) Databank [FJV*09] was established in the UK in 2006. It allows external researchers to access billions of EHRs on datasets such as outpatient, critical care and primary GP care in the UK. Access to additional restricted datasets such as bowel screening, breast test and cervical screening in Wales is granted with additional approval from data providers. We include this noteworthy SAIL Databank as a context data source as the access is granted via project collaboration only.

Study of Health in Pomerania (SHIP) [JHL*01, VAS*11b] started after the German reunification in the 1990s, as a population-based epidemiological study. The study includes 7,008 women and men aged 20 - 79 years, with a wide range of medical data being collected. We include SHIP as a context data source due to its lack of accessibility since the study is primarily archived in German.

Groningen Initiative to Analyse Type 2 Diabetes Treatment (GIANTT) [GIA] is a project aimed at the quality of care for people with type 2 diabetes in Groningen, the Netherlands since 2004. The primary data source is from local general practices. We include GIANTT as a context data source only due to its restricted accessibility since the study is in Dutch. GIANTT also charges an access fee.es an access fee.es an access fee.es an access fee.

7. Future Research Challenges and Discussion

In this section, potential future research directions are derived from the discussion of the challenges reported in the literature. Future work and challenges are often discussed at the end of each research paper. Table 18 summarizes a list of the top 10 most popular future challenges we extract from the reviewed literature, ordered by their popularity. We observe that the top future challenges are to tackle scalability as data size grows, conduct additional in-depth and effective evaluations and improve the efficiency in screen space utilization. Another popular challenge is the interoperability between different EHR Vis systems, which can be potentially addressed by adopting a common terminology standard such as the UMLS. Finally, the ability to increase system usability while simultaneously introducing advanced interactive user options, is a popular future research direction.

Table 18: ***Challenge table:*** *A summary of future challenges identified in the literature, ordered by the publication year on the x-axis and the frequency on the y-axis. We use 1-2 words to represent these challenges in the table header, and describe them in detail in Section 7.*

	Challenges										
Literature	**Scalability**	**Evaluation**	**Screen space**	**Interoperability**	**Usability**	**Interaction**	**Dimensionality**	**Uncertainty**	**Clustering**	**Access**	**Year**
Plaisant et al.[PMS*98b]						🟥	🟥				1998
Horn et al.[HPU01]	🟥										2001
Bade et al.[BSM04]	🟥					🟥					2004
Goren-Bar et al.[GSG*04]						🟥					2004
Hinum et al.[HMA*05]				🟥		🟥					2005
Fails et al.[FKSS06]	🟥				🟥						2006
Bui et al.[BAK07]	🟥			🟥			🟥			🟥	2007
Wang et al.[WPS*09]					🟥			🟥			2009
Rind et al.[RMA*10]		🟥			🟥				🟥		2010
Faiola and Newlon[FN11]			🟥								2011
Gotz et al.[GSCE11]	🟥	🟥									2011
Gschwandtner et al.[GAK*11]					🟥						2011
Wongsuphasawat et al.[WGP*11]				🟥							2011
Zhang et al.[ZAR*11]			🟥								2011
Alonso and McCormick[AM12]	🟥							🟥			2012
Sopan et al.[SNK*12]	🟥	🟥		🟥							2012
Wongsuphasawat and Gotz[WG12]		🟥	🟥				🟥				2012
Monroe et al.[MLL*13]	🟥					🟥					2013
Ramírez-Ramírez et al.[RGT*13]				🟥							2013
Borland et al.[BWH14]			🟥								2014
Gotz and Stavropoulos[GS14]		🟥									2014
Kamaleswaran et al.[KPT*14]	🟥						🟥	🟥			2014
Malik et al.[MDM*14a]						🟥					2014
Bernard et al.[BSB*15]			🟥								2015
Bernard et al.[BSM*15]	🟥								🟥		2015
Federico et al.[FUS*15]	🟥					🟥					2015
Klemm et al.[KLG*15]	🟥	🟥									2015
Glueck et al.[GHC*16]										🟥	2016
Jiang et al.[JFB*16]	🟥		🟥			🟥					2016
Kamaleswaran et al.[KCJM16]			🟥		🟥				🟥		2016
Loorak et al.[LPK*16]								🟥		🟥	2016
Ola and Sedig[OS16]	🟥	🟥		🟥				🟥			2016
Dabek et al.[DJC17]			🟥				🟥				2017
Glueck et al.[GGC*17]	🟥			🟥							2017
Tong et al.[TML*17]	🟥						🟥				2017
Tong et al.[TRL*17]	🟥	🟥					🟥				2017
Glueck et al.[GND*18]		🟥			🟥						2018
Guo et al.[GXZ*18]					🟥	🟥					2018
Tong et al.[TML18]		🟥									2018
Trivedi et al.[TPC*18]		🟥		🟥							2018
Alemzadeh et al.[ANI*19]	🟥				🟥						2019
Bernard et al.[BSKR19]	🟥		🟥						🟥		2019
Glicksberg et al.[GOT*19]			🟥	🟥							2019
Guo et al.[GJG*19]	🟥										2019
Kwon et al.[KCK*19]	🟥	🟥			🟥						2019
McNabb and Laramee[ML19]		🟥									2019
Sultanum et al.[SSBC19]	🟥	🟥		🟥	🟥						2019
Zhang et al.[ZCD19]		🟥									2019
Jin et al.[JCG*20]	🟥		🟥		🟥			🟥			2020
Kwon et al.[KAS*20]			🟥		🟥			🟥			2020
Wang et al.[WLLP21]	🟥				🟥					🟥	2021
Total unique papers: 51	24	15	12	10	13	9	7	7	4	4	

Table 19: *Visualization techniques applied in the literature. We follow the classification of visualization techniques by Keim [Kei02], and categorize bar chart, line chart and pie chart as standard 2d display.* Ⓔ *indicates the technique is applied in the literature.* 🟢 *indicates a customized variant of the technique is applied in the literature.*

	Visualization Techniques																					
Literature	Area	Box and Whisker	Bubble	Cartogram	Chord	Choropleth Map	Glyph	Heatmap	Histogram	Map	Matrix	Parallel Coordinates	Parallel Sets	Sankey	Scatterplot	Standard 2D Displays	Stream Graph	Sunburst	Timeline	Tree Diagram	Treemap	Year
Plaisant et al.[PMS*98b]																			🟢			1998
Horn et al.[HPU01]							🔴									🔴						2001
Bade et al.[BSM04]		🟢																				2004
Goren-Bar et al.[GSG*04]																🔴			🔴			2004
Hinum et al.[HMA*05]							🔴															2005
Fails et al.[FKSS06]							🔴									🔴			🔴			2006
Bui et al.[BAK07]																🔴						2007
Wang et al.[WPS*09]							🔴		🔴										🔴			2009
Rind et al.[RMA*10]																🔴						2010
Faiola and Newlon[FN11]																			🔴			2011
Gotz et al.[GSCE11]																					🔴	2011
Gschwandtner et al.[GAK*11]							🔴		🔴							🔴				🔴		2011
Wongsuphasawat et al.[WGP*11]							🔴												🟢			2011
Zhang et al.[ZAR*11]																		🟢				2011
Alonso and McCormick[AM12]								🔴		🔴						🔴						2012
Sopan et al.[SNK*12]						🔴		🔴								🔴						2012
Wongsuphasawat and Gotz[WG12]														🔴								2012
Monroe et al.[MLL*13]							🔴															2013
Ramírez-Ramírez et al.[RGT*13]										🔴						🔴						2013
Borland et al.[BWH14]					🟢							🟢										2014
Gotz and Stavropoulos[GS14]			🔴				🔴		🔴				🔴	🔴								2014
Kamaleswaran et al.[KPT*14]								🔴														2014
Malik et al.[MDM*14a]							🔴									🔴			🟢			2014
Bernard et al.[BSB*15]																🔴						2015
Bernard et al.[BSM*15]	🔴	🔴							🔴							🔴						2015
Federico et al.[FUS*15]								🔴								🔴				🔴		2015
Klemm et al.[KLG*15]								🟢														2015
Glueck et al.[GHC*16]							🔴									🔴		🔴				2016
Jiang et al.[JFB*16]	🔴					🔴										🔴						2016
Kamaleswaran et al.[KCJM16]			🔴				🔴				🔴						🔴					2016
Loorak et al.[LPK*16]									🔴		🔴	🔴				🔴						2016
Ola and Sedig[OS16]			🟢		🟢	🔴	🔴	🔴		🟢			🟢	🟢		🔴					🔴	2016
Dabek et al.[DJC17]							🔴				🔴					🔴		🔴	🔴	🔴		2017
Glueck et al.[GGC*17]							🔴		🔴		🔴							🔴				2017
Tong et al.[TML*17]				🔴												🔴					🔴	2017
Tong et al.[TRL*17]				🔴												🔴					🔴	2017
Glueck et al.[GND*18]							🔴								🔴	🔴		🔴	🔴			2018
Guo et al.[GXZ*18]							🟢									🔴			🟢		🔴	2018
Tong et al.[TML18]				🔴																		2018
Trivedi et al.[TPC*18]											🔴									🟢		2018
Alemzadeh et al.[ANI*19]					🔴						🔴	🔴				🔴						2019
Bernard et al.[BSKR19]		🔴					🔴		🔴							🔴						2019
Glicksberg et al.[GOT*19]																🔴						2019
Guo et al.[GJG*19]							🟢									🔴					🟢	2019
Kwon et al.[KCK*19]	🔴						🔴		🔴						🔴							2019
McNabb and Laramee[ML19]	🟢						🟢									🔴						2019
Sultanum et al.[SSBC19]							🔴				🔴							🔴	🔴			2019
Zhang et al.[ZCD19]		🟢					🔴									🔴						2019
Jin et al.[JCG*20]							🟢						🟢			🔴						2020
Kwon et al.[KAS*20]					🔴		🔴					🟢	🔴			🔴	🔴					2020
Wang et al.[WLLP21]							🔴				🔴				🔴							2021
Total unique paper: 51	4	4	3	3	4	3	25	6	8	3	8	4	4	3	3	30	2	6	11	4	6	

Scalability (22 papers, 45%) and data **dimensionality (7 papers, 14%)** are reported as a future challenge 28 times in total. As the result of data growth exceeds the capacity of existing EHR Vis systems [LK06]. Apart from the handing of high-dimensional and multivariate EHR data, maintaining the system availability in a real world scenario where multiple users are accessing the system concurrently, is a trending future research direction [SNK*12]. From the table, we can see this has been a persistent theme.

While scalability a challenge for all visualization systems, we make note of how the following challenges are inherent to EHR visualization.

In-depth **evaluation (14 papers, 29%)** and validation including quantitative studies, qualitative studies and validation is reported 14 times as the second most popular future research direction. An in-depth evaluation and validation helps to reveal the weakness and potential improvements for the system. We examine and describe the evaluation techniques adopted by the literature in Section 5. Some 14 papers report the lack of evaluation or an insufficient number of participants in their studies. The recruitment of qualified participants is challenging, these participants often do not have the time to complete lengthy and thorough evaluations. The table of challenges indicates this as a prominent theme in recent years.

Limited **screen space (12 papers, 24%)** constrains the content visualized and reduces the effectiveness of an EHR system [GOT*19]. As the probability of using multiple views increases in EHR Vis systems, we categorize this challenge as a domain-specific one. Features with less significance are often hidden to make space for others [BSKR19]. This may result in over-simplification and missing potential insights [MLL*13]. This is highly related to the challenge of visual aggregation and **clustering (4 papers, 8%)** of multiple patients and requires more advanced **interaction (9 papers, 18%)** techniques to explore and navigate the data, especially the temporal dimension. Table 7 indicates that interaction is a popular future challenge in earlier years.

Data **interoperability (10 papers, 20%)** between EHR Vis systems and institutions continues to lag [MIT16] and is reported 10 times as a future challenge. This increases the difficulty for researchers to incorporate data from heterogeneous sources in varying formats [OS16]. Although Table 3 indicates that some papers focus on same UMLS terms, these EHR Vis systems are built specifically for their given datasets and do not offer interoperability. This is a very EHR-specific challenge that can be potentially addressed by promoting collaboration between different research groups on same topics, and adopting a common terminology standard such as the UMLS. Table 7 indicates limited screen space and data interoperability as re-occurring challenges over the last 10 years.

System **usability (12 papers, 24%)** and human factors are reported by 11 papers as a future challenge direction. Low usability often results in a longer learning curve that requires more training time for users [WPS*09, KCJM16]. This in turn may increase the occurrence of human errors. Due to the domain expertise required, it is difficult to conduct a full usability test on EHR Vis systems.

Data quality and **uncertainty (7 papers, 14%)** is another challenge reported in 7 papers. Data often contains missing or incorrect values, this requires further investigation during data collection and pre-processing [OPS16].

Open data access (4 papers, 8%) is reported 3 times, as the authors of most papers we review are collaborating with domain experts or institutions. However, access to high quality data still remains a big challenge for many researchers [MIT16]. We attempt to address this challenge here in Section 6. Even though the sensitive nature of EHR data requires special permission, open data access and accessibility are not mentioned more often in the literature. This is likely due to the collaborations formed between visualization and medical experts: in Section 5, we find 59% of the papers choose to collaborate with medical experts, who also provide EHR data for visualization researchers.

More advanced visual designs: Table 19 shows an overview of visualization techniques applied in all papers included in this STAR. We observe that standard 2D displays and glyph are the most popular techniques among 21 techniques found across all EHR Vis systems. This implies that using advanced visual techniques to mitigate scalability challenge brought by EHR data dimensionality, remains understudied.

8. Conclusions

In this STAR, we present an up-to-date overview of research papers, with an in-depth investigation of 99 in the field of EHR and PopHR Visualization and Visual Analytics. We investigate some of the most commonly used terminology in the field and categorize the literature based on six re-occurring research themes. Our STAR differs from the eight related surveys, by including 29 more recent publications, as well as a novel classification that utilizes UMLS, as a means to improve the understanding of recent development in research and foster potential interdisciplinary collaborations. We then investigate the evaluation techniques adopted by the literature. Furthermore, we invest over two months in investigating a collection of 34 high quality open access datasets, aims to serve as a starting point for potential researchers. Lastly, our interactive EHR STAR Browser enables the reader to easily navigate through all literature and data sources collected in this STAR.

Acknowledgments

This work is funded by the grant EP/S010238/2 from the Engineering and Physical Sciences Research Council (EPSRC). EPSRC is a British Research Council that provides government funding for grants to undertake research and postgraduate degrees in engineering and the physical sciences.

References

[Act09] The American Recovery and Reinvestment Act of 2009. jan 2009. URL: https://www.congress.gov/bill/111th-congress/house-bill/1/text.

[AKDA15] Agrawal R., Kadadi A., Dai X., Andres F.: Challenges and opportunities with big data visualization. In

Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems (New York, NY, USA, oct 2015), vol. 13, ACM, pp. 169–173. https://doi.org/10.1145/2857218.2857256.

[AKOS18] Abukhodair F., Khashoggi K., O'Connell T., Shaw C.: RadStream: An interactive visual display of radiology workflow for delay detection in the clinical imaging process. *2017 IEEE Workshop on Visual Analytics in Healthcare, VAHC 2017* (2018), 69–76. https://doi.org/10.1109/VAHC.2017.8387543.

[Alp10] Alpaydin E.: *Introduction to Machine Learning*, second ed. MIT Press, Cambridge, 2010.

[AM06] Aigner W., Miksch S.: CareVis: Integrated visualization of computerized protocols and temporal patient data. *Artificial Intelligence in Medicine 37*, 3 (2006), 203–218. https://doi.org/10.1016/j.artmed.2006.04.002.

[AM12] Alonso W. J., McCormick B. J.: EPIPOI: A user-friendly analytical tool for the extraction and visualization of temporal parameters from epidemiological time series. *BMC Public Health 12*, 1 (dec 2012), 982. https://doi.org/10.1186/1471-2458-12-982.

[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 Graphics Forum 39*, 1 (feb 2019), 63–75. https://doi.org/10.1111/cgf.13662.

[AU02] Atkinson P., Unwin D.: Density and local attribute estimation of an infectious disease using MapInfo. *Computers & Geosciences 28*, 9 (nov 2002), 1095–1105. https://doi.org/10.1016/S0098-3004(02)00026-2.

[BAK07] Bui A. A. T., Aberle D. R., Kangarloo H.: TimeLine: Visualizing Integrated Patient Records. *IEEE Transactions on Information Technology in Biomedicine 11*, 4 (jul 2007), 462–473. https://doi.org/10.1109/TITB.2006.884365.

[Ber99] Bertin J.: Graphics and Graphic Information Processing. In *Readings in Information Visualization*, Card S. K., Mackinlay J. D., Shneiderman B., (Eds.). Morgan Kaufmann, San Francisco, 1999, pp. 62–65.

[BGD05] Brodbeck D., Gasser R., Degen M.: Enabling large-scale telemedical disease management through interactive visualization. *European Notes in Medical Informatics* (2005).

[BHW*09] Bashyam V., Hsu W., Watt E., Bui A. A. T., Kangarloo H., Taira R. K.: Problem-centric Organization and Visualization of Patient Imaging and Clinical Data. *RadioGraphics 29*, 2 (mar 2009), 331–343. https://doi.org/10.1148/rg.292085098.

[BMM*06] Blanton J. D., Manangan A., Manangan J., Hanlon C. A., Slate D., Rupprecht C. E.: Development of a GIS-based, real-time Internet mapping tool for rabies surveillance. *International Journal of Health Geographics 5* (2006), 1–8. https://doi.org/10.1186/1476-072X-5-47.

[Bod04] Bodenreider O.: The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research 32*, DATABASE ISS. (jan 2004), 267D–270. https://doi.org/10.1093/nar/gkh061.

[BRMF19] Bernardini M., Romeo L., Misericordia P., Frontoni E.: Discovering the Type 2 Diabetes in Electronic Health Records using the Sparse Balanced Support Vector Machine. *IEEE Journal of Biomedical and Health Informatics* (2019), 1–1. https://doi.org/10.1109/jbhi.2019.2899218.

[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 the 2015 Workshop on Visual Analytics in Healthcare - VAHC '15* (New York, New York, USA, 2015), vol. 25-October, ACM Press, pp. 1–8. https://doi.org/10.1145/2836034.2836035.

[BSKR19] Bernard J., Sessler D., Kohlhammer J., Ruddle R. A.: Using dashboard networks to visualize multiple patient histories: A design study on post-operative prostate cancer. *IEEE Transactions on Visualization and Computer Graphics 25*, 3 (2019), 1615–1628. https://doi.org/10.1109/TVCG.2018.2803829.

[BSM04] Bade R., Schlechtweg S., Miksch S.: Connecting time-oriented data and information to a coherent interactive visualization. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04* (New York, New York, USA, 2004), no. May 2014, ACM Press, pp. 105–112. https://doi.org/10.1145/985692.985706.

[BSM*15] Bernard J., Sessler D., May T., Schlomm T., Pehrke D., Kohlhammer J.: A visual-interactive system for prostate cancer cohort analysis. *IEEE Computer Graphics and Applications 35*, 3 (2015), 44–55. https://doi.org/10.1109/MCG.2015.49.

[Bus04] Bush G. W.: Executive Order 13335—Incentives for the Use of Health Information Technology and Establishing the Position of the National Health Information Technology Coordinator. *Federal Register 69*, 84 (2004), 24059–24061. URL: https://georgewbush-whitehouse.archives.gov/news/releases/2004/04/20040427-4.html.

[BWH14] Borland D., West V. L., Hammond W. E.: Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates. *Proceedings of the 2014 Workshop on Visual Analytics in Healthcare* (2014), 19–24.

[CAD*14] Carroll L. N., Au A. P., Detwiler L. T., Fu T.-c., Painter I. S., Abernethy N. F.: Visualization and analytics tools for infectious disease epidemiology: A systematic review. *Journal of Biomedical Informatics 51*, 1 (oct 2014), 287–298. https://doi.org/10.1016/j.jbi.2014.04.006.

[CARA*20] Chen M., Abdul-Rahman A., Archambault D., Dykes J., Slingsby A., Ritsos P. D., Torsney-Weir T., Turkay C., Bach B., Brett A., Fang H., Jianu R., Khan S., Laramee R. S., Nguyen P. H., Reeve R., Roberts J. C., Vidal F., Wang Q., Wood J., Xu K.: RAMPVIS: Towards a New Methodology for Developing Visualisation Capabilities for Large-scale Emergency Responses. *arXiv Preprint* (2020). arXiv:2012.04757.

[CBC*17] Cowie M. R., Blomster J. I., Curtis L. H., Duclaux S., Ford I., Fritz F., Goldman S., Janmohamed S., Kreuzer J., Leenay M., Michel A., Ong S., Pell J. P., Southworth M. R., Stough W. G., Thoenes M., Zannad F., Zalewski A.: Electronic health records to facilitate clinical research. *Clinical Research in Cardiology 106*, 1 (2017), 1–9. https://doi.org/10.1007/s00392-016-1025-6.

[CCDW17] Caballero H. S. G., Corvo A., Dixit P. M., Westenberg M. A.: Visual analytics for evaluating clinical pathways. In *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)* (oct 2017), IEEE, pp. 39–46. https://doi.org/10.1109/VAHC.2017.8387499.

[CCT03] Chittaro L., Combi C., Trapasso G.: Data mining on temporal data: a visual approach and its clinical application to hemodialysis. *Journal of Visual Languages & Computing 14*, 6 (dec 2003), 591–620. https://doi.org/10.1016/j.jvlc.2003.06.003.

[CKS*09] Connors J., Krzywinski M., Schein J., Gascoyne R., Horsman D., Jones S. J., Marra M. A.: Circos: An information aesthetic for comparative genomics. *Genome Research 19*, 604 (2009), 1639–1645. https://doi.org/10.1101/gr.092759.109.19.

[CXR18] Chen Y., Xu P., Ren L.: Sequence Synopsis: Optimize Visual Summary of Temporal Event Data. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 45–55. https://doi.org/10.1109/TVCG.2017.2745083.

[DGG*07] Da Silva F. A., Gagliardi H. F., Gallo E., Madope M. A., Neto V. C., Pisa I. T., Alves D.: IntegraEPI: A grid-based epidemic surveillance system. *Studies in Health Technology and Informatics 126*, February (2007), 197–206.

[DGM*11] Driscoll T., Gabbard J. L., Mao C., Dalay O., Shukla M., Freifeld C. C., Hoen A. G., Brownstein J. S., Sobral B. W.: Integration and visualization of host–pathogen data related to infectious diseases. *Bioinformatics 27*, 16 (aug 2011), 2279–2287. https://doi.org/10.1093/bioinformatics/btr391.

[DJC17] Dabek F., Jimenez E., Caban J. J.: A timeline-based framework for aggregating and summarizing electronic health records. In *2017 IEEE Workshop on Visual Analytics in Healthcare, VAHC 2017* (2017), pp. 55–61. https://doi.org/10.1109/VAHC.2017.8387501.

[DMPM15] Dunne C., Muller M., Perra N., Martino M.: VoroGraph: Visualization tools for epidemic analysis. *Conference on Human Factors in Computing Systems - Proceedings 18* (2015), 255–258. https://doi.org/10.1145/2702613.2725459.

[DSP*17] Du F., Shneiderman B., Plaisant C., Malik S., Perer A.: Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus. *IEEE Transactions on Visualization and Computer Graphics 23*, 6 (jun 2017), 1636–1649. https://doi.org/10.1109/TVCG.2016.2539960.

[Els] Elsevier: Mendeley - Reference Management Software & Researcher Network. URL: https://www.mendeley.com/.

[Eur17] European Commission: Health - Horizon 2020, 2017. URL: https://ec.europa.eu/programmes/horizon2020/en/area/health.

[Eva16] Evans R. S.: Electronic Health Records: Then, Now, and in the Future. *Yearbook of Medical Informatics 25*, S 01 (aug 2016), S48–S61. https://doi.org/10.15265/IYS-2016-s006.

[FH99] Friedman C., Hripcsak G.: Natural language processing and its future in medicine. *Academic Medicine 74*, 8 (aug 1999), 890–5. https://doi.org/10.1097/00001888-199908000-00012.

[FKSS06] Fails J., Karlson A., Shahamat L., Shneiderman B.: A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across Multiple Histories. In *2006 IEEE Symposium On Visual Analytics And Technology* (oct 2006), IEEE, pp. 167–174. https://doi.org/10.1109/VAST.2006.261421.

[FMRB14] Freifeld C., Mandl K., Reis B., Brownstein J.: HealthMap: Global Infectious Disease Monitoring through. *Journal of the American Medical Informatics Association 15*, 2 (2014), 150–157. https://doi.org/10.1197/jamia.M2544. Introduction.

[FN11] Faiola A., Newlon C.: Advancing Critical Care in the ICU: A Human-Centered Biomedical Data Visualization Systems. In *International Conference on Ergonomics and Health Aspects of Work with Computers* (2011), vol. 6779 LNCS, pp. 119–128. https://doi.org/10.1007/978-3-642-21716-6_13.

[FP10] Friedman D. J., Parrish R. G.: The population health record: concepts, definition, design, and implementation. *Journal of the American Medical Informatics Association 17*, 4 (jul 2010), 359–366. https://doi.org/10.1136/jamia.2009.001578.

[FUS*15] Federico P., Unger J., Sacchi L., Klimov D., Miksch S.: Gnaeus: utilizing clinical guidelines for knowledge-assisted visualisation of EHR cohorts. *EuroVis Workshop on Visual Analytics* (2015). https://doi.org/10.2312/eurova.20151108.

[GAK*11] Gschwandtner T., Aigner W., Kaiser K., Miksch S., Seyfang A.: CareCruiser: Exploring and visualizing plans, events, and effects interactively. In *IEEE Pacific Visualization Symposium 2011, PacificVis 2011 - Proceedings* (2011), pp. 43–50. https://doi.org/10.1109/PACIFICVIS.2011.5742371.

[GAS*14] Gálvez J. A., Ahumada L., Simpao A. F., Lin E. E., Bonafide C. P., Choudhry D., England W. R., Jawad A. F., Friedman D., Sesok-Pizzini D. A., Rehman M. A.: Visual analytical tool for evaluation of 10-year perioperative transfusion practice at a children's hospital. *Journal of the American Medical Informatics Association 21*, 3 (may 2014), 529–534. https://doi.org/10.1136/amiajnl-2013-002241.

[GAW*11] Gleicher M., Albers D., Walker R., Jusufi I., Hansen C. D., Roberts J. C.: Visual comparison for information visualization. *Information Visualization 10*, 4 (oct 2011), 289–309. https://doi.org/10.1177/1473871611416549.

[GB16] Gotz D., Borland D.: Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization. *IEEE Computer Graphics and Applications 36*, 3 (may 2016), 90–96. https://doi.org/10.1109/MCG.2016.59.

[GBMG17] Gesulga J. M., Berjame A., Moquiala K. S., Galido A.: Barriers to Electronic Health Record System Implementation and Information Systems Resources: A Structured Review. *Procedia Computer Science 124* (2017), 544–551. https://doi.org/10.1016/j.procs.2017.12.188.

[GBSGA*04] Goren-Bar D., Shahar Y., Galperin-Aizenberg M., Boaz D., Tahan G.: Knave II: The definition and implementation of an intelligent tool for visualization and exploration of time-oriented clinical data. In *Proceedings of the working conference on Advanced visual interfaces - AVI '04* (New York, New York, USA, 2004), ACM Press, p. 171. https://doi.org/10.1145/989863.989889.

[GGC*17] Glueck M., Gvozdik A., Chevalier F., Khan A., Brudno M., Wigdor D.: PhenoStacks: Cross-Sectional Cohort Phenotype Comparison Visualizations. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 191–200. https://doi.org/10.1109/TVCG.2016.2598469.

[GHC*16] Glueck M., Hamilton P., Chevalier F., Breslav S., Khan A., Wigdor D., Brudno M.: PhenoBlocks: Phenotype Comparison Visualizations. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 101–110. https://doi.org/10.1109/TVCG.2015.2467733.

[GJG*19] Guo S., Jin Z., Gotz D., Du F., Zha H., Cao N.: Visual Progression Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (jan 2019), 417–426. https://doi.org/10.1109/TVCG.2018.2864885.

[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 (nov 2012), 954–959. https://doi.org/10.1136/amiajnl-2011-000486.

[GMA*08] Gao S., Mioc D., Anton F., Yi X., Coleman D. J.: Online GIS services for mapping and sharing disease information. *International Journal of Health Geographics 7*, 1 (2008), 8. https://doi.org/10.1186/1476-072X-7-8.

[GNDV*18] Glueck M., Naeini M. P., Doshi-Velez F., Chevalier F., Khan A., Wigdor D., Brudno M.: PhenoLines: Phenotype Comparison Visualizations for Disease Subtyping via Topic Models. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 371–381. https://doi.org/10.1109/TVCG.2017.2745118.

[Gooc] Google: Google Scholar. URL: https://scholar.google.com/.

[GOT*19] Glicksberg B. S., Oskotsky B., Thangaraj P. M., Giangreco N., Badgeley M. A., Johnson K. W., Datta D., Rudrapatna V. A., Rappoport N., Shervey M. M., Miotto R., Goldstein T. C., Rutenberg E., Frazier R., Lee N., Israni S., Larsen R., Percha B., Li L., Dudley J. T., Tatonetti N. P., Butte A. J.: PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model. *Bioinformatics 35*, 21 (nov 2019), 4515–4518. https://doi.org/10.1093/bioinformatics/btz409.

[GS14] Gotz D., Stavropoulos H.: DecisionFlow: Visual analytics for high-dimensional temporal event sequence data. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1783–1792. https://doi.org/10.1109/TVCG.2014.2346682.

[GSCE11] Gotz D., Sun J., Cao N., Ebadollahi S.: Visual cluster analysis in support of clinical decision intelligence. *AMIA ... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium* (2011), 481–490.

[GT05] Gunter T. D., Terry N. P.: The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. *Journal of Medical Internet Research 7*, 1 (mar 2005), e3. https://doi.org/10.2196/jmir.7.1.e3.

[GTC*10] Goldsmith M.-R., Transue T. R., Chang D. T., Tornero-Velez R., Breen M. S., Dary C. C.: PAVA: physiological and anatomical visual analytics for mapping of tissue-specific concentration and time-course data. *Journal of Pharmacokinetics and Pharmacodynamics 37*, 3 (jun 2010), 277–287. https://doi.org/10.1007/s10928-010-9160-6.

[Guo07] Guo D.: Visual analytics of spatial interaction patterns for pandemic decision support. *International Journal of Geographical Information Science 21*, 8 (sep 2007), 859–877. https://doi.org/10.1080/13658810701349037.

[GXZ*18] Guo S., Xu K., Zhao R., Gotz D., Zha H., Cao N.: EventThread: Visual Summarization and Stage Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (jan 2018), 56–65. https://doi.org/10.1109/TVCG.2017.2745320.

[Hal08] Hallett C.: Multi-modal presentation of medical histories. In *Proceedings of the 13th international conference on Intelligent user interfaces - IUI '08* (New York, New York, USA, 2008), ACM Press, p. 80. https://doi.org/10.1145/1378773.1378785.

[HAP11] Hripcsak G., Albers D. J., Perotte A.: Exploiting time in electronic health record correlations. *Journal of the American*

Medical Informatics Association 18, Supplement_1 (dec 2011), i109–i115. https://doi.org/10.1136/amiajnl-2011-000463.

[HBAS17] HEART T., BEN-ASSULI O., SHABTAI I.: A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy. *Health Policy and Technology 6*, 1 (mar 2017), 20–25. https://doi.org/10.1016/j.hlpt.2016.08.002.

[HDE*08] HEITGERD J. L., DENT A. L., ELMORE K. A., KAPLAN B., HOLT J. B., METZLER M. M., MELFI K., STANLEY J. M., HIGHSMITH K., KANAREK N., FREDERICKSON COMER K.: Community health status indicators: adding a geospatial component. *Preventing chronic disease 5*, 3 (jul 2008), A96.

[HHH16] HOGAN T., HINRICHS U., HORNECKER E.: The Elicitation Interview Technique: Capturing People's Experiences of Data Representations. *IEEE Transactions on Visualization and Computer Graphics 22*, 12 (dec 2016), 2579–2593. https://doi.org/10.1109/TVCG.2015.2511718.

[HHN00] HAVRE S., HETZLER B., NOWELL L.: ThemeRiver: visualizing theme changes over time. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings* (2000), IEEE Comput. Soc, pp. 115–123. https://doi.org/10.1109/INFVIS.2000.885098.

[HMA*05] HINUM K., MIKSCH S., AIGNER W., OHMANN S., POPOW C., POHL M., RESTER M.: Gravi++: Interactive information visualization to explore highly structured temporal data. *Journal of Universal Computer Science 11*, 11 (2005), 1792–1805. https://doi.org/10.3217/jucs-011-11-1792.

[HPU01] HORN W., POPOW C., UNTERASINGER L.: Support for fast comprehension of ICU data: Visualization using metaphor graphics. *Methods of Information in Medicine* (2001). https://doi.org/10.1055/s-0038-1634202.

[HZC*07] HU P. J.-H., ZENG D., CHEN H., LARSON C., CHANG W., TSENG C., MA J.: System for Infectious Disease Information Sharing and Analysis: Design and Evaluation. *IEEE Transactions on Information Technology in Biomedicine 11*, 4 (jul 2007), 483–492. https://doi.org/10.1109/TITB.2007.893286.

[Iak98] IAKOVIDIS I.: Towards personal health record: current situation, obstacles and trends in implementation of electronic healthcare record in Europe. *International Journal of Medical Informatics 52*, 1-3 (1998), 105–115. https://doi.org/10.1016/s1386-5056(98)00129-4.

[IEE] IEEE: IEEE Xplore Digital Library. URL: https://ieeexplore.ieee.org/Xplore/home.jsp.

[IHK*17] ISENBERG P., HEIMERL F., KOCH S., ISENBERG T., XU P., STOLPER C. D., SEDLMAIR M., CHEN J., MOLLER T., STASKO J.: Vispubdata.org: A Metadata Collection About IEEE Visualization (VIS) Publications. *IEEE Transactions on Visualization and Computer Graphics 23*, 9 (sep 2017), 2199–2206. https://doi.org/10.1109/TVCG.2016.2615308.

[IIJ*13] ISENBERG T., ISENBERG P., JIAN Chen, SEDLMAIR M., MOLLER T.: A Systematic Review on the Practice of Evaluating Visualization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (dec 2013), 2818–2827. https://doi.org/10.1109/TVCG.2013.126.

[IZCC08] ISENBERG P., ZUK T., COLLINS C., CARPENDALE S.: Grounded evaluation of information visualizations. In *Proceedings of the 2008 conference on BEyond time and errors novel evaLuation methods for Information Visualization - BELIV '08* (New York, New York, USA, 2008), ACM Press, p. 1. https://doi.org/10.1145/1377966.1377974.

[JCG*20] JIN Z., CUI S., GUO S., GOTZ D., SUN J., CAO N.: CarePre: An Intelligent Clinical Decision Assistance System. *ACM Transactions on Computing for Healthcare 1*, 1 (mar 2020), 1–20. https://doi.org/10.1145/3344258.

[JFB*16] JIANG S., FANG S., BLOOMQUIST S., KEIPER J., PALAKAL M., XIA Y., GRANNIS S.: Healthcare Data Visualization: Geospatial and Temporal Integration. In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (2016), vol. 2, SCITEPRESS - Science and and Technology Publications, pp. 212–219. https://doi.org/10.5220/0005714002120219.

[JS12] JOSHI R., SZOLOVITS P.: Prognostic physiology: modeling patient severity in Intensive Care Units using radial domain folding. *AMIA ... Annual Symposium proceedings. AMIA Symposium 2012* (2012), 1276–83.

[KAS*20] KWON B. C., ANAND V., SEVERSON K. A., GHOSH S., SUN Z., FROHNERT B. I., LUNDGREN M., NG K.: DPVis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways. *IEEE Transactions on Visualization and Computer Graphics 2626*, c (2020), 1–1. http://arxiv.org/abs/1904.11652 arXiv:1904.11652, https://doi.org/10.1109/TVCG.2020.2985689.

[KCJM16] KAMALESWARAN R., COLLINS C., JAMES A., MCGREGOR C.: PhysioEx: Visual Analysis of Physiological Event Streams. *Computer Graphics Forum 35*, 3 (2016), 331–340. https://doi.org/10.1111/cgf.12909.

[KCK*19] KWON B. C., CHOI M. J., KIM J. T., CHOI E., KIM Y. B., KWON S., SUN J., CHOO J.: RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics* (2019). https://doi.org/10.1109/TVCG.2018.2865027.

[KDBB19] KOLECK T. A., DREISBACH C., BOURNE P. E., BAKKEN S.: Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *Journal of the American Medical Informatics Association 26*, 4 (2019), 364–379. https://doi.org/10.1093/jamia/ocy173.

[Kei02] KEIM D.: Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 1–8. https://doi.org/10.1109/2945.981847.

[KLG*15] KLEMM P., LAWONN K., GLABER S., NIEMANN U., HEGENSCHEID K., VOLZKE H., PREIM B.: 3D Regression Heat Map Analysis of Population Study Data. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (jan 2015), 81–90. https://doi.org/10.1109/TVCG.2015.2468291.

[KM01] KOSARA R., MIKSCH S.: Metaphors of movement: a visualization and user interface for time-oriented, skeletal plans. *Artificial Intelligence in Medicine 22*, 2 (may 2001), 111–131. https://doi.org/10.1016/S0933-3657(00)00103-2.

[KNK10] KUMASAKA N., NAKAMURA Y., KAMATANI N.: The textile plot: A new Linkage disequilibrium display of multiple-Single Nucleotide Polymorphism genotype data. *PLoS ONE 5*, 4 (2010), 1–12. https://doi.org/10.1371/journal.pone.0010207.

[KPT*14] KAMALESWARAN R., PUGH J. E., THOMMANDRAM A., JAMES A., MCGREGOR C.: Visualizing Neonatal Spells: Temporal Visual Analytics of High Frequency Cardiorespiratory Physiological Event Streams. In *IEEE VIS 2014 Workshop on Visualization of Electronic Health Records* (2014), pp. 1–4.

[KRN*19] KIM E., RUBINSTEIN S. M., NEAD K. T., WOJCIESZYNSKI A. P., GABRIEL P. E., WARNER J. L.: The Evolving Use of Electronic Health Records (EHR) for Research. *Seminars in Radiation Oncology 29*, 4 (oct 2019), 354–361. https://doi.org/10.1016/j.semradonc.2019.05.010.

[KSTM10] KLIMOV D., SHAHAR Y., TAIEB-MAIMON M.: Intelligent selection and retrieval of multiple time-oriented records. *Journal of Intelligent Information Systems 35*, 2 (oct 2010), 261–300. https://doi.org/10.1007/s10844-009-0100-0.

[Lar11] LARAMEE R. S.: How to Read a Visualization Research Paper: Extracting the Essentials. *IEEE Computer Graphics and Applications 31*, 3 (may 2011), 78–82. URL: http://ieeexplore.ieee.org/document/5754296/, https://doi.org/10.1109/MCG.2011.44.

[Lar14] LARAMEE R. S.: eHealth On the Horizon. In *Vizualizing Electronic Health Record Data (Workshop)* (2014), no. 1, pp. 2–4.

[LFL*11] LEWIS S. L., FEIGHNER B. H., LOSCHEN W. A., WOJCIK R. A., SKORA J. F., COBERLY J. S., BLAZES D. L.: SAGES: A Suite of Freely-Available Software Tools for Electronic Disease Surveillance in Resource-Limited Settings. *PLoS ONE 6*, 5 (may 2011), e19750. https://doi.org/10.1371/journal.pone.0019750.

[Lid01] LIDDY E.: Natural Language Processing, 2001.

[LK06] LARAMEE R. S., KOSARA R.: Challenges and Unsolved Problems. In *Human-Centered Visualization Environments*, vol. 4417. Springer-Verlag, 2006, pp. 231–254.

[LPK*16] LOORAK M. H., PERIN C., KAMAL N., HILL M., CARPENDALE S.: TimeSpan: Using Visualization to Explore Temporal Multi-dimensional Data of Stroke Patients. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 409–418. https://doi.org/10.1109/TVCG.2015.2467325.

[LRS12] LIVNAT Y., RHYNE T., SAMORE M.: Epinome: A Visual-Analytics Workbench for Epidemiology Data. *IEEE Computer Graphics and Applications 32*, 2 (mar 2012), 89–95. https://doi.org/10.1109/MCG.2012.31.

[MBS*12] MANE K. K., BIZON C., SCHMITT C., OWEN P., BURCHETT B., PIETROBON R., GERSING K.: VisualDecisionLinc: A visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. *Journal of Biomedical Informatics 45*, 1 (feb 2012), 101–106. https://doi.org/10.1016/j.jbi.2011.09.003.

[MDM*14a] MALIK S., DU F., MONROE M., ONUKWUGHA E., PLAISANT C., SHNEIDERMAN B.: An Evaluation of Visual Analytics Approaches to Comparing Cohorts of Event Sequences. *Proc. of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records* (2014), 1–6.

[MDM*14b] MALIK S., DU F., MONROE M., ONUKWUGHA E., PLAISANT C., SHNEIDERMAN B.: An Evaluation of Visual Analytics Approaches to Comparing Cohorts of Event Sequences. *Proc. of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records* (2014), 1–6.

[MDM*15] MALIK S., DU F., MONROE M., ONUKWUGHA E., PLAISANT C., SHNEIDERMAN B.: Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics. In *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15* (New York, New York, USA, 2015), vol. 2015-Janua, ACM Press, pp. 38–49. https://doi.org/10.1145/2678025.2701407.

[MIT16] MIT Critical Data: *Secondary Analysis of Electronic Health Records*. Springer International Publishing, Cham, 2016. https://doi.org/10.1007/978-3-319-43742-2.

[ML17] MCNABB L., LARAMEE R. S.: Survey of Surveys (SoS) - Mapping The Landscape of Survey Papers in Information Visualization. In *Computer Graphics Forum* (jun 2017), vol. 36, The Eurographs Association & John Wiley & Sons, Ltd., pp. 589–617. https://doi.org/10.1111/cgf.13212.

[ML19] MCNABB L., LARAMEE R. S.: Multivariate Maps—A Glyph-Placement Algorithm to Support Multivariate Geospatial Visualization. *Information 10*, 10 (sep 2019), 302. https://doi.org/10.3390/info10100302.

[MLK16] MASOODIAN M., LUZ S., KAVENGA D.: Nu-view: A visualization system for collaborative Co-located analysis of geospatial disease data. *ACM International Conference Proceeding Series 01-05-Febr* (2016). https://doi.org/10.1145/2843043.2843374.

[MLL*13] MONROE M., LAN R., LEE H., PLAISANT C., SHNEIDERMAN B.: Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2227–2236. https://doi.org/10.1109/TVCG.2013.200.

[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 & Computing 22*, 4 (aug 2011), 268–278. https://doi.org/10.1016/j.jvlc.2011.04.002.

[MSD*16] MAURIELLO M. L., SHNEIDERMAN B., DU F., MALIK S., PLAISANT C.: Simplifying overviews of temporal event sequences. *Conference on Human Factors in Computing Systems - Proceedings 07-12-May-* (2016), 2217–2224. https://doi.org/10.1145/2851581.2892440.

[Mun09] MUNZNER T.: A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (nov 2009), 921–928. https://doi.org/10.1109/TVCG.2009.111.

[MWP*12] MONROE M., WONGSUPHASAWAT K., PLAISANT C., SHNEIDERMAN B., MILLSTEIN J., GOLD S.: Exploring Point and Interval Event Patterns: Display Methods and Interactive Visual Query. *HCIL Technical Report, Dept Computer Science, University of Maryland*, May (2012), 1–10.

[Natb] National Cancer Institute: Definition of personal health record. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/personal-health-record.

[Natc] National Cancer Institute: Surveillance, Epidemiology, and End Results Program. URL: https://seer.cancer.gov/.

[Nau10] NAUMOVA E. N.: Visual analytics for immunologists: Data compression and fractal distributions. *Self/Nonself 1*, 3 (jul 2010), 241–249. https://doi.org/10.4161/self.1.3.12876.

[NHSb] NHS Digital: Personal Health Records definition. URL: https://digital.nhs.uk/services/personal-health-records-adoption-service/personal-health-records-adoption-toolkit/initiating-a-personal-health-record/personal-health-records-definition.

[Obs20] Observational Health Data Sciences and Informatics: *The Book of OHDSI*, 1 ed. 2020. URL: http://book.ohdsi.org.

[OPS16] ONUKWUGHA E., PLAISANT C., SHNEIDERMAN B.: Data Visualization Tools for Investigating Health Services Utilization Among Cancer Patients. In *Oncology Informatics*. Elsevier Inc., 2016, pp. 207–229. https://doi.org/10.1016/b978-0-12-802115-6.00011-2.

[OS16] OLA O., SEDIG K.: Beyond simple charts: Design of visualizations for big health data. *Online Journal of Public Health Informatics 8*, 3 (dec 2016). https://doi.org/10.5210/ojphi.v8i3.7100.

[PFH07] PIECZKIEWICZ D. S., FINKELSTEIN S. M., HERTZ M. I.: Design and evaluation of a web-based interactive visualization system for lung transplant home monitoring data. *AMIA ... Annual Symposium proceedings. AMIA Symposium* (oct 2007), 598–602.

[PL20] PREIM B., LAWONN K.: A Survey of Visual Analytics for Public Health. *Computer Graphics Forum 39*, 1 (feb 2020), 543–580. https://doi.org/10.1111/cgf.13891.

[PMR*96] PLAISANT C., MILASH B., ROSE A., WIDOFF S., SHNEIDERMAN B.: LifeLines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems common ground* (New York, New York, USA, 1996), ACM Press, pp. 221–ff. https://doi.org/10.1145/238386.238493.

[PMS*98a] PLAISANT C., MUSHLIN R., SNYDER A., LI J., HELLER D., SHNEIDERMAN B.: LifeLines: using visualization to enhance navigation and analysis of patient records. *Proceedings. AMIA Symposium 48*, 9 (sep 1998), 76–80. https://doi.org/10.1016/B978-155860915-0/50038-X.

[PMS*98b] PLAISANT C., MUSHLIN R., SNYDER A., LI J., HELLER D., SHNEIDERMAN B.: LifeLines: using visualization to enhance navigation and analysis of patient records. In *Proceedings. AMIA Symposium* (1998), pp. 76–80.

[PS12] PERER A., SUN J.: MatrixFlow: temporal network visual analytics to track symptom evolution during disease progression. *AMIA ... Annual Symposium proceedings. AMIA Symposium 2012* (2012), 716–25.

[RBL*13] RAJWAN Y. G., BARCLAY P. W., LEE T., SUN I.-F., PASSARETTI C., LEHMANN H.: Visualizing Central Line-Associated Blood Stream Infection (CLABSI) Outcome Data to Health Care Consumers and Practitioners for Decision Making – Evaluation Study. *Online Journal of Public Health Informatics 5*, 2 (jun 2013), 1–18. https://doi.org/10.5210/ojphi.v5i2.4364.

[REA*08] REINHARDT M., ELIAS J., ALBERT J., FROSCH M., HARMSEN D., VOGEL U.: EpiScanGIS: an online geographic surveillance system for meningococcal disease. *International Journal of Health Geographics 7*, 1 (2008), 33. https://doi.org/10.1186/1476-072X-7-33.

[Res] ResearchGate: Search Publications | ResearchGate. URL: https://www.researchgate.net/search/publications.

[RFG*17] RIND A., FEDERICO P., GSCHWANDTNER T., AIGNER W., DOPPLER J., WAGNER M.: Visual Analytics of Electronic Health Records with a Focus on Time. In *New Perspectives in Medical Records*, Capello F., Rinaldi G., Gatti G., (Eds.). 2017, ch. 5, pp. 65–77. https://doi.org/10.1007/978-3-319-28661-7.

[RJ76] ROBERTSON S. E., JONES K. S.: Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 3 (may 1976), 129–146. https://doi.org/10.1002/asi.4630270302.

[RMA*10] RIND A., MIKSCH S., AIGNER W., TURIC T., POHL M.: VisuExplore: gaining new medical insights from visual exploration. *Proceedings of the 1st International Workshop on Interactive Systems in Healthcare (WISH@CHI2010)* (2010), 149–152.

[ROC*18] RAJKOMAR A., OREN E., CHEN K., DAI A. M., HAJAJ N., HARDT M., LIU P. J., LIU X., MARCUS J., SUN M., SUNDBERG P., YEE H., ZHANG K., ZHANG Y., FLORES G., DUGGAN G. E., IRVINE J., LE Q., LITSCH K., MOSSIN A., TANSUWAN J., WANG D., WEXLER J., WILSON J., LUDWIG D., VOLCHENBOUM S. L., CHOU K., PEARSON M., MADABUSHI S., SHAH N. H., BUTTE

A. J., Howell M. D., Cui C., Corrado G. S., Dean J.: Scalable and accurate deep learning with electronic health records. *npj Digital Medicine 1*, 1 (dec 2018), 18. https://doi.org/10.1038/s41746-018-0029-1.

[RRGT*13] Ramírez-Ramírez L. L., Gel Y. R., Thompson M., de Villa E., McPherson M.: A new surveillance and spatio-temporal visualization tool SIMID: SIMulation of Infectious Diseases using random networks and GIS. *Computer Methods and Programs in Biomedicine 110*, 3 (jun 2013), 455–470. https://doi.org/10.1016/j.cmpb.2013.01.007.

[RST10] Roque F. S., Slaughter L., Tkatšenko A.: A Comparison of Several Key Information Visualization Systems for Secondary Use of Electronic Health Record Content. *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, June (2010), 76–83.

[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. https://doi.org/10.1561/1100000039.

[SAD*14] Simpao A. F., Ahumada L. M., Desai B. R., Bonafide C. P., Galvez J. A., Rehman M. A., Jawad A. F., Palma K. L., Shelov E. D.: Optimization of drug-drug interaction alert rules in a pediatric hospital's electronic health record system using a visual analytics dashboard. *Journal of the American Medical Informatics Association 22*, 2 (oct 2014), 361–369. https://doi.org/10.1136/amiajnl-2013-002538.

[SAGR14] Simpao A. F., Ahumada L. M., Gálvez J. A., Rehman M. A.: A Review of Analytics and Clinical Informatics in Health Care. *Journal of Medical Systems 38*, 4 (apr 2014), 45. https://doi.org/10.1007/s10916-014-0045-x.

[SKD12] Stubbs B., Kale D. C., Das A.: Sim•TwentyFive: an interactive visualization system for data-driven decision support. *AMIA ... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium 2012* (2012), 891–900.

[SMB*10] Steenwijk M. D., Milles J., Buchem M. A., Reiber J. H., Botha C. P.: Integrated Visual Analysis for Heterogeneous Datasets in Cohort Studies. *IEEE VisWeek Workshop on Visual Analytics in Health Care* (2010).

[SNK*12] Sopan A., Noh A. S. I., Karol S., Rosenfeld P., Lee G., Shneiderman B.: Community Health Map: A geospatial and multivariate data visualization tool for public health datasets. *Government Information Quarterly 29*, 2 (2012), 223–234. https://doi.org/10.1016/j.giq.2011.10.002.

[SNLOB17] Shaban-Nejad A., Lavigne M., Okhmatovskaia A., Buckeridge D. L.: PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data. *Annals of the New York Academy of Sciences 1387*, 1 (jan 2017), 44–53. https://doi.org/10.1111/nyas.13271.

[SP19] Shneiderman B., Plaisant C.: Interactive Visual Event Analytics: Opportunities and Challenges. *Computer 52*, 1 (jan 2019), 27–35. https://doi.org/10.1109/MC.2018.2890217.

[SSBC19] Sultanum N., Singh D., Brudno M., Chevalier F.: Doccurate: A Curation-Based Approach for Clinical Text Visualization. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 142–151. https://doi.org/10.1109/TVCG.2018.2864905.

[STBR18] Shickel B., Tighe P. J., Bihorac A., Rashidi P.: Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics 22*, 5 (2018), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063.

[SZ00] Stasko J., Zhang E.: Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings* (2000), IEEE Comput. Soc, pp. 57–65. https://doi.org/10.1109/INFVIS.2000.885091.

[TAB*06] Tang P. C., Ash J. S., Bates D. W., Overhage J. M., Sands D. Z.: Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption. *Journal of the American Medical Informatics Association 13*, 2 (mar 2006), 121–126. https://doi.org/10.1197/jamia.M2025.

[Thea] The Association for Computing Machinery: ACM Digital Library. URL: https://dl.acm.org/.

[Thee] The U.S. Centers for Medicare and Medicaid Services: Electronic Health Records. URL: https://www.cms.gov/Medicare/E-Health/EHealthRecords.

[The19] The Allen Institute for Artificial Intelligence: Semantic Scholar - An academic search engine for scientific articles, 2019. URL: https://www.semanticscholar.org/.

[The20b] The Scottish COVID-19 Response Consortium: The Scottish COVID-19 Response Consortium (SCRC), 2020. URL: https://www.gla.ac.uk/research/az/scrc/#d.en.727867http://www.gla.ac.uk/research/az/suerc/.

[TML*17] Tong C., McNabb L., Laramee R. S., Lyons J., Walters A. M., Thayer D., Berridge D.: Time-oriented Cartographic Treemaps for Visualization of Public Healthcare Data Chao. In *Computer Graphics & Visual Computing* (2017).

[TML18] Tong C., Mcnabb L., Laramee R. S.: Cartograms with Topological Features. In *Computer Graphics & Visual Computing* (2018).

[TPC*18] Trivedi G., Pham P., Chapman W. W., Hwa R., Wiebe J., Hochheiser H.: NLPReViz: an interactive tool for natural language processing on clinical text. *Journal of the American Medical Informatics Association: JAMIA 25*, 1 (jan 2018), 81–87. https://doi.org/10.1093/jamia/ocx070.

[TRL*17] Tong C., Roberts R., Laramee R. S., Berridge D., Thayer D.: Cartographic Treemaps for Visualization of Public Healthcare Data. In *Computer Graphics & Visual Computing* (2017).

[WAM01] WEBER M., ALEXA M., MULLER W.: Visualizing time-series on spirals. In *IEEE Symposium on Information Visualization, 2001. INFOVIS 2001.* (2001), IEEE, pp. 7–13. https://doi.org/10.1109/INFVIS.2001.963273.

[WBH15] WEST V. L., BORLAND D., HAMMOND W. E.: Innovative information visualization of electronic health record data: A systematic review. *Journal of the American Medical Informatics Association 22*, 2 (mar 2015), 330–339. https://doi.org/10.1136/amiajnl-2014-002955.

[WG12] WONGSUPHASAWAT K., GOTZ D.: Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (dec 2012), 2659–2668. https://doi.org/10.1109/TVCG.2012.225.

[WGP*11] WONGSUPHASAWAT K., GUERRA GÓMEZ J. A., PLAISANT C., WANG T. D., TAIEB-MAIMON M., SHNEIDERMAN B.: LifeFlow: Visualizing an Overview of Event Sequences. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (New York, New York, USA, 2011), ACM Press, p. 1747. https://doi.org/10.1145/1978942.1979196.

[Wil10] WILLISON B.: Advancing Meaningful Use: Simplifying Complex Clinical Metrics Through Visual Representation. *PIIM Research* (2010).

[WLLP21] WANG Q., LARAMEE R. S., LACEY A., PICKRELL W. O.: LetterVis: a letter-space view of clinic letters. *The Visual Computer 37*, 9-11 (sep 2021), 2643–2656. https://doi.org/10.1007/s00371-021-02171-w.

[WP17] World Health Organziation, Pan American Health Organization: Handbook for Electronic Health Records Implementation. 75. URL: http://www.paho.org/ict4health/images/docs/DRAFT-Handbook_EHR_Implementation.pdf.

[WPS*09] WANG T., PLAISANT C., SHNEIDERMAN B., SPRING N., ROSEMAN D., MARCHAND G., MUKHERJEE V., SMITH M.: Temporal Summaries: Supporting Temporal Categorical Searching, Aggregation and Comparison. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (nov 2009), 1049–1056. https://doi.org/10.1109/TVCG.2009.187.

[WS09] WONGSUPHASAWAT K., SHNEIDERMAN B.: Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In *VAST 09 - IEEE Symposium on Visual Analytics Science and Technology, Proceedings* (2009). https://doi.org/10.1109/VAST.2009.5332595.

[WZM*16] WANG X. M., ZHANG T. Y., MA Y. X., XIA J., CHEN W.: A Survey of Visual Analytic Pipelines. *Journal of Computer Science and Technology 31*, 4 (2016), 787–804. https://doi.org/10.1007/s11390-016-1663-1.

[XW05] XU R., WUNSCHII D.: Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks 16*, 3 (may 2005), 645–678. https://doi.org/10.1109/TNN.2005.845141.

[YHH*08] YI Q., HOSKINS R. E., HILLRINGHOUSE E. A., SORENSEN S. S., OBERLE M. W., FULLER S. S., WALLACE J. C.: Integrating open-source technologies to build low-cost information systems for improved access to public health data. *International Journal of Health Geographics 7*, 1 (2008), 29. https://doi.org/10.1186/1476-072X-7-29.

[ZAR*11] ZHANG Z., AHMED F., RAMAKRISHNAN A. M. I. V., ZHAO R., VICCELLIO A., MUELLER K.: AnamneVis: a framework for the visualization of patient history and medical diagnostics chains. *IEEE VAHC Workshop*, January (2011), 1–4.

[ZBA*13] ZHIYUAN Zhang, BING Wang, AHMED F., RAMAKRISHNAN I. V., RONG Zhao, VICCELLIO A., MUELLER K.: The Five Ws for Information Visualization with Application to Healthcare Informatics. *IEEE Transactions on Visualization and Computer Graphics 19*, 11 (nov 2013), 1895–1910. https://doi.org/10.1109/TVCG.2013.89.

[ZCD19] ZHANG Y., CHANANA K., DUNNE C.: IDMVis: Temporal Event Sequence Visualization for Type 1 Diabetes Treatment Decision Support. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 512–522. https://doi.org/10.1109/TVCG.2018.2865076.

[ZXG19] ZHANG X., XIAO J., GU F.: Applying support vector machine to electronic health records for cancer classification. In *2019 Spring Simulation Conference, SpringSim 2019* (apr 2019), IEEE, pp. 1–9. https://doi.org/10.23919/SpringSim.2019.8732906.

Data References

[VIS04] IEEE Visualization 2004 Contest, 2004. URL: http://vis.computer.org/vis2004contest/, https://doi.org/10.1109/visual.2004.43.

[VAS10a] VAST Challenge 2010 MC2 - Characterization of Pandemic Spread, 2010. URL: https://www.cs.umd.edu/hcil/varepository/VASTChallenge2010/challenges/MC2-CharacterizationofPandemicSpread/.

[VAS10b] VAST Challenge 2010 MC3 - Tracing the Mutations of a Disease, 2010. URL: https://www.cs.umd.edu/hcil/varepository/VASTChallenge2010/challenges/MC3-TracingtheMutationsofaDisease/.

[VAS11a] VAST Challenge 2011 MC1 - Characterization of an Epidemic Spread, 2011. URL: http://www.cs.umd.edu/hcil/varepository/VASTChallenge2011/challenges/MC1-CharacterizationofanEpidemicSpread/.

[Big] Big Cities Health Coalition: Data Platform — Big Cities Health Coalition. URL: https://www.bigcitieshealth.org/city-data/.

[DDG20] Dong E., Du H., Gardner L.: An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* (feb 2020). https://doi.org/10.1016/S1473-3099(20)30120-1.

[Dep] Department of Epidemiology Erasmus University Medical Center: The Rotterdam Study. URL: http://www.erasmus-epidemiology.nl/research/ergo.htm.

[DG17] Dua D., Graff C.: {UCI} Machine Learning Repository, 2017. URL: http://archive.ics.uci.edu/ml/index.phphttp://archive.ics.uci.edu/ml.

[Eur] European Union: European Data Portal. URL: https://www.europeandataportal.eu/en.

[Eur16] European Parliament and Council of the European Union: Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive), 2016. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504&from=EN.

[FAI] FAIRsharing: FAIRsharing. URL: https://fairsharing.org/.

[FJV*09] Ford D. V., Jones K. H., Verplancke J.-P., Lyons R. A., John G., Brown G., Brooks C. J., Thompson S., Bodger O., Couch T., Leake K.: The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Services Research 9*, 1 (dec 2009), 157. https://doi.org/10.1186/1472-6963-9-157.

[GAG*00] Goldberger A. L., Amaral L. A. N., Glass L., Hausdorff J. M., Ivanov P. C., Mark R. G., Mietus J. E., Moody G. B., Peng C.-K., Stanley H. E.: PhysioBank, PhysioToolkit, and PhysioNet. *Circulation 101*, 23 (jun 2000). https://doi.org/10.1161/01.CIR.101.23.e215.

[GAL*19] Gourevitch M. N., Athens J. K., Levine S. E., Kleiman N., Thorpe L. E.: City-Level Measures of Health, Health Determinants, and Equity to Foster Population Health Improvement: The City Health Dashboard. *American Journal of Public Health 109*, 4 (apr 2019), 585–592. https://doi.org/10.2105/AJPH.2018.304903.

[GIA] GIANTT: GIANTT – Groningen Initiative to Analyse Type 2 diabetes Treatment. URL: https://www.giantt.nl/.

[Gooa] Google: Dataset Search. URL: https://datasetsearch.research.google.com/.

[Goob] Google: Google. URL: https://www.google.com/.

[Har19] Harvard Medical School: N2C2: National NLP Clinical Challenges, 2019. URL: https://n2c2.dbmi.hms.harvard.edu/.

[Hea] HealthData.gov: HealthData.gov. URL: https://healthdata.gov/.

[Hea19] Health Data Research UK: Health Data Research Innovation Gateway, 2019. URL: https://healthdatagateway.org/.

[IHM15] IHME University of Washington: GBD Compare, 2015. URL: http://vizhub.healthdata.org/gbd-compare.

[JHL*01] John U., Hensel E., Lü, Demann 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 SPM 46*, 3 (may 2001), 186–194. https://doi.org/10.1007/BF01324255.

[Joh20a] Johns Hopkins University: Coronavirus COVID-19 (2019-nCoV), 2020. URL: https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6.

[Joh20b] Johns Hopkins University: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE, 2020. URL: https://github.com/CSSEGISandData/COVID-19.

[JPS*16] Johnson A. E., Pollard T. J., Shen L., Lehman L. W. H., Feng M., Ghassemi M., Moody B., Szolovits P., Anthony Celi L., Mark R. G.: MIMIC-III, a freely accessible critical care database. *Scientific Data 3* (may 2016). https://doi.org/10.1038/sdata.2016.35.

[Lif] Lifelines Biobank: Lifelines Biobank. URL: https://www.lifelines.nl/.

[Mae] Maelstrom Research: Maelstrom Catalogue. URL: https://www.maelstrom-research.org/maelstrom-catalogue.

[Nata] National Cancer Institute: Definition of electronic health record. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/electronic-health-record.

[NHSa] NHS Commissioning Board: NHS England Data Catalogue. URL: https://data.england.nhs.uk/.

[NHSc] NHS Scotland Open Data: Datasets - NHS Scotland Open Data. URL: https://www.opendata.nhs.scot/dataset.

[Puba] Public Health England: PHE data and analysis tools - GOV.UK. URL: https://www.gov.uk/guidance/phe-data-and-analysis-tools.

[Pubb] Public Health Wales: Data - Public Health Wales. URL: https://phw.nhs.wales/data/.

[Re3] Re3data.org: Registry of Research Data Repositories. URL: https://www.re3data.org/, https://doi.org/10.17616/R3D.

[SBW] Shkolnikov V., Barbieri M., Wilmoth J.: Human Mortality Database. URL: https://www.mortality.org/.

[SRP*08] Stolk R. P., Rosmalen J. G. M., Postma D. S., de Boer R. A., Navis G., Slaets J. P. J., Ormel J., Wolffenbuttel B. H. R.: Universal risk factors for multifactorial diseases. *European Journal of Epidemiology 23*, 1 (jan 2008), 67–74. https://doi.org/10.1007/s10654-007-9204-4.

[Theb] The Government Digital Service: Find open data - data.gov.uk. URL: https://data.gov.uk/.

[Thec] The Office of the National Coordinator for Health Information Technology: Health IT Data. URL: https://dashboard.healthit.gov/datadashboard/data.php.

[Thed] The Open Data Team: Open Data NI. URL: https://www.opendatani.gov.uk/.

[Thef] The U.S. General Services Administration: Data.gov. URL: https://www.data.gov/.

[The20a] The Allen Institute for Artificial Intelligence: COVID-19 Open Research Dataset Challenge (CORD-19) | Kaggle, 2020. URL: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/kernels.

[Tra] Trails: Tracking Adolescents' Individual Lives Survey. URL: https://www.trails.nl/en.

[UK] UK Biobank: UK Biobank. URL: https://www.ukbiobank.ac.uk/.

[VA 18] VA Community: VAST Challenge 2018, 2018. URL: http://www.vacommunity.org/VAST+Challenge+2018.

[VAS*11b] Volzke 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., Junger M., Mayerle J., Kraft M., Lerch M. M., Dorr M., Reffelmann T., Empen K., Felix S. B., Obst A., Koch B., Glaser S., Ewert R., Fietze I., Penzel T., Doren M., Rathmann W., Haerting J., Hannemann M., Ropcke J., Schminke U., Jurgens C., Tost F., Rettig R., Kors J. A., Ungerer S., Hegenscheid K., Kuhn J.-P., Kuhn J., Hosten N., Puls R., Henke J., Gloger O., Teumer A., Homuth G., Volker 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 (apr 2011), 294–307. https://doi.org/10.1093/ije/dyp394.

[vPCB18] van Panhuis W. G., Cross A., Burke D. S.: Project Tycho 2.0: a repository to improve the integration and reuse of data for global population health. *Journal of the American Medical Informatics Association 25*, 12 (dec 2018), 1608–1617. https://doi.org/10.1093/jamia/ocy123.